Spoken Arabic Dialect Identification Using Phonotactic Modeling

Fadi Biadsy and **Julia Hirschberg**
Department of Computer Science
Columbia University, New York, USA
{fadi,julia}@cs.columbia.edu

Nizar Habash
Center for Computational Learning Systems
Columbia University, New York, USA
habash@ccls.columbia.edu

Abstract

The Arabic language is a collection of multiple variants, among which Modern Standard Arabic (MSA) has a special status as the formal written standard language of the media, culture and education across the Arab world. The other variants are informal spoken dialects that are the media of communication for daily life. Arabic dialects differ substantially from MSA and each other in terms of phonology, morphology, lexical choice and syntax. In this paper, we describe a system that automatically identifies the Arabic dialect (Gulf, Iraqi, Levantine, Egyptian and MSA) of a speaker given a sample of his/her speech. The phonotactic approach we use proves to be effective in identifying these dialects with considerable overall accuracy — 81.60% using 30s test utterances.

1 Introduction

For the past three decades, there has been a great deal of work on the automatic identification (ID) of languages from the speech signal alone. Recently, accent and dialect identification have 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 predetermined language, given a sample of his/her speech. The dialect-identification problem has been viewed as more challenging than that of language ID due to the greater similarity between dialects of the same language. Our goal in this paper is to analyze the effectiveness of a phonotactic approach, i.e. making use primarily of the rules that govern phonemes and their sequences in a language — a techniques which has often been employed by the language ID community — for the identification of Arabic dialects.

The Arabic language has multiple variants, including Modern Standard Arabic (MSA), the formal written standard language of the media, culture and education, and the informal spoken dialects that are the preferred method of communication in daily life. While there are commercially available Automatic Speech Recognition (ASR) systems for recognizing MSA with low error rates (typically trained on Broadcast News), these recognizers fail when a native Arabic speaker speaks in his/her regional dialect. Even in news broadcasts, speakers often *code switch* between MSA and dialect, especially in conversational speech, such as that found in interviews and talk shows. Being able to identify dialect vs. MSA as well as to identify which dialect is spoken during the recognition process will enable ASR engines to adapt their acoustic, pronunciation, morphological, and language models appropriately and thus improve recognition accuracy.

Identifying the regional dialect of a speaker will also provide important benefits for speech technology beyond improving speech recognition. It will allow us to infer the speaker's regional origin and ethnicity and to adapt features used in speaker identification to regional original. It should also prove useful in adapting the output of text-to-speech synthesis to produce regional speech as well as MSA – important for spoken dialogue systems' development.

In Section 2, we describe related work. In Section 3, we discuss some linguistic aspects of Arabic dialects which are important to dialect identification. In Section 4, we describe the Arabic dialect corpora employed in our experiments. In Section 5, we explain our approach to the identification of Arabic dialects. We present our experimental results in Section 6. Finally, we conclude in Section 7 and identify directions for future research.

2 Related Work

A variety of cues by which humans and machines distinguish one language from another have been explored in previous research on language identi-

Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, pages 53–61,
Athens, Greece, 31 March, 2009.

fication. Examples of such cues include phone inventory and phonotactics, prosody, lexicon, morphology, and syntax. Some of the most successful approaches to language ID have made use of phonotactic variation. For example, the Phone Recognition followed by Language Modeling (PRLM) approach uses phonotactic information to identify languages from the acoustic signal alone (Zissman, 1996). In this approach, a phone recognizer (not necessarily trained on a related language) is used to tokenize training data for each language to be classified. Phonotactic language models generated from this tokenized training speech are used during testing to compute language ID likelihoods for unknown utterances.

Similar cues have successfully been used for the identification of regional dialects. Zisssman et al. (1996) show that the PRLM approach yields good results classifying Cuban and Peruvian dialects of Spanish, using an English phone recognizer trained on TIMIT (Garofolo et al., 1993). The recognition accuracy of this system on these two dialects is 84%, using up to 3 minutes of test utterances. Torres-Carrasquillo et al. (2004) developed an alternate system that identifies these two Spanish dialects using Gaussian Mixture Models (GMM) with shifted-delta-cepstral features. This system performs less accurately (accuracy of 70%) than that of (Zissman et al., 1996). Alorfi (2008) uses an ergodic HMM to model phonetic differences between two Arabic dialects (Gulf and Egyptian Arabic) employing standard MFCC (Mel Frequency Cepstral Coefficients) and delta features. With the best parameter settings, this system achieves high accuracy of 96.67% on these two dialects. Ma et al. (2006) use multi-dimensional pitch flux features and MFCC features to distinguish three Chinese dialects. In this system the pitch flux features reduce the error rate by more than 30% when added to a GMM based MFCC system. Given 15s of test-utterances, the system achieves an accuracy of 90% on the three dialects.

