



Discriminative Phonotactics for Dialect Recognition Using Context-Dependent Phone Classifiers

Fadi Biadisy*, Hagen Soltau†, Lidia Mangu†,
Jiri Navratil†, Julia Hirschberg*

*Department of Computer Science, Columbia University, New York, NY

{fadi, julia}@cs.columbia.edu

† IBM T. J. Watson Research Center, Yorktown Heights, NY

{hsoltau, mangu, jiri}@us.ibm.com

Abstract

In this paper, we introduce a new approach to dialect recognition that relies on context-dependent (CD) phonetic differences between dialects as well as phonotactics. Given a speech utterance, we obtain the phone sequence using a CD-phone recognizer. We then identify the most likely dialect of these CD-phones using SVM classifiers. Augmenting these phones with the output of these classifiers, we extract augmented phonotactic features which are subsequently given to a logistic regression classifier to obtain a dialect detection score. We test our approach on the task of detecting four Arabic dialects from 30s utterances. We compare our performance to two baselines, PRLM and GMM-UBM, as well as to our own improved version of GMM-UBM which employs fMLLR adaptation. Our approach performs significantly better than all three baselines at 5% absolute Equal Error Rate (EER). The overall EER of our system is 6%.

1. Introduction

The last few decades have seen considerable progress in automatically identifying the language of a speaker given a sample of his/her speech. Accent and dialect recognition have more recently begun to receive attention from the speech science and technology communities. The task of dialect identification is the recognition of a speaker's regional dialect, within a pre-determined language, given the acoustic signal alone. Dialect identification for Arabic dialects, in particular, can help in Automatic Speech Recognition (ASR), since speakers with different dialects pronounce some words differently, consistently altering certain phones and even morphemes. Identifying regional dialect prior to ASR allows for the use of a more restricted pronunciation dictionary in decoding, resulting in a reduced search space and lower language modeling perplexity. Moreover, identifying the dialect first will enable the ASR system to adapt its acoustic, morphological, and language models appropriately.

Identifying the regional dialect of a speaker will also provide important benefits for speech technology beyond improving ASR. It will allow us to infer the speakers regional origin and ethnicity, and to adapt the output of text-to-speech synthesis to produce regional speech — important for spoken dialogue systems' development.

The problem of dialect recognition has been viewed as more challenging than that of language recognition due to the greater similarity between dialects of the same language. Although dialects may differ in any dimension(s) of the linguistic spectrum including, morphological, lexical, syntactic, phonetic and

phonological differences, these differences are likely to be more subtle across dialects than those across languages.

In this work, we first attempt to identify the phonetic cues that distinguish different Arabic dialects by training discriminative classifiers. We use these classifiers to extract *augmented* phonotactic features, which are then used to identify the dialect of the speaker. We conduct a series of experiments to test our approach on four Arabic dialects of spontaneous telephone conversations and compare our results to three baselines. In Section 2, we describe related work in language and dialect recognition. The Arabic dialect corpora employed in our experiments are described in Section 3. In Section 4, we describe the front-end and context-dependent phone recognize, we have built for our approach. We discuss the context-dependent classifier in Section 5, and then describe how to use these classifiers to identify linguistic differences between pairs of dialects in Section 6. We described our discriminative phonotactic approach to dialect recognition in Section 7 and discuss experimental results in Section 8. Finally, in Section 9, we conclude and identify directions for future work.

2. Related Work

Some successful approaches to language identification have made use of phonotactic variation. For example, the parallel Phone Recognition followed by Language Modeling (parallel PRLM) [1] approach uses phonotactic information to identify languages using n-gram language models over phones. Zissman et al. [2] show that the PRLM approach yields good results classifying Cuban and Peruvian dialects of Spanish in the Miami corpus, using an English phone recognizer. We have used the parallel PRLM with 9 phone recognizers trained on different languages to distinguish among the four Arabic dialects we examine in this work, as well as Modern Standard Arabic (MSA), in [3]. Shen et al. [4] describe a dialect recognition system that made use of adapted phonetic models per dialect applied in a PRLM framework to distinguish American vs. Indian English and two Mandarin dialects (Mainland and Taiwanese).

Gaussian Mixture Models - Universal Background model (GMM-UBM) has also achieved considerable success in speaker and language recognition [5, 6]. Torres-Carrasquillo et al. [7] developed a system using GMM-UBM with shifted delta cepstral (SDC) features. The system performed worse than that of Zissman et al [2] on the Miami corpus, but performs well on two Mandarin dialects and two Spanish dialects from Call-Home. Discriminative training has proven quite useful in recent language recognition systems (e.g., [8, 9]). Torres-Carrasquillo

et al. [10] showed that a GMM-UBM based model discriminatively trained with SDC features with an eigen-channel compensation component and vocal-tract normalization (VTLN) provides good results for the recognition of American vs. Indian English, four Chinese dialects, and three Arabic dialects (Gulf, Iraqi, and Levantine). Alorfi explores ergodic HMMs to model phonetic differences between two Arabic dialects (Gulf and Egyptian Arabic) employing standard MFCC features [11].

In addition to phonotactic and acoustic-based systems, dialects may also differ in their prosodic structure (e.g., [12, 13]). We observed in our previous work that four Arabic dialects significantly differ in their rhythmic and syllabic structure as well as in vowel durations and some intonational patterns, such as pitch peak alignments within syllables [14]. We showed that modeling a sequence of prosodic features extracted from automatically obtained pseudo syllables using ergodic HMMs significantly improves the results of Parallel PRLM.

