Effective Arabic Dialect Classification Using Diverse Phonotactic Models

Murat Akbacak[1] *Dimitra Vergyri*[1] *Andreas Stolcke*[2] *Nicolas Scheffer*[1] *Arindam Mandal*[1]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
[2]Microsoft Speech Labs, Mountain View, CA, USA
{murat,dverg,scheffer,arindam}@speech.sri.com, anstolck@microsoft.com

Abstract

We study the effectiveness of recently developed language recognition techniques based on speech recognition models for the discrimination of Arabic dialects. Specifically, we investigate dialect-specific and cross-dialectal phonotactic models, using both language models and support vector machines (SVMs). Techniques are evaluated both alone and in combination with a cepstral system with joint factor analysis (JFA), using a four-dialect data set employing 30-second telephone speech samples. We find good complementarity from different features and modeling paradigms, and achieve 2% average equal error rate for pairwise classification.

1. Introduction

Over the last few years, as the speech community is targeting more natural and inherently variable speech recognition tasks, accent and dialect identification have become increasingly important research areas. Accent and dialect information is valuable in itself to allow inferences about the speaker's geographic or cultural background. Furthermore, such information enables effective adaptation of speech and language processing systems, e.g., by switching to specialized acoustic, pronunciation, or language models in speech recognition. It is thus important to be able to label unannotated data with high accuracy to enable the application of dialect-specific models. Several languages of great practical importance, such as Chinese and Arabic, comprise a variety of regional dialects that can differ significantly from each other, and can even be mutually unintelligible.

Earlier studies employed Gaussian mixture models (GMMs) for dialect and accent classification, e.g., to identify native multi-accented Mandarin [1, 2]. Recently, approaches originally developed for language recognition (or language identification (LID)) have been applied successfully to the dialect identification task. Typically, LID techniques are based on a combination of two main modeling approaches. In one approach, short-term cepstral features, typically including shifted delta cepstral (SDC) features [3], are modeled using a universal background model (GMM-UBM) with joint factor analysis (JFA) [4, 5]. For the second approach, one or more language-specific, unconstrained phone recognizers are used to tokenize the speech into phone sequences, which are then characterized using statistical language models [6] ([parallel] phone recognition language modeling, [P]PRLM). Scores from both kinds of system are combined and calibrated using a variety of techniques, such as Gaussian back ends [7] or multiclass logistic regression [8]. Both of these approaches have been applied to dialect identification: Torres-Carrasquillo *et al.* [7], for example, showed that a GMM-UBM based model discriminatively trained with SDC features with an eigenchannel compensation component and vocal-tract normalization produces good results for discriminating American versus Indian English, four Chinese dialects, and three Arabic dialects. Zissman *et al.* [9] found PRLM approaches to be very effective for discriminating between Cuban and Peruvian dialects of Spanish, and similarly Shen *et al.* [10] for discriminating between American and Indian English, and Mainland and Taiwanese Mandarin.

Our own prior work has focused on the use of advanced modeling techniques from speech and speaker recognition to the LID task [11]. A key improvement came from the application of multilingual phone recognizers for phonotactic models. In this paper we investigate how these techniques fare when applied to the classification of confusable languages, or more specifically, regional dialects of the same language.

We evaluate different features, models, and classification techniques on the task of Arabic dialect identification. Arabic is a Semitic language characterized by a wide range of spoken regional dialects. The greatest differences are those between dialects according to geographical/linguistic regions. These can be divided in several ways, but the following typology is commonly used: (1) Levantine, spoken in Lebanon, Syria and Palestine, (2) Iraqi or Mesopotamian, spoken in Iraq, (3) Arabian peninsula (Gulf) dialects, spoken in Saudi Arabia and the Gulf area, (4) Egyptian, (5) Maghreb dialects, spoken in North Africa. Some of these varieties are mutually unintelligible with others. These *colloquial* dialects represent spoken forms of the language, and are in contrast to *Modern Standard Arabic* (MSA), which is both written and spoken, used in public and formal discourse, and serves as a lingua franca between speakers of different dialects.

In this paper, we focus on discriminating among four Arabic colloquial dialects for which we have similar data sets available from the Linguistic Data Consortium (LDC): Levantine, Iraqi, Gulf, and Egyptian. We start by examining the use of the cepstral GMM-UBM approach with JFA for this task. Then, we explore phone-recognition-based systems, using a phone recognizer trained on different individual languages, as well as a multilingual and a Pan-Arabic phone set, and corresponding phone recognizers for PRLM and support vector machine (SVM) systems. We compare the effectiveness of individual systems, as well as their benefit in system combination at the score level.

2. Data

The experiments reported here use the data set prepared by Biadsy [12, 13]. All data was drawn from the LDC conversational telephone speech (CTS) collections for Levantine, Iraqi, Gulf (Appen Pty collection), and Egyptian (CallHome/CallFriend collection) Arabic. Unfortunately, the Egyptian data collection preceded the others by about a decade, leading to likely differences in collection protocol and channel properties (e.g., preva-

Table 1: Dialect data used for our experiments.