Intonational cues have been shown to be good indicators to human subjects identifying regional dialects. Peters et al. (2002) show that human subjects rely on intonational cues to identify two German dialects (Hamburg urban dialects vs. Northern Standard German). Similarly, Barakat et al. (1999) show that subjects distinguish between Western vs. Eastern Arabic dialects significantly above chance based on intonation alone.

Hamdi et al. (2004) show that rhythmic differences exist between Western and Eastern Arabic. The analysis of these differences is done by comparing percentages of vocalic intervals (%V) and the standard deviation of intervocalic intervals (ΔC) across the two groups. These features have been shown to capture the complexity of the syllabic structure of a language/dialect in addition to the existence of vowel reduction. The complexity of syllabic structure of a language/dialect and the existence of vowel reduction in a language are good correlates with the rhythmic structure of the language/dialect, hence the importance of such a cue for language/dialect identification (Ramus, 2002).

As far as we could determine, there is no previous work that analyzes the effectiveness of a phonotactic approach, particularly the parallel PRLM, for identifying Arabic dialects. In this paper, we build a system based on this approach and evaluate its performance on five Arabic dialects (four regional dialects and MSA). In addition, we experiment with six phone recognizers trained on six languages as well as three MSA phone recognizers and analyze their contribution to this classification task. Moreover, we make use of a discriminative classifier that takes all the perplexities of the language models on the phone sequences and outputs the hypothesized dialect. This classifier turns out to be an important component, although it has not been a standard component in previous work.

3 Linguistic Aspects of Arabic Dialects

3.1 Arabic and its Dialects

MSA is the official language of the Arab world. It is the primary language of the media and culture. MSA is syntactically, morphologically and phonologically based on Classical Arabic, the language of the Qur'an (Islam's Holy Book). Lexically, however, it is much more modern. It is not a native language of any Arabs but is the language of education across the Arab world. MSA is primarily written not spoken.

The Arabic dialects, in contrast, are the true native language forms. They are generally restricted in use to informal daily communication. They are not taught in schools or even standardized, although there is a rich popular dialect culture of folktales, songs, movies, and TV shows. Dialects are primarily spoken, not written. However, this is changing as more Arabs gain access to elec-

tronic media such as emails and newsgroups. Arabic dialects are loosely related to Classical Arabic. They are the result of the interaction between different ancient dialects of Classical Arabic and other languages that existed in, neighbored and/or colonized what is today the Arab world. For example, Algerian Arabic has many influences from Berber as well as French.

Arabic dialects vary on many dimensions – primarily, geography and social class. Geolinguistically, the Arab world can be divided in many different ways. The following is only one of many that covers the main Arabic dialects:

- **Gulf Arabic** (GLF) includes the dialects of Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates, and Oman.
- **Iraqi Arabic** (IRQ) is the dialect of Iraq. In some dialect classifications, Iraqi Arabic is considered a sub-dialect of Gulf Arabic.
- **Levantine Arabic** (LEV) includes the dialects of Lebanon, Syria, Jordan, Palestine and Israel.
- **Egyptian Arabic** (EGY) covers the dialects of the Nile valley: Egypt and Sudan.
- **Maghrebi Arabic** covers the dialects of Morocco, Algeria, Tunisia and Mauritania. Libya is sometimes included.

  Yemenite Arabic is often considered its own class. Maltese Arabic is not always considered an Arabic dialect. It is the only Arabic variant that is considered a separate language and is written with Latin script.

Socially, it is common to distinguish three subdialects within each dialect region: city dwellers, peasants/farmers and Bedouins. The three degrees are often associated with a class hierarchy from rich, settled city-dwellers down to Bedouins. Different social associations exist as is common in many other languages around the world.

The relationship between MSA and the dialect in a specific region is complex. Arabs do not think of these two as separate languages. This particular perception leads to a special kind of coexistence between the two forms of language that serve different purposes. This kind of situation is what linguists term *diglossia*. Although the two variants have clear domains of prevalence: formal written (MSA) versus informal spoken (dialect), there is a large gray area in between and it is often filled with a mixing of the two forms.

In this paper, we focus on classifying the dialect of audio recordings into one of five varieties: MSA, GLF, IRQ, LEV, and EGY. We do not address other dialects or diglossia.