3. Corpora

When training a system to recognize languages or dialects, it is essential to use training and testing corpora recorded under similar acoustic conditions. Otherwise, the trained models may capture channel specific information as opposed to linguistic differences. In this work, we test our approach on the following four Arabic dialects.¹

- Iraqi Arabic, including three sub-dialects: Baghdadi, Northern, and Southern.
- Gulf Arabic, including three sub-dialects: Omani, UAE, and Saudi Arabic.
- Levantine Arabic, contains four sub-dialects: Jordanian, Lebanese, Palestinian, and Syrian Arabic.
- Egyptian Arabic, including primarily Cairene Arabic.

The data are spontaneous telephone conversations, produced by native speakers of the dialects, speaking with family members, friends, and unrelated individuals, sometimes about predetermined topics. We use the speech of the 478 speakers from the Iraqi Arabic Conversational Telephone Speech corpus [15], holding out 20% of the speakers for testing. We use the 976 speakers from the Gulf Arabic Conversational Telephone Speech corpus [16], again holding out 20% of the speakers for testing. Our Levantine data consists of 985 speakers from the Levantine Arabic Conversational Telephone Speech corpus [17], also holding out 20% of the speakers for testing. These three corpora were collected by the same company (Appen Pty Ltd) and appear to have been collected under similar conditions. Each of the corpora contains male and female speakers speaking by landline or mobile phones. Since it is likely that the distribution of these categories may influence the trained models, we decided to equalize the number of test speakers in each category. So, our test set for each of the three dialects include: 25% are selected randomly from the set of female speakers speaking on mobile phones; 25% selected from male speakers speaking on mobile phones; 25% selected from females speaking on landline phones; and 25% selected from males speaking over landlines. For the Egyptian dialect corpus, we use the 280 speakers in CallHome Egyptian and its supplement [18] for training. Attempting to test our system on different acoustic conditions, we

¹See [3] for the linguistic background pertaining to these four dialects

employ a completely different corpus for testing: 120 speakers from CallFriend Egyptian [19].² The Egyptian data also includes male and female speakers, but it is not clear if the speakers used landlines, mobile phones, or both. All corpora are provided by the Linguistic Data Consortium (LDC). Although some of the data have been annotated phonetically and/or orthographically by LDC, we do not make use of these annotations for our work.

To identify speech regions in the audio files, we segmented the files based on silence using Praat [20] using a silence threshold of -35db with a minimum silence interval of 0.5s and minimum sounding intervals of 0.5s. All segments were used in training. In this paper, we present results from testing our system on 30-second cuts. Each cut consists of consecutive speech segments totaling 30s in length.³ Multiple cuts are extracted from each speaker. For Iraqi, we have a total of 477 30s test cuts, and 801, 818, 1912 30s test cuts for Gulf, Levantine, and Egyptian, respectively.⁴

4. Context-Dependent Phone Recognizer

To support our approach to dialect recognition, we first build a continuous HMM-based triphone context-dependent (CD) phone recognizer using IBM's Attila system [21]. This phone recognizer is trained on Modern Standard Arabic (MSA) using 50 hours of GALE speech data of broadcast news and broadcast conversations. Our phone recognizer consists of 230 context-dependent acoustic models and a total of 20,000 Gaussians.⁵ The number of Gaussians per state is variable (about 80 Gaussians on average).

We use one acoustic model for silence, one for non-vocal noise and another to model vocal noise. Therefore, in total, we have 227 CD-phones. The set of CD-phones is automatically generated by using a decision tree which asks questions about left and right contexts of each triphone. Contexts with the smallest acoustic difference are clustered together.

The front-end is a 13-dimensional Perceptual Linear Prediction (PLP) front-end with cepstral mean and variance normalization (CMVN). Each frame is spliced together with four preceding and four succeeding frames and then Linear Discriminant Analysis (LDA) is performed to yield 40-dimensional feature vectors. We use the LDA matrix derived for IBM's Attila Arabic ASR system here [21].

All CD-phone HMMs consist of 3 states, except for the the MSA short vowels (/a/ /i/ /u/) which consist of only 2 states.⁶ All state observation densities are Gaussian Mixture Models (GMM). We utilize a unigram language model of phones trained on MSA. We do not use higher order of n-gram to avoid bias for any particular dialect. The pronunciation dictionary and MSA phonetic inventory used in this work are generated as described in [22].

The phone-recognizer is a two-pass system. In the first pass, we obtain the most likely phone sequence hypothesis. The second pass uses this hypothesis to perform model adaptation,

²Because the Egyptian corpora were not collected by the same company which collected the other three corpora.

³N.B. It is sometimes necessary to truncate speaker turns to achieve exactly 30 seconds.

⁴Our training/testing splits and segmentations are available on www.cs.columbia.edu/speech/corpora

⁵We use *only* 230 CD models since we build a classifier for each CD-phone type in Section 5, otherwise we will have data sparsity issue.

⁶It has been previously shown that 2 states for short vowels as opposed to 3 significantly improves ASR word error rate [21].