Dialect	Train		Test		
	Spkrs	Hours	Spkrs	Hours	# 30s cuts
Iraqi	382	32.63	96	3.9	477
Gulf	781	20.53	195	6.7	801
Levantine	788	31.95	197	6.8	818
Egyptian	280	28.34	120	16	1912

lence of cellular phones) that have the potential to confound dialect recognition results. Still, we decided to keep Biadsy's data set as originally defined to allow comparison of results, and will later break out results by dialect.

Speech files were segmented based on silence, using the Praat tool and a silence threshold of -35 dB, with a minimum silence interval of 0.5 seconds and minimum sounding intervals of 0.5 seconds. Similar to the NIST Language recognition evaluation (LRE), conversation excerpts of consecutive speech segments totaling about 30 seconds were used as test samples. Full conversations were used for training. Statistics for training and test data are summarized in Table 1. The test set was balanced for gender (male/female), speakers, and channel (landline/mobile), where possible (channel information was not available for Egyptian). Phone recognition models, as well as JFA for the GMM system, were trained on other datasets, as described below.

3. Method

3.1. Cepstral GMM System

We first developed a GMM system that employed the eigenchannel framework [14] for speaker and channel compensation. We used a 56-dimensional MFCC-based feature vector consisting of energy and a shifted-delta cepstrum in 7-1-3-7 configuration [3]. Deltas were computed over the whole file (including nonspeech regions), and feature frames were removed based on SRI's speech/nonspeech segmenter. We estimated a 2048-component GMM-UBM and a 300-dimensional eigenchannel subspace, both trained on separate training data, that of the LRE07 task as described in [11]. Scores for each test sample were produced using dot product scoring [15].

3.2. Phonotactic Systems

The other state-of-the-art approach for dialect identification is phonotactic modeling, where phone N-gram statistics are extracted from the output of a phone recognizer (PR) and used as input to a classifier. All acoustic models used for phonetic recognition in this work were trained using a perceptual linear prediction (PLP) front end with up to third-order difference features, vocal tract length normalization, dimensionality reduction via heteroscedastic linear discriminant analysis (HLDA), and triphone acoustic models trained using the minimum phone error (MPE) criterion. The decoding process uses an open phone loop (no phonotactic constraints) to generate phone lattices, from which phone N-grams with posterior probability weighted frequencies are extracted. Following [16], we also perform constrained MLLR speaker adaptation prior to decoding.

3.2.1. Acoustic Modeling Choices for Phone Recognition

In previous studies for LID, it has been shown that running multiple phone recognizers in parallel and combining final classification scores at the end leads to improvement over a single phonotactic system. In this work, we use our existing phone recognizers (for Levantine and MSA Arabic) in our baseline phonotactic systems, and then compare them against two non-standard phone recognizers: (1) a multilingual phone recognizer (ML-PR) trained on four fairly diverse languages including only one Arabic dialect (Egyptian), and (2) a "Pan-Arabic" phone recognizer (PA-PR) trained on data from the four target dialects plus MSA.

The multilingual phone recognizer (ML-PR) is a universal phone recognizer developed previously for an LID task [11], where it was found to perform substantially better than three language-specific recognizers. The system is based on a shared set of 52 phones that give a reasonable representation of four fairly diverse languages: American English (trained on 123h native and 108h nonnative speech), Mandarin (103h), Spanish (19h), and Egyptian Arabic (17h), while glossing over fine-grained distinctions in the language-specific phone sets (such as tone in Mandarin). For training purposes, the word-level transcripts in each language were mapped to the multilingual phone set, and the training data was pooled.

The Pan-Arabic phone recognizer (PA-PR) is trained on data from a subset of the target dialects and from MSA, based on available transcribed training data: 17h of Egyptian, 61h of Levantine, 727h of Iraqi, and 1123h of Modern Standard Arabic. The latter is obtained from broadcast data downsampled to 8 kHz to match the telephone data. The available dictionaries for these dialects are mapped to a single phone set, and an attempt was made to homogenize the pronunciation conventions encoded in these dictionaries. For example, extra short vowels and geminates encoded in the original Egyptian dictionary were removed, since pronunciations for the other dialects were only partially vowelized based on linguistic rules applied to the word forms in Arabic script.