3.2 Phonological Variations among Arabic Dialects

Although Arabic dialects and MSA vary on many different levels — phonology, orthography, morphology, lexical choice and syntax — we will focus on phonological difference in this paper.[1] MSA's phonological profile includes 28 consonants, three short vowels, three long vowels and two diphthongs (/ay/ and /aw/). Arabic dialects vary phonologically from standard Arabic and each other. Some of the common variations include the following (Holes, 2004; Habash, 2006):

The MSA consonant (/q/) is realized as a glottal stop /'/ in EGY and LEV and as /g/ in GLF and IRQ. For example, the MSA word /t̲ari:q/ 'road' appears as /t̲ari:'/ (EGY and LEV) and /t̲ari:g/ (GLF and IRQ). Other variants also are found in sub dialects such as /k/ in rural Palestinian (LEV) and /dj/ in some GLF dialects. These changes do not apply to modern and religious borrowings from MSA. For instance, the word for 'Qur'an' is never pronounced as anything but /qur'a:n/.

The MSA alveolar affricate (/dj/) is realized as /g/ in EGY, as /j/ in LEV and as /y/ in GLF. IRQ preserves the MSA pronunciation. For example, the word for 'handsome' is /djami:l/ (MSA, IRQ), /gami:l/ (EGY), /jami:l/ (LEV) and /yami:l/ (GLF).

The MSA consonant (/k/) is generally realized as /k/ in Arabic dialects with the exception of GLF, IRQ and the Palestinian rural sub-dialect of LEV, which allow a /č/ pronunciation in certain contexts. For example, the word for 'fish' is /samak/ in MSA, EGY and most of LEV but /simač/ in IRQ and GLF.

The MSA consonant /θ/ is pronounced as /t/ in LEV and EGY (or /s/ in more recent borrowings from MSA), e.g., the MSA word /θala:θa/ 'three' is pronounced /tala:ta/ in EGY and /tla:te/ in LEV. IRQ and GLF generally preserve the MSA pronunciation.

[1] It is important to point out that since Arabic dialects are not standardized, their orthography may not always be consistent. However, this is not a relevant point to this paper since we are interested in dialect identification using audio recordings and without using the dialectal transcripts at all.

The MSA consonant /δ/ is pronounced as /d/ in LEV and EGY (or /z/ in more recent borrowings from MSA), e.g., the word for 'this' is pronounced /ha:δa/ in MSA versus /ha:da/ (LEV) and /da/ EGY. IRQ and GLF generally preserve the MSA pronunciation.

The MSA consonants /$\underline{d}$/ (emphatic/velarized d) and /$\underline{\delta}$/ (emphatic /δ/) are both normalized to /$\underline{d}$/ in EGY and LEV and to /$\underline{\delta}$/ in GLF and IRQ. For example, the MSA sentence /$\underline{\delta}$alla ya$\underline{d}$rubu/ 'he continued to hit' is pronounced /$\underline{d}$all yu$\underline{d}$rub/ (LEV) and /$\underline{\delta}$all yu$\underline{\delta}$rub/ (GLF). In modern borrowings from MSA, /$\underline{\delta}$/ is pronounced as /$\underline{z}$/ (emphatic z) in EGY and LEV. For instance, the word for 'police officer' is /$\underline{\delta}$a:bi$\underline{t}$/ in MSA but /$\underline{z}$a:bi$\underline{t}$/ in EGY and LEV.

In some dialects, a loss of the emphatic feature of some MSA consonants occurs, e.g., the MSA word /la$\underline{t}$i:f/ 'pleasant' is pronounced as /lati:f/ in the Lebanese city sub-dialect of LEV. Emphasis typically spreads to neighboring vowels: if a vowel is preceded or succeeded directly by an emphatic consonant (/$\underline{d}$/, /$\underline{s}$/, /$\underline{t}$/, /$\underline{\delta}$/) then the vowel becomes an emphatic vowel. As a result, the loss of the emphatic feature does not affect the consonants only, but also their neighboring vowels.

Other vocalic differences among MSA and the dialects include the following: First, short vowels change or are completely dropped, e.g., the MSA word /yaktubu/ 'he writes' is pronounced /yiktib/ (EGY and IRQ) or /yoktob/ (LEV). Second, final and unstressed long vowels are shortened, e.g., the word /ma$\underline{t}$a:ra:t/ 'airports' in MSA becomes /ma$\underline{t}$ara:t/ in many dialects. Third, the MSA diphthongs /aw/ and /ay/ have mostly become /o:/ and /e:/, respectively. These vocalic changes, particularly vowel drop lead to different syllabic structures. MSA syllables are primarily light (CV, CV:, CVC) but can also be (CV:C and CVCC) in utterance-final positions. EGY syllables are the same as MSA's although without the utterance-final restriction. LEV, IRQ and GLF allow heavier syllables including word initial clusters such as CCV:C and CCVCC.