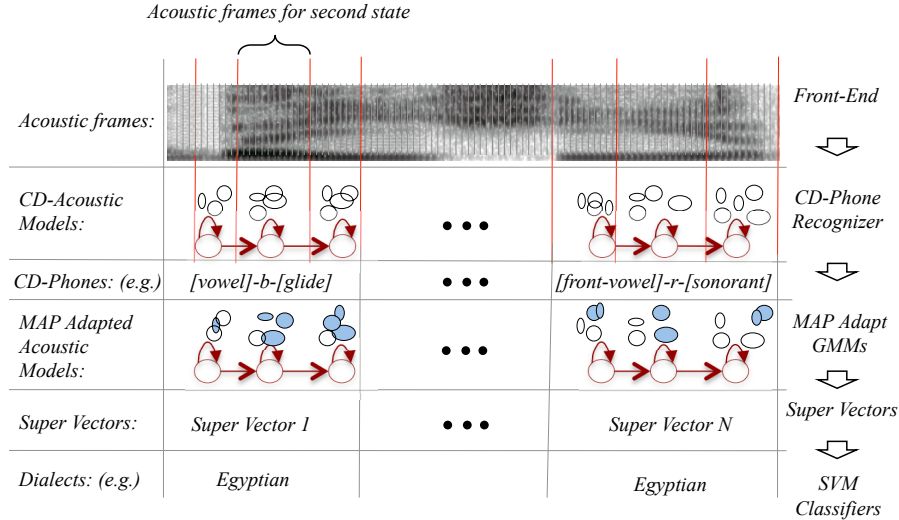


Figure 1: Dialect Classification of Context-Dependent Phones

followed by decoding. In this work, we first apply feature space Maximum Likelihood Linear Regression (fMLLR) followed by MLLR adaptation, given the most likely phone sequence hypothesis. The resulting CD-phones are exemplified by the CD-phone /r/: [Voiced-Consonant & !Glide]-/r/-[Front Vowel].

5. Context-Dependent Phone Classifiers

As noted above, dialects typically differ in some number of phonetic realizations in context. For example, the /r/ in Scottish English is trilled in some phonetic contexts but approximant in others. However, /r/ is typically approximant in American English. In this section, we describe an approach that allows us to classify each CD-phone instance in an utterance as belonging to one of our dialects. This approach is similar in spirit to the GMM-SVM approach introduced by Campbell et al. [23] for speaker verification. However, in our approach, we target the acoustic differences at the level of CD-phones as opposed to the differences in the entire speaker’s acoustic data.

As illustrated in Figure 1, the first step in our approach, after front-end processing, is to obtain the context-dependent phone sequence of a given speech segment. To do this, we run the CD-phone recognizer described in Section 4 to obtain the most likely phone sequence hypothesis. In the second step, for each CD-phone instance in the sequence, we extract the phone acoustic features aligned to each HMM state in the corresponding acoustic model. Note that these features are extracted after normalization (CMVN) and fMLLR transformation. In other words, for each CD-phone instance in the phone sequence, we have one sequence of acoustic frames aligned to the first state in the HMM, and another frame sequence aligned to the second state. If the HMM has three states, then we have also another frame sequence aligned to the third state. See the second row in Figure 1.

Recall that there is a GMM for each HMM state. For each CD-phone instance in the utterance, we use the acoustic frames aligned to the HMM states to MAP (Maximum A-Posteriori) adapt each GMM in each state to the corresponding frames. That is, if the HMM has three states, then we get three new adapted GMMs for each CD-phone instance in the utterance; see the fourth row in Figure 1. In our implementation, we only adapt the means of the Gaussians using a relevance factor of $r = 0.1$. In the context of the GMM-UBM approach, the HMM

can be viewed as the universal background model of the CD-phone type.

Now, to classify a CD-phone instance as belonging to one of our dialects, we adopt the GMM-SVM approach [23] — but at the level of phone instances. We represent each CD-phone instance in the utterance by a *Supervector* which is the result of stacking all the mean vectors of the two or all three adapted GMMs of the CD-phone HMM. The intuition is that some of these adapted means ‘summarize’ the spectral characteristics of the CD-phone instance. We previously observed that the duration of vowels and certain consonants significantly differ across Arabic dialects. Therefore we also include the phone duration as an additional feature in the Supervector of each CD-phone [14].⁷

During training, we apply the procedure described above on the training data to obtain a set of Supervectors for each CD-phone type from each dialect. Using these sets of Supervectors, we train a binary discriminative classifier for each CD-phone type for each pair of dialects. From our 227 CD-phones, we thus have a total of 227 binary classifiers for each pair of dialects. In our implementation, we train SVM classifiers with an RBF kernel.⁸ We have found that an SVM with such a kernel performs significantly better than an SVM with a linear kernel and also better than a logistic regression classifier for the vast majority of the 227 classifiers. During testing, given a CD-phone instance with its frame alignment, we apply the procedure described above to extract its Supervector, and then run the corresponding SVM classifier to classify this CD-phone into one of our dialects.

6. Automatic Extraction of Linguistic Knowledge

There are several uses of our CD-phone classification framework. First, we can utilize it to automatically extract linguistic knowledge, specifically the phonetic cues that may distinguish one of our dialects from another. We are particularly interested

⁷One could add additional prosodic features to the phone vector (such as F0 shape features). We hypothesize that such features would be particularly useful for tonal dialects.

⁸In our implementation, we use LibSVM and LibLinear toolkits [24].

in knowing what phones in which contexts are realized differently across dialects. An empirical measure of the classification performance of each CD-phone classifier provides us with a measure of how the realization of a CD-phone is distinguishable across pairs of dialects.⁹

To extract these phonetic cues, we conducted the following experiment. We split the training speaker set of each dialect into halves. We used the first half to train the CD-phone classifiers for each pair of dialects and the second to test each classifier’s performance. We randomly balanced the number of test instances so that a chance baseline is 50%. For each classifier, using the test instances, we applied the binomial test procedure to identify which CD-phone classifier performed on the test set with a significance level of 0.05. We report on this performance in Table 1 where we show the weighted accuracy of the classifiers that performed significantly better than chance for each dialect pair. We observe that the Egyptian dialect has the highest number of top performing classifiers under our definition.

Dialect Pair	Num. of * classifiers	Weighted accuracy (%)
Egyptian/Iraqi	195	70.9
Egyptian/Gulf	196	69.1
Egyptian/Levantine	199	68.6
Levantine/Iraqi	172	63.96
Gulf/Iraqi	166	61.77
Levantine/Gulf	179	61.53

Table 1: Number of CD-phone classifiers out of the 227 that performed significantly higher than chance for each pair of dialects. * significance level of 0.05.

We report the accuracy of the CD-phone classification results in Table 2-4 for the 10 most and 3 least accurate classifiers for some of our dialect pairs (with significance level of 0.05).¹⁰ The third column in these tables contains the number of instances used in the classification task per dialect. The top 10 CD-phones can be viewed as those that best distinguish between a pair of dialects. We found, for example, that some consonants in the context of central vowels can be useful cues to distinguish Arabic dialects. Moreover, the phoneme /k/ is one of the top 10 cues for distinguishing between Iraqi and Levantine. This might be due to the consistent replacement of the MSA /k/ sound to /ch/ by rural sub-dialects of Levantine.¹¹ These empirical findings can be useful for dialectologists as well as speech scientists and engineers.

It should be noted that substantially more accurate phonetic cues can be obtained by making use of orthographic transcripts in the system instead of using a phone recognizer. In other words, we can do forced-alignment to obtain the phone sequences and then train/analyze the CD-phone classifiers from that. However, we currently lack such orthographic transcripts and/or a pronunciation dictionary that maps our colloquial dialect transcripts onto a shared phonetic inventory.

7. Dialect-Recognition Framework

The task of dialect recognition is the identification of a speakers regional dialect given a sample of his/her speech. We now show

⁹Note that other methods (such as Kullback-Leibler divergence) can be used to quantify differences between adapted dialect acoustic models. However our approach uses held out data instead of “distance” between models. Also our accuracy measures can be more easily interpreted.

¹⁰See [22] for the MSA phonetic symbols used in this work.

¹¹Note that /ch/ and /k/ are modeled as one phoneme in the phone recognizer.

CD-Phone ([l-context]-phone-[r-context])	Accuracy	#
-sh-	71.1	6302
[SIL]-a-*	70.3	3935
[SIL]-?-[Central Vowel]	68.7	1323
-j-	68.5	3722
!Central Vowel]-s-![High Vowel]	68.5	1975
[Nasal]-A-[Anterior]	68.1	5459
!SIL & !Central Vowel]-E-[Central Vowel]	67.8	3687
[Central Vowel]-m-[Central Vowel]	66.7	2639
!Voiced Cons. & !Glottal & !Pharyngeal & !Nasal & !Trill & !w & !Emphatic]-A-[Anterior]	66.4	11857
*-k-[Central Vowel]	66.4	1433
...
!SIL & !Central Vowel]-G-[Central Vowel]	57.5	852
!A]-h-[Back Vowel]	57.0	409
!Vowel & !SIL]-m-[Central Vowel & !Back Vowel]	56.2	300

Table 2: The 10 most and 3 least accurate CD-phone classifiers for Levantine/Iraqi dialects (with significance level of 0.05)

CD-Phone ([l-context]-phone-[r-context])	Accuracy	#
!Central Vowel & !Unvoiced Cons.]-t-[SIL]	71.2	473
-sh-	67.9	6302
[SIL]-w-[Central Vowel]	67.3	745
!Central Vowel]-H-[Central Vowel]	67.0	1234
[SIL]-a-*	66.5	3935
!Central Vowel]-s-![High Vowel]	66.2	1975
[SIL]-b-[Central Vowel & !Front Vowel]	66.1	505
!Central Vowel & !SIL]-b-[Central Vowel]	66.1	750
!SIL & !Central Vowel]-E-[Central Vowel]	65.8	1480
!SIL & !Central Vowel]-E-[Central Vowel]	65.7	3687
...
[Strident]-u-[*]	55.7	380
[Glottal Stop]-a-*	55.3	515
[Pharyngeal]-A-[SIL & !Anterior]	55.1	484

Table 3: The 10 most and 3 least accurate CD-phone classifiers for Gulf/Iraqi dialects

CD-Phone ([l-context]-phone-[r-context])	Accuracy	#
-sh-	80.2	8127
[Central Vowel]-H-[Central Vowel]	77.4	1980
[SIL]-f-[Front Vowel]	76.5	612
[SIL]-m-[Central Vowel]	75.8	2547
*-T-[Central Vowel Vowel]	75.5	1145
!Central Vowel]-s-![High Vowel]	75.3	3396
[SIL]-a-*	75.1	7411
[h]-A-[Anterior]	74.5	1370
!Central Vowel & !Unvoiced Cons.]-t-[SIL]	74.4	857
[SIL]-w-[Central Vowel]	74.1	1534
...
[Front Vowel]-h-[!Back & !Central Vowels]	59.0	183
[Central Vowel]-?-[Central Vowel]	58.4	353
!Vowel & !SIL]-m-[SIL]	57.5	389

Table 4: The 10 most and 3 least accurate CD-phone classifiers for Egyptian/Gulf dialects

how we can employ our CD-phone classification framework to distinguish among dialects.

We have shown in our previous work that Arabic dialects significantly differ in terms of their phonotactic distribution. Particularly, we have shown that Phone Recognition followed by Language Modeling (PRLM) [1] distinguishes Arabic dialects with a high identification accuracy for four Arabic dialects [3]. In this section, we show how we can use the CD-phone classifiers described above to augment the phonotactic approach for dialect recognition. We term this new approach *discriminative phonotactics*.

Given an utterance, we first run our CD-phone recognizer to obtain the most likely CD-phone sequence hypothesis and frame alignment. Then, for each CD-phone in the sequence, we extract its Suprvector and run the corresponding SVM classi-

fier, as described in Section 5. We next attach the classification output to the CD-phone identity itself. If, for example, a CD-phone is [Voiced Cons.]-*r*-[Central Vowel] and the classification output is *Iraqi*, then we get [Voiced Cons.]-*r*-[Central Vowel]_{*Iraqi*}. We apply this procedure for the entire CD-phone sequence, and we denote the output as the *annotated CD-phone sequence*. Note that the idea of appending extra information to the phone identity was suggested by Zissman [1], who attached duration tags (Long/Short) to vowels based on their duration.