4 Corpora

When training a system intended to classify languages or dialects, it is of course important to use training and testing corpora recorded under similar acoustic conditions. We are able to obtain corpora from the Linguistic Data Consortium (LDC) with similar recording conditions for four Arabic dialects: Gulf Arabic, Iraqi Arabic, Egyptian Arabic, and Levantine Arabic. These are corpora of spontaneous telephone conversations produced by native speakers of the dialects, speaking with family members, friends, and unrelated individuals, sometimes about predetermined topics. Although, the data have been annotated phonetically and/or orthographically by LDC, in this paper, we do not make use of any of annotations.

We use the speech files of 965 speakers (about 41.02 hours of speech) from the Gulf Arabic conversational telephone Speech database for our Gulf Arabic data (Appen Pty Ltd, 2006a).[2] From these speakers we hold out 150 speakers for testing (about 6.06 hours of speech).[3] We use the Iraqi Arabic Conversational Telephone Speech database (Appen Pty Ltd, 2006b) for the Iraqi dialect, selecting 475 Iraqi Arabic speakers with a total duration of about 25.73 hours of speech. From these speakers we hold out 150 speakers[4] for testing (about 7.33 hours of speech). Our Levantine data consists of 1258 speakers from the Arabic CTS Levantine Fisher Training Data Set 1-3 (Maamouri, 2006). This set contains about 78.79 hours of speech in total. We hold out 150 speakers for testing (about 10 hours of speech) from Set 1.[5] For our Egyptian data, we use CallHome Egyptian and its Supplement (Canavan et al., 1997) and CallFriend Egyptian (Canavan and Zipperlen, 1996). We use 398 speakers from these corpora (75.7 hours of speech), holding out 150 speakers for testing.[6] (about 28.7 hours of speech.)

Unfortunately, as far as we can determine, there is no data with similar recording conditions for MSA. Therefore, we obtain our MSA training data from TDT4 Arabic broadcast news. We use about 47.6 hours of speech. The acoustic signal was processed using forced-alignment with the transcript to remove non-speech data, such as music. For testing we again use 150 speakers, this time identified automatically from the GALE Year 2 Distillation evaluation corpus (about 12.06 hours of speech). Non-speech data (e.g., music) in the test

[2] We excluded very short speech files from the corpora.

[3] The 24 speakers in *devtest* folder and the last 63 files, after sorting by file name, in *train2c* folder (126 speakers). The sorting is done to make our experiments reproducible by other researchers.

[4] Similar to the Gulf corpus, the 24 speakers in *devtest* folder and the last 63 files (after sorting by filename) in *train2c* folder (126 speakers)

[5] We use the last 75 files in Set 1, after sorting by name.

[6] The test speakers were from *evaltest* and *devtest* folders in CallHome and CallFriend.

corpus was removed manually. It should be noted that the data includes read speech by anchors and reporters as well as spontaneous speech spoken in interviews in studios and though the phone.

5 Our Dialect ID Approach

Since, as described in Section 3, Arabic dialects differ in many respects, such as phonology, lexicon, and morphology, it is highly likely that they differ in terms of phone-sequence distribution and phonotactic constraints. Thus, we adopt the phonotactic approach to distinguishing among Arabic dialects.

5.1 PRLM for dialect ID

As mentioned in Section 2, the PRLM approach to language identification (Zissman, 1996) has had considerable success. Recall that, in the PRLM approach, the phones of the training utterances of a dialect are first identified using a single phone recognizer.[7] Then an n-gram language model is trained on the resulting phone sequences for this dialect. This process results in an n-gram language model for each dialect to model the dialect distribution of phone sequence occurrences. During recognition, given a test speech segment, we run the phone recognizer to obtain the phone sequence for this segment and then compute the perplexity of each dialect n-gram model on the sequence. The dialect with the n-gram model that minimizes the perplexity is hypothesized to be the dialect from which the segment comes.

Parallel PRLM is an extension to the PRLM approach, in which multiple (k) parallel phone recognizers, each trained on a different language, are used instead of a single phone recognizer (Zissman, 1996). For training, we run all phone recognizers in parallel on the set of training utterances of each dialect. An n-gram model on the outputs of each phone recognizer is trained for each dialect. Thus if we have m dialects, k x m n-gram models are trained. During testing, given a test utterance, we run all phone recognizers on this utterance and compute the perplexity of each n-gram model on the corresponding output phone sequence. Finally, the perplexities are fed to a combiner to determine the hypothesized dialect. In our implementation, we employ a logistic regression classifier as our back-end combiner. We have experimented with different classifiers such as SVM, and neural networks, but logistic regression classifier was superior. The system is illustrated in Figure 1.

We hypothesize that using multiple phone recognizers as opposed to only one allows the system to capture subtle phonetic differences that might be crucial to distinguish dialects. Particularly, since the phone recognizers are trained on different languages, they may be able to model different vocalic and consonantal systems, hence a different phonetic inventory. For example, an MSA phone recognizer typically does not model the phoneme /g/; however, an English phone recognizer does. As described in Section 3, this phoneme is an important cue to distinguishing Egyptian Arabic from other Arabic dialects. Moreover, phone recognizers are prone to many errors; relying upon multiple phone streams rather than one may lead to a more robust model overall.

5.2 Phone Recognizers

In our experiments, we have used phone recognizers for English, German, Japanese, Hindi, Mandarin, and Spanish, from a toolkit developed by Brno University of Technology.[8] These phone recognizers were trained on the OGI multilanguage database (Muthusamy et al., 1992) using a hybrid approach based on Neural Networks and Viterbi decoding without language models (open-loop) (Matejka et al., 2005).

Since Arabic dialect identification is our goal, we hypothesize that an Arabic phone recognizer would also be useful, particularly since other phone recognizers do not cover all Arabic consonants, such as pharyngeals and emphatic alveolars. Therefore, we have built our own MSA phone recognizer using the HMM toolkit (HTK) (Young et al., 2006). The monophone acoustic models are built using 3-state continuous HMMs without state-skipping, with a mixture of 12 Gaussians per state. We extract standard Mel Frequency Cepstral Coefficients (MFCC) features from 25 ms frames, with a frame shift of 10 ms. Each feature vector is 39D: 13 features (12 cepstral features plus energy), 13 deltas, and 13 double-deltas. The features are normalized using cepstral mean normalization. We use the Broadcast News TDT4 corpus (Arabic Set 1; 47.61 hours of speech; downsampled to 8Khz) to train our acoustic models. The

[7] The phone recognizer is typically trained on one of the languages being identified. Nonetheless, a phone recognize trained on any language might be a good approximation, since languages typically share many phones in their phonetic inventory.

[8] www.fit.vutbr.cz/research/groups/speech/sw/phnrec

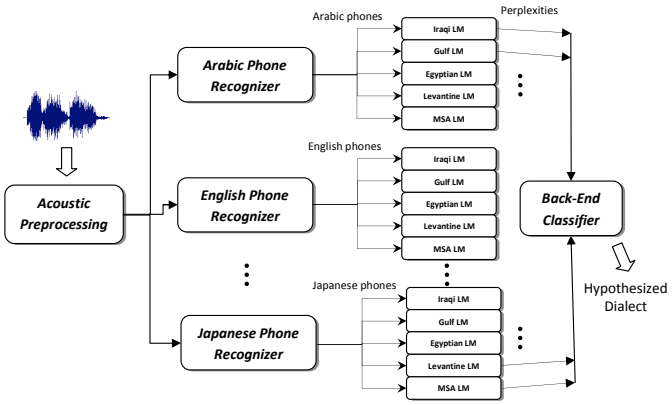

Figure 1: Parallel Phone Recognition Followed by Language Modeling (PRLM) for Arabic Dialect Identification.

pronunciation dictionary is generated as described in (Biadsy et al., 2009). Using these settings we build three MSA phone recognizers: (1) an open-loop phone recognizer which does not distinguish emphatic vowels from non-emphatic **(ArbO)**, (2) an open-loop with emphatic vowels **(ArbOE)**, and (3) a phone recognizer with emphatic vowels and with a bi-gram phone language model **(ArbLME)**. We add a new pronunciation rule to the set of rules described in (Biadsy et al., 2009) to distinguish emphatic vowels from non-emphatic ones (see Section 3) when generating our pronunciation dictionary for training the acoustic models for the the phone recognizers. In total we build 9 (Arabic and non-Arabic) phone recognizers.

6 Experiments and Results

In this section, we evaluate the effectiveness of the parallel PRLM approach on distinguishing Arabic dialects. We first run the nine phone recognizers described in Section 5 on the training data described in Section 4, for each dialect. This process produces nine sets of phone sequences for each dialect. In our implementation, we train a tri-gram language model on each phone set using the SRILM toolkit (Stolcke, 2002). Thus, in total, we have 9 x *(number of dialects)* tri-grams.

In all our experiments, the 150 test speakers of each dialect are first decoded using the phone recognizers. Then the perplexities of the corresponding tri-gram models on these sequences are computed, and are given to the logistic regression classifier. Instead of splitting our held-out data into test and training sets, we report our results with 10-fold cross validation.