Now the task is classifying an annotated CD-phone sequence to one of the dialects. One could simply adopt the PRLM approach using the annotated CD-phone sequences instead of raw phone sequences. However, we decided to treat the problem as one of purely text classification. Instead of applying a generative model (n-grams for each dialect), we train a discriminative classifier for each pair of dialects. In our implementation, we employ logistic regression classifiers. These classifiers are trained on the following list of *textual features* extracted from the annotated phone sequence:

- Frequency of annotated CD-Phone bigrams, e.g.,
“[Nasal]-*r*-[Vowel]_{*Iraqi*} [Voiced Cons.]-*a*-[Liquid]_{*Gulf*}”
- Frequency of bigrams with only one annotated CD-Phone, e.g.,
“[Nasal]-*r*-[Vowel] [Voiced Cons.]-*a*-[Liquid]_{*Gulf*}”
- Frequency of annotated unigrams, e.g.,
“[Central Vowel]-*E*-[Central Vowel]_{*Egyptian*}”
- Frequency of not annotated CD-Phone unigrams and bigrams, e.g.,
“[Nasal]-*r*-[Vowel] [Voiced Cons.]-*a*-[Liquid]”
- Frequency of context *independent* phone *trigrams*, e.g.,
“*s A I*”

We normalize the feature vector by the norm of the vector to address duration differences across samples. Most of our features are *CD-phone* unigrams and bigrams. This is because the classification is performed at the level of CD-phone not — context-independent (CI). Moreover, using CD bi-phones captures phonetic context better than CI bi-phones but less successfully than CI quad-phones. In fact, we found that using a PRLM with bigram models trained on CD-phone sequences, instead of trigrams trained on CI phones, performs slightly better.

There is a commonly held belief that discriminative classifiers are almost always to be preferred over generative classifiers due to modeling directly the posterior probability, or a map from input to class label. It has also been shown empirically that logistic regression and maximum entropy have typically lower asymptotic error than native Bayes for multiple classification tasks as well as for text classification [25, 26]. Moreover, the advantage of using a discriminative classifier over an n-gram model in our case is due to the noisy identity tags attached to phones. An n-gram model trained on such sequences may not be robust; however a logistic classifier with a regularizer or SVM classifier will focus on the informative features and attempt to avoid irrelevant features that do not contribute to the classification task.

In addition, using a classification framework allows us to include different types of features at any level — even global features, which cannot be modeled using an n-gram model. In our experiments, we find that logistic regression with L_2 -regularizer performs slightly worse than SVM with a linear kernel. However, surprisingly, logistic-regression with a L_2 -regularizer typically performs slightly better than a logistic re-

gression with a L_1 -regularizer, although the L_1 -regularizer is known for its feature selection capability [27]. For our detection task, we are interested in using confidence scores. Therefore, we choose logistic regression with a L_2 -regularizer. We will make use of the posterior probability provided by logistic regression as our detection scores, described below.

8. Dialect Recognition Experiments

In this section, we evaluate our discriminative phonotactics approach on the task of Arabic dialect recognition. We compare it to three baselines: a standard phonotactics approach (i.e., PRLM), a standard GMM-UBM approach, and finally our own improved version of GMM-UBM which applies fMLLR adaptation. We adopt the NIST language/dialect and speaker recognition evaluation framework to report detection results instead of identification. In the detection task, we are given a hypothesis and a set of test trials. We are asked to give a decision for each test trial to accept or reject the hypothesis, along with a confidence score. Using these scores, we report our results using Detection Error Tradeoff (DET) figures, which plots false alarms versus miss probabilities, and Equal Error Rate (EER), the error rate when both false alarm and miss probabilities are equal [28]. To plot an overall DET, our results are pooled across each pair of dialects with dialect prior equalized to discount the impact of different number of test trials in each dialect.¹²

8.1. Scoring

We denote the feature vector extracted for a given test trial r , as \mathcal{O}_r . Every test trial is given a confidence score of belonging to target dialect D_t . Assuming that the dialect priors are equal, the posterior probability of \mathcal{O}_r can be reduced to the expression in (1). We use these posterior probabilities to represent our scores, similar to [9]; where \mathcal{D} is the set of dialects of interest, $p(\mathcal{O}_r|\lambda_{D_x})$ represents the likelihood of \mathcal{O}_r given the model λ_{D_x} of dialect D_x , and τ_r normalizes duration differences across trials.

Since we do pairwise detection, for score computation we can make use of the knowledge that an utterance belongs to either the target or non-target dialect. In this paper, we test two scoring schemes: ALLSCORING and PAIRSCORING. In ALLSCORING, we normalize by the likelihoods of \mathcal{O}_r under every model to represent the final score — i.e., \mathcal{D} in (1) contains all our dialects. In PAIRSCORING, we normalize by the likelihoods under the target and non-target dialect models only — i.e., \mathcal{D} in (1) contains only D_t and D_{nt} , the non-target dialect.

$$P(D_t|\mathcal{O}_r) = \frac{p(\mathcal{O}_r|\lambda_{D_t})^{\tau_r}}{\sum_{D_x \in \mathcal{D}} p(\mathcal{O}_r|\lambda_{D_x})^{\tau_r}} \quad (1)$$

8.2. Phonotactics Baseline

As noted above, we have previously shown that a phonotactic approach, particularly the PRLM approach is effective in identifying Arabic dialects. Moreover, since our discriminative phonotactic approach captures phonotactic features as well, we think it is essential to compare both. For PRLM, every non-silent segment in the training data of all dialects is tokenized to the most likely CI phone sequence hypothesis, using the CD-phone recognizer described in Section 4. Afterwards, using the

¹²We use the NIST scoring software developed for LRE07: www.itl.nist.gov/iad/mig/tests/lre/2007

CI phone sequences of dialect D_x , we train a phonotactic back-off trigram model with Witten-Bell smoothing for this dialect, denoted as λ_{D_x} , using the SRILM toolkit [29].