We have conducted three experiments to evaluate our system. The first is to compare the performance of our system to Alorfi's (2008) on the same two dialects (Gulf and Egyptian Arabic). The second is to attempt to classify four colloquial Arabic dialects. In the third experiment, we include MSA as well in a five-way classification task.

6.1 Gulf vs. Egyptian Dialect ID

To our knowledge, Alorfi's (2008) work is the only work dealing with the automatic identification of Arabic dialects. In this work, an Ergodic HMM is used to model phonetic differences between Gulf and Egyptian Arabic using MFCC and delta features. The test and training data used in this work was collected from TV soap operas containing both the Egyptian and Gulf dialects and from twenty speakers from CallHome Egyptian database. The best accuracy reported by Alorfi (2008) on identifying the dialect of 40 utterances of duration of 30 seconds each of 40 male speakers (20 Egyptians and 20 Gulf speakers) is 96.67%.

Since we do not have access to the test collection used in (Alorfi, 2008), we test a version of our system which identifies these two dialects only on our 150 Gulf and 150 Egyptian speakers, as described in Section 4. Our best result is 97.00% (Egyptian and Gulf F-Measure = 0.97) when using only the features from the ArbOE, English, Japanese, and Mandarin phone recognizers. While our accuracy might not be *significantly* higher than that of Alorfi's, we note a few advantages of our experiments. First, the test sets of both dialects are from telephone conversations, with the same recording conditions, as opposed to a mix of different genres. Second, in our system we test 300 speakers as oppose to 40, so our results may be more reliable. Third, our test data includes female

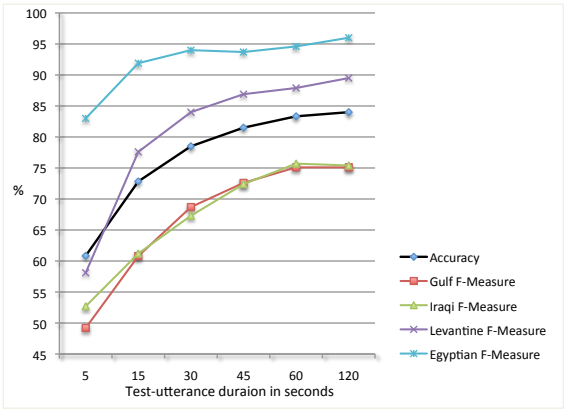

Figure 2: The accuracies and F-Measures of the four-way classification task with different test-utterance durations

speakers as well as male, so our results are more general.

6.2 Four Colloquial Arabic Dialect ID

In our second experiment, we test our system on four colloquial Arabic dialects (Gulf, Iraqi, Levantine, and Egyptian). As mentioned above, we use the phone recognizers to decode the training data to train the 9 tri-gram models per dialect (9x4=36 tri-gram models). We report our 10-fold cross validation results on the test data in Figure 2. To analyze how dependent our system is on the duration of the test utterance, we report the system accuracy and the F-measure of each class for different durations (5s – 2m). The longer the utterance, the better we expect the system to perform. We can observe from these results that regardless of the test-utterance duration, the best distinguished dialect among the four dialects is Egyptian (F-Measure of 94% with 30s test utterances), followed by Levantine (F-Measure of 84% with 30s), and the most confusable dialects, according to the classification confusion matrix, are those of the Gulf and Iraqi Arabic (F-Measure of 68.7%, 67.3%, respectively with 30s). This confusion is consistent with dialect classifications that consider Iraqi a sub-dialect of Gulf Arabic, as mentioned in Section 3.

We were also interested in testing which phone recognizers contribute the most to the classification task. We observe that employing a subset of the phone recognizers as opposed to all of them provides us with better results. Table 1 shows which phone recognizers are selected empirically, for each test-utterance duration condition.[9]

[9]Starting from all phone recognizers, we remove one recognizer at a time; if the cross-validation accuracy decreases, we add it back. We have experimented with an automatic feature selection methods, but with the empirical ('greedy') selection we typically obtain higher accuracy.