During testing, we calculate the scores as in (1), where \mathcal{O}_r represents the most likely CI phone sequence of trial r , and $p(\mathcal{O}_r|\lambda_{D_x})$ represents the likelihood of \mathcal{O}_r given the phonotactic trigram model λ_{D_x} of dialect D_x , and τ_r is the inverse of the number of phones in the sequence \mathcal{O}_r . Using our test data for the four dialects, and employing ALLSCORING scheme, we found that the overall EER obtained by pooling the six pairs of dialects is 23.0%. Interestingly, when we use the target and non-target models only, i.e., using PAIRSCORING, we achieve a significant improvement: the EER is 17.3%, as shown in Figure 2.

8.3. GMM-UBM Baseline

Gaussian Mixture Model-Universal Background Model (GMM-UBM) is one of the most widely employed approaches in the Speaker and Language Recognition communities [5, 6]. Since, in our discriminative phonotactic approach, the CD-phone classifiers rely upon acoustic features, we believe that it is also essential to compare the performance of our approach to an approach that utilizes such features. For GMM-UBM, we use the same front end described in Section 4 to extract the 40-dimensional PLP features, followed by CMVN. We use an equal number of training frames from three dialects (Iraqi, Gulf, and Levantine) to ML (Maximum Likelihood) train the UBM with 2048 Gaussian components, using the Expectation-Maximization algorithm. Although it has been shown that broader temporal (e.g., Shifted Delta Cepstral) features typically outperform the standard cepstral features [30], we use the same front-end used in the CD-phone classifiers to allow for a simple comparison. Moreover, our features are extracted from a relatively wide context; recall that our 40D PLP features span over 9 frames followed by LDA.

A GMM (λ_{D_x}) is created for each dialect (D_x) by MAP adapting only the means of the UBM using the entire training data for that dialect. We run the MAP adaptation in 5 iterations with a relevance factor of $r = 16$. These settings are similar to [10]. In this work, we do not employ fast scoring.

During testing, we calculate the scores as in (1), where \mathcal{O}_r represents the sequence of 40D PLP features of trial r , and $p(\mathcal{O}_r|\lambda_{D_x})$ represents the likelihood of \mathcal{O}_r given GMM λ_{D_x} of dialect D_x , and τ_r is the inverse of the number of frames in the sequence \mathcal{O}_r .

Similar to the PRLM approach, we use the test data of the four dialects, described in Section 3, to test the performance of the GMM-UBM approach. Again we test the two scoring schemes described above. Using ALLSCORING, which uses all the scores from all GMM models, we obtain an EER of 20%, and, similar to the phonotactics approach, we get a significant improvement when utilizing PAIRSCORING: an EER of 15.3%. Interestingly, this GMM-UBM significantly outperforms the PRLM approach, as shown in Figure 2.

8.4. GMM-UBM with fMLLR Adaptation

It has been shown that the GMM-UBM approach can be improved by applying some normalization/transformation techniques for the acoustic signal. For example, Wong and Sridharan [31] and Torres-Carrasquillo et al. [30] have shown that Vocal Tract Normalization (VTNLN), to remove speaker-depednet features, improves language and dialect recognition results. In addition, channel compensation techniques to retain only lan-

guage dependent information have been shown to significantly improve performance (c.f. [30]).

In this paper, we apply the fMLLR adaptation technique to transform the feature space given the phone hypotheses. Specifically, we first run the CD-phone recognizer to obtain the most likely phone sequences.¹³ Afterwards, we use the CD-phone sequences to transform the acoustic data. Finally, we use the transformed frames as new features in our GMM-UBM approach. To the best of our knowledge, fMLLR has not been employed for the task of language/dialect recognition in such framework.

Applying the same settings of the GMM-UBM experiment above, but with fMLLR adaptation, we achieve an EER of 15.8% with the ALLSCORING scheme. Similar to both PRLM and GMM-UBM, we obtain significantly better results when utilizing PAIRSCORING: an EER of 11.0%. The GMM-UBM approach with fMLLR, interestingly, provides us with significantly higher results when compared to the PLRM and GMM-UBM without adaptation (see Figure 2).

8.5. Discriminative Phonotactics

To explain how we evaluate our discriminative phonotactic approach, recall that we train two types of models for each pair of dialects: the SVM CD-phone classifiers (see Section 5) and a logistic regression classifier, which relies on features extracted in part from the predictions of the SVM classifiers. To train these two models, we divide our training speakers into two sets (SETI and SETII). Initially, we run the CD-phone recognizer on both sets to obtain a CD-phone sequence for each speaker. We then train the SVM CD-phone classifiers using SETI (see Section 5). Afterwards, we use these SVM classifiers to annotate the CD-phone sequences of SETII. Note that we are interested in testing our system on 30s cuts, but our training files are substantially longer. Therefore, all files in SETII are divided into segments of approximately 30s each. We then extract the textual features, described in Section 7, for each of these 30s segments, producing one feature vector for each segment. Using these vectors, we finally train a logistic regression classifier for each pair of dialects.