Dur.	Acc. (%)	Phone Recognizers
5s	60.83	ArbOE+ArbLME+G+H+M+S
15s	72.83	ArbOE+ArbLME+G+H+M
30s	78.50	ArbO+H+S
45s	81.5	ArbE+ArbLME+H+G+S
60s	83.33	ArbOE+ArbLME+E+G+H+M
120s	84.00	ArbOE+ArbLME+G+M

Table 1: Accuracy of the four-way classification (four colloquial Arabic dialects) and the best combination of phone recognizers used per test-utterances duration; The phone recognizers used are: E=English, G=German, H=Hindi, M=Mandarin, S=Spanish, ArbO=open-loop MSA without emphatic vowels, ArbOE=open-loop MSA with emphatic vowels, ArbLME=MSA with emphatic vowels and bi-gram phone LM

We observe that the MSA phone recognizers are the most important phone recognizers for this task, usually when emphatic vowels are modeled. In all scenarios, removing all MSA phone recognizers leads to a significant drop in accuracy. German, Mandarin, Hindi, and Spanish typically contribute to the classification task, but English, and Japanese phone recognizers are less helpful. It is possible that the more useful recognizers are able to capture more of the distinctions among the Arabic dialects; however, it might also be that the overall quality of the recognizers also varies.

6.3 Dialect ID with MSA

Considering MSA as a dialectal variant of Arabic, we are also interested in analyzing the performance of our system when including it in our classification task. In this experiment, we add MSA as the fifth dialect. We perform the same steps described above for training, using the MSA corpus described in Section 4. For testing, we use also our 150 hypothesized MSA speakers as our test set. Interestingly, in this five-way classification, we observe that the F-Measure for the MSA class in the cross-validation task is always above 98% regardless of the test-utterance duration, as shown in Figure 3.

It would seem that MSA is rarely confused with any of the colloquial dialects: it appears to have a distinct phonotactic distribution. This explanation is supported by linguists, who note that MSA differs from Arabic dialects in terms of its phonology, lexicon, syntax and morphology, which appears to lead to a profound impact on its phonotactic distribution. Similar to the four-way classification task,

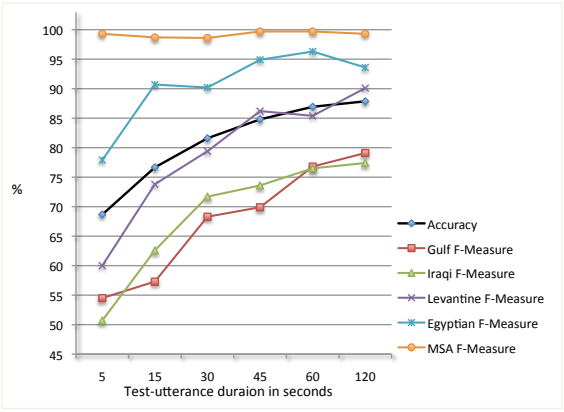

Figure 3: The accuracies and F-Measures of the five-way classification task with different test-utterance durations

Dur.	Acc. (%)	Phone Recognizers
5s	68.67	ArbO+ArbLME+H+M
15s	76.67	ArbLME+G+H+J+M
30s	81.60	ArbO+ArbOE+E+G+H+J+M+S
45s	84.80	ArbOE+ArbLME+E+G+H+J+M+S
60s	86.93	ArbOE+ArbLME+G+J+M+S
120s	87.86	ArbO+ArbLME+E+S

Table 2: Accuracy of the five-way classification (4 colloquial Arabic dialects + MSA) and the best combination of phone recognizers used per test-utterances duration; The phone recognizers used are: E=English, G=German, H=Hindi, J=Japanese, M=Mandarin, S=Spanish, ArbO=open-loop MSA without emphatic vowels, ArbOE=open-loop MSA with emphatic vowels, ArbLME=MSA with emphatic vowels and bi-gram phone LM

Egyptian was the most easily distinguished dialect (F-Measure=90.2%, with 30s test utterance) followed by Levantine (79.4%), and then Iraqi and Gulf (71.7% and 68.3%, respectively). Due to the high MSA F-Measure, the five-way classifier can also be used as a binary classifier to distinguish MSA from colloquial Arabic (Gulf, Iraqi, Levantine, and Egyption) reliably.

It should be noted that our classification results for MSA might be inflated for several reasons: (1) The MSA test data were collected from Broadcast News, which includes read (anchor and reporter) speech, as well as telephone speech (for interviews). (2) The identities of the test speakers in the MSA corpus were determined automatically, and so might not be as accurate.

As a result of the high identification rate of MSA, the overall accuracy in the five-way classification task is higher than that of the four-way classification. Table 2 presents the phone recognizers selected the accuracy for each test utterance duration. We observe here that the most important phone recognizers are those trained on MSA (ArbO, ArbOE, and/or ArbLME). Removing them completely leads to a significant drop in accuracy. In this classification task, we observe that all phone recognizers play a role in the classification task in some of the conditions.