Note that one way to utilize the entire training data is to use the *second* set for training the SVM classifiers and the *first* set to train the logistic regression classifier. For classification, we simply use the average of the posteriors of both logistic classifiers; we denote this as a *cross training* method.

Recall that, during testing, given a trial r , we first run the CD-phone recognizer to obtain the most likely CD-phone sequence. We then extract a Supervector for each CD-phone. Each Supervector is classified using the corresponding SVM classifier to obtain a dialect label. Attaching the labels to the phones in the CD-phone sequence, we then extract our textual features to obtain a feature vector \mathcal{F}_r . On the hypothesis that each trial is either a target dialect, D_t or non-target D_{nt} , we simply use the posterior probability provided by the corresponding logistic regression model ($\Theta_{D_t D_{nt}}$) to represent our trial score: $p(D_t|\mathcal{F}_r; \Theta_{D_t D_{nt}})$.

As shown in Figure 2, using the discriminative phonotactic approach with the cross-training method, described above, the overall EER after pooling all test trials across dialect is 6.0%. The discriminative phonotactic approach achieves significantly higher results than all baseline approaches above,

¹³Recall that our phone recognizer employs fMLLR followed by MLLR.

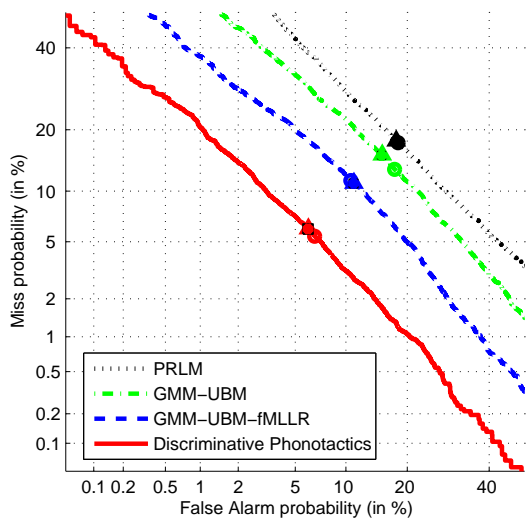


Figure 2: Overall DET for each of the four approaches for the four dialects with the best scoring scheme for all approaches.

PRLM, GMM-UBM, and our own GMM-UBM-fMLLR. Without cross-training, the EER is 6.9%.

We also compare the detection of each dialect against the rest separately, to determine whether the discriminative phonotactic approach outperforms the best baseline (GMM-UBM with fMLLR) in every dialect. As shown in Figure 3, we can see that, for all dialects, the discriminative phonotactic approach is superior when compared to the GMM-UBM with fMLLR. In addition, we can see that the Egyptian dialect is the most distinguishable dialect across all dialects for both GMM-UBM and discriminative phonotactics. This is consistent with our previous work [14, 22]. This could be due to several reasons: (1) According to linguists, the Egyptian Arabic has distinguishable linguistic cues (e.g., syllabic structure is simple); (2) our Egyptian dialect corpus contains mostly Cairene Arabic as opposed to the other dialect corpora which include multiple sub-dialects; (3) the Egyptian test corpus was not collected by the same company which collected the other three dialect corpora. Therefore it is possible that different recording conditions have inflated the results, although our test utterances are from a completely different corpus than the training data.

We have also conducted more experiments in which we exclude the Egyptian dialect from our test trials. For the discriminative phonotactics approach, we obtain 10.5%; we obtain 17.6% for the GMM-UBM with fMLLR adaptation; we obtain 23.1% for the GMM-UBM without adaptation, and 21.5% using the PRLM approach — all using PAIRSCORING scheme.

9. Discussion

In this paper, we have introduced a new approach to dialect recognition that automatically identifies context-dependent phonetic differences between dialects. Our approach discovers which phones in what contexts significantly distinguish between dialect pairs. In developing this approach, we have applied the GMM-SVM approach [23], but at the level of context-dependent (CD) phones to build discriminative classifiers to identify the likely dialect of each CD-phone. We use this infor-

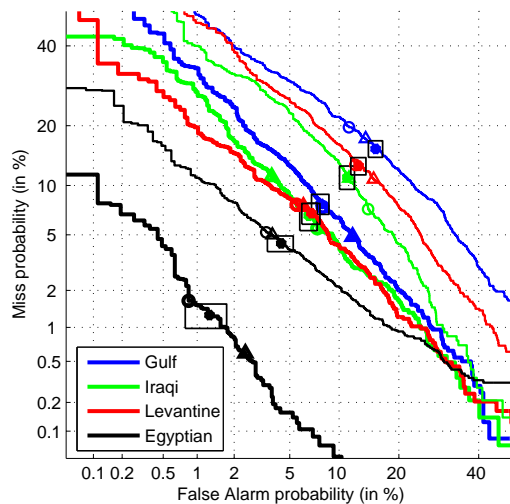


Figure 3: Comparing the overall DET of each dialect against the rest, with two approaches: Discriminative Phonotactics (thicker lines) vs. GMM-UBM-fMLLR

mation to *augment* the phonotactic sequences with dialect labels and then train a discriminative classifier to classify dialects using these augmented phonotactic sequences. Thus, discriminative phonotactic approach can be viewed as taking advantage of both phonotactic and acoustic information in a discriminative manner.

We have analyzed the performance of our discriminative phonotactic approach on detecting four Arabic dialects, with 30 second trials. We have shown that our discriminative phonotactic approach significantly outperforms two well-known baselines (PRLM and GMM-UBM) as well as our own improved version of GMM-UBM which applies fMLLR adaptation to transform the acoustic feature vectors. Discriminative phonotactics achieves an EER of 6.0%, an improvement of 5% (in absolute EER) over our best baseline (GMM-UBM-fMLLR).