7 Conclusions and Future Work

In this paper, we have shown that four Arabic colloquial dialects (Gulf, Iraqi, Levantine, and Egyptian) plus MSA can be distinguished using a phonotactic approach with good accuracy. The parallel PRLM approach we employ thus appears to be effective not only for language identification but also for Arabic dialect ID.

We have found that the most distinguishable dialect among the five variants we consider here is MSA, independent of the duration of the test-utterance (F-Measure is always above 98.00%). Egyptian Arabic is second (F-Measure of 90.2% with 30s test-utterances), followed by Levantine (F-Measure of 79.4%, with 30s test). The most confusable dialects are Iraqi and Gulf (F-Measure of 71.7% and 68.3%, respectively, with 30s test-utterances). This high degree of Iraqi-Gulf confusion is consistent with some classifications of Iraqi Arabic as a sub-dialect of Gulf Arabic. We have obtained a total accuracy of 81.60% in this five-way classification task when given 30s-duration utterances. We have also observed that the most useful phone streams for classification are those of our Arabic phone recognizers — typically those with emphatic vowels.

As mentioned above, the high F-measure for MSA may be due to the MSA corpora we have used, which differs in genre from the dialect corpora. Therefore, one focus of our future research will be to collect MSA data with similar recording conditions to the other dialects to validate our results. We are also interested in including prosodic features, such as intonational, durational, and rhythmic features in our classification. A more long-term and general goal is to use our results to improve ASR for cases in which code-switching occurs between MSA and other dialects.

Acknowledgments

We thank Dan Ellis, Michael Mandel, and Andrew Rosenberg for useful discussions. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 (approved for public release, distribution unlimited). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

F. S. Alorfi. 2008. PhD Dissertation: Automatic Identification Of Arabic Dialects Using Hidden Markov Models. In *University of Pittsburgh*.

Appen Pty Ltd. 2006a. Gulf Arabic Conversational Telephone Speech Linguistic Data Consortium, Philadelphia.

Appen Pty Ltd. 2006b. Iraqi Arabic Conversational Telephone Speech Linguistic Data Consortium, Philadelphia.

M. Barkat, J. Ohala, and F. Pellegrino. 1999. Prosody as a Distinctive Feature for the Discrimination of Arabic Dialects. In *Proceedings of Eurospeech'99*.

F. Biadsy, N. Habash, and J. Hirschberg. 2009. Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules. In *Proceedings of NAACL/HLT 2009, Colorado, USA*.

A. Canavan and G. Zipperlen. 1996. CALLFRIEND Egyptian Arabic Speech Linguistic Data Consortium, Philadelphia.

A. Canavan, G. Zipperlen, and D. Graff. 1997. CALLHOME Egyptian Arabic Speech Linguistic Data Consortium, Philadelphia.

J. S. Garofolo et al. 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus Linguistic Data Consortium, Philadelphia.

N. Habash. 2006. On Arabic and its Dialects. *Multilingual Magazine*, 17(81).

R. Hamdi, M. Barkat-Defradas, E. Ferragne, and F. Pellegrino. 2004. Speech Timing and Rhythmic Structure in Arabic Dialects: A Comparison of Two Approaches. In *Proceedings of Interspeech'04*.

C. Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press. Revised Edition.

B. Ma, D. Zhu, and R. Tong. 2006. Chinese Dialect Identification Using Tone Features Based On Pitch Flux. In *Proceedings of ICASP'06*.

M. Maamouri. 2006. Levantine Arabic QT Training Data Set 5, Speech Linguistic Data Consortium, Philadelphia.

P. Matejka, P. Schwarz, J. Cernocky, and P. Chytil. 2005. Phonotactic Language Identification using High Quality Phoneme Recognition. In *Proceedings of Eurospeech'05*.

Y. K. Muthusamy, R.A. Cole, and B.T. Oshika. 1992. The OGI Multi-Language Telephone Speech Corpus. In *Proceedings of ICSLP'92*.

J. Peters, P. Gilles, P. Auer, and M. Selting. 2002. Identification of Regional Varieties by Intonational Cues. An Experimental Study on Hamburg and Berlin German. 45(2):115–139.

F. Ramus. 2002. Acoustic Correlates of Linguistic Rhythm: Perspectives. In *Speech Prosody*.

A. Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *ICASP'02*, pages 901–904.

P. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds. 2004. Dialect identification using Gaussian Mixture Models. In *Proceedings of the Speaker and Language Recognition Workshop, Spain*.

S. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2006. The HTK Book, version 3.4.

M. A. Zissman, T. Gleason, D. Rekart, and B. Losiewicz. 1996. 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.

M. A. Zissman. 1996. Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. *IEEE Transactions of Speech and Audio Processing*, 4(1).