We speculate that our approach may also be effective on shorter segments of speech, since we target differences at the level of CD-phones. Our plans for future work include comparing the performance of our system on 3s and 10s utterances to the corresponding baselines. Since it is well-known that backend classifiers typically improve detection results, we would also like to see the impact of such a component on our approach. As noted above, our Gulf, Iraqi and Levantine corpora contain 3-4 sub-dialects each. We hypothesize that teasing these apart may improve the CD-phone classifiers by capturing phonetic differences within these sub-dialects. We intend to integrate confidence scores for the CD-phone classifiers into our discriminative phonotactic approach. In addition, as mentioned above, vocal tract normalization and channel compensation techniques improve language and dialect recognition systems; therefore, we also will test the impact of such techniques on our approach, particularly the SVM nuisance attribute projection (NAP) [32].

Acknowledgments: We would like to acknowledge the support of DARPA under Grant HR0011-06-2-0001 for funding part of this work. We also thank Hong-kwang Kuo, Jason Pelecanos, Michael Picheny, George Saon, and Kapil Thadani for useful discussions.

10. References

- [1] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Transactions of Speech and Audio Processing*, vol. 4, no. 1, 1996.
- [2] M.A. Zissman, T. Gleason, D. Rekart, and B. Losiewicz, "Automatic Dialect Identification of Extemporaneous Conversational, Latin American Spanish Speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, USA, 1996.
- [3] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic Dialect Identification Using Phonotactic Modeling," in *Proceedings of EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, 2009.
- [4] W. Shen, N. Chen, and D. Reynolds, "Dialect recognition using adapted phonetic models," in *Proceedings of INTERSPEECH*, Brisbane, Australia, 2008.
- [5] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19 – 41, 2000.
- [6] E. Wong, J. Pelecanos, S. Myers, and S. Sridharan, "Language identification using efficient gaussian mixture model analysis," in *Australian International Conference on Speech Science and Technology*, 2000.
- [7] P.A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using Gaussian Mixture Models," in *Proceedings of the Speaker and Language Recognition Workshop, Spain*, 2004.
- [8] B. Burget, P. Matejka, and J. Cernock, "Discriminative training techniques for acoustic language identification," in *Proceedings of ICASSP'06*, France, 2006.
- [9] P. Matejka, L. Burget, P. Schwarz, and J. Cernocky, "Brno university of technology system for nist 2005 language recognition evaluation," in *Proceedings of Odyssey*, 2006.
- [10] P.A. Torres-Carrasquillo, D. Sturim, D. Reynolds, and A. McCree, "Eigen-channel Compensation and Discriminatively Trained Gaussian Mixture Models for Dialect and Accent Recognition," in *INTERSPEECH*, Brisbane, Australia, 2008.
- [11] F. S. Alorfi, "PhD Dissertation: Automatic Identification Of Arabic Dialects Using Hidden Markov Models," in *University of Pittsburgh*, 2008.
- [12] M. Barkat, J. Ohala, and F. Pellegrino, "Prosody as a Distinctive Feature for the Discrimination of Arabic Dialects," in *Proceedings of Eurospeech'99*, 1999.
- [13] R. Hamdi, M. Barkat-Defradas, E. Ferragne, and F. Pellegrino, "Speech Timing and Rhythmic Structure in Arabic Dialects: A Comparison of Two Approaches," in *Proceedings of Interspeech'04*, 2004.
- [14] F. Biadsy and J. Hirschberg, "Using Prosody and Phonotactics in Arabic Dialect Identification," in *Proceedings of INTERSPEECH'09*, Brighton, UK, 2009.
- [15] Appen Pty Ltd, "Iraqi Arabic Conversational Telephone Speech – Linguistic Data Consortium, Philadelphia," Sydney, Australia 2006.
- [16] Appen Pty Ltd, "Gulf Arabic Conversational Telephone Speech – Linguistic Data Consortium, Philadelphia," Sydney, Australia, 2006.
- [17] Appen Pty Ltd, "Levantine Arabic Conversational Telephone Speech – Linguistic Data Consortium, Philadelphia," Sydney, Australia, Jan 2007.
- [18] A. Canavan, G. Zipperlen, and D. Graff, "CALLHOME Egyptian Arabic Speech Linguistic Data Consortium, Philadelphia," 1997.
- [19] A. Canavan and G. Zipperlen, "CALLFRIEND Egyptian Arabic Speech Linguistic Data Consortium, Philadelphia," 1996.
- [20] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," 2001, Software available at www.praat.org.
- [21] H. Soltau, G. Saon, B. Kingsbury, H.-K.Kuo, D. Povey, and A. Emami, "Advances in arabic speech transcription at IBM under DARPA GALE program," *EEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 5, pp. 884–895, 2009.
- [22] F. Biadsy, N. Habash, and J. Hirschberg, "Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules," in *Proceedings of NAACL/HLT 2009, Colorado, USA*, 2009.
- [23] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.
- [24] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," 2001, Software available at www.csie.ntu.edu.tw/~cjlin/libsvm.
- [25] A.Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems 14*, 2002.
- [26] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, Stockholm, Sweden, 1999.
- [27] A.Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," in *Proceedings of the 21 st International Conference on Machine Learning*, Banff, Canada, 2004.
- [28] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proceedings of the European Conference on Speech Communication and Technology*, 1997.
- [29] A. Stolcke, "SRILM - an Extensible Language Modeling Toolkit," in *ICASP'02*, 2002, pp. 901–904.
- [30] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr., "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proceedings of ICSLP*, 2002.
- [31] E. Wong and S. Sridharan, "Methods to improve gaussian mixture model based language identification system," in *ICSLP*, 2002.
- [32] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP variability compensation," in *Proceedings of ICASSP'06*, France, May 2006.