3.2.2. Phone N-gram Modeling Choices: PRLM vs. PRSVM

We explored two modeling choices in phonotactic dialect recognition that have been used in LID and speaker recognition systems to model phone N-gram statistics. In the first approach, phone recognition is followed by language-modeling (PRLM), and dialect recognition scores are obtained by computing the length-normalized log likelihood ratio of the target phone language model relative to the nontarget language model (as estimated from the union of all nontarget language data). In the second approach, N-gram phone statistics are modeled via SVMs (PRSVM). The relative frequencies of the most frequent phone N-grams (as determined on the combined training set) are assembled into a feature vector. Each relative frequency is scaled by the inverse square root of the global frequency of that N-gram, so as to implement the term frequency (pseudo) log-likelihood ratio (TFLLR) kernel proposed by [17]. The resulting feature vectors are modeled and scored by linear-kernel SVMs. In both PRLM and PRSVM systems, phone N-gram frequencies are extracted as posterior-probability weighted counts from phone lattices.

3.3. System Calibration and Combination

For all systems, the calibration process is a multiclass logistic regression using Niko Brümmer's FoCal Multiclass toolkit [18]. System combination is performed at the score-level in a similar manner using the same optimization function.

Consistent with the NIST LRE confusable language pair evaluations, and following [13], the primary metric used in our experiments is the average equal error rate (avgEER) of the six

Table 2: Results with different classification models, for 4-way and pairwise (2-way) classification, before and after calibration.

Systems	%avgEER			
	4-way		2-way	
	raw	calibr.	raw	calibr
MSA-PRLM	18.61	18.31	15.88	15.71
LEV-PRLM	11.22	9.69	8.12	7.81
ML-PRLM	12.33	11.26	10.87	9.72
PA-PRLM	11.55	10.06	7.97	7.85
MSA-PRSVM	14.51	14.64	12.75	12.69
LEV-PRSVM	9.55	9.53	7.94	7.87
ML-PRSVM	10.28	10.16	8.03	7.90
PA-PRSVM	11.15	10.72	9.05	8.93
GMM	4.77	4.05	3.53	3.13

pairwise dialect comparisons. However, we also report avgEER in a 4-way (1-versus-3) dialect detection task, as this might be more relevant for applications.

For evaluation purposes, we split the test set into two subsets and performed two-fold cross-validation, i.e., calibration parameters, combination weights, and calibration thresholds are estimated on the complement of the half of the data that is being scored. The results reported here are accumulated from the two folds. The optimization criterion used for calibration and combination is the avgEER for 4-way classification, and we use the same scores to estimate the pairwise results (which are thus sub-optimal).

4. Results

Table 2 summarizes results with various PRLM and PRSVM configurations, and with the GMM system, both with raw and calibrated scores. The GMM system serves as the baseline for comparison and combination with the other systems. The 3.13% avgEER baseline result achieved after calibration for pairwise dialect comparison is significantly better than the cepstral GMM systems reported for the same task in [13] (with the GMM system giving 11%). Since the GMM system was of interest to us only as a baseline, we didn't investigate the reasons for the better result; likely factors are different acoustic features, different UBM training data, and different channel compensation and calibration approaches. However, the best system reported in [13] is a novel SVM-based classifier using phone-specific GMM supervectors as features, yielding 3.96% avgEER.

We can also see from the table that calibration consistently improves results. Even though calibration is optimizing the 4-way classification problem, we see that calibrated scores also perform better for pairwise classification.

We also observe that the LEV-PR models give the best results for both PRLM and PRSVM system. While Pan-Arabic (PA) models were second best for the PRLM case, the ML models performed better in the PRSVM system. The LEV and PA models, which are trained on data relevant to this task, are a better fit to the phone recognition task. The MSA models are the least well matched, since the training data differs acoustically and stylistically from the test data (down-sampled broadcast news compared to telephone conversations), and indeed give the worst performance. All the systems are different enough so that improvements are observed when combined with each other.

Also shown in Table 2, phonotactic SVM modeling can give substantial gains over statistical language models in our dialect identification task. This is consistent with prior studies in speaker and language recognition based on phone N-grams [19, 8, 11]. The only exception is the PA-PRSVM system, when compared to the PA-PRLM system. A possible explanation for this result is that generative modeling might be sufficient when used with matched models such as the PA recognition models.

Table 3: Results after different system combinations.

Systems	%avgEER	
	4-way	2-way
MSA-{PRLM+PRSVM}	14.09	12.06
LEV-{PRLM+PRSVM}	8.47	6.63
ML-{PRLM+PRSVM}	7.08	5.62
PA-{PRLM+PRSVM}	6.69	5.36
{LEV+MSA}-{PRLM+PRSVM}	6.70	5.43
{LEV+MSA+PA}-{PRLM+PRSVM}	4.86	3.93
ALL-{PRLM+PRSVM}	4.70	3.83
GMM+LEV-PRSVM	2.93	2.29
GMM+MSA-PRSVM	3.70	3.19
GMM+ML-PRSVM	2.75	2.30
GMM+PA-PRLM	3.03	2.31
GMM+ALL-{PRLM+PRSVM}	2.47	2.01

We now analyze the system combination results for the different systems shown in Table 3. In the first part of the table, we observe among combinations between different PR systems that the Pan-Arabic PRLM and PRSVM combination brings the highest improvement. It is indeed better than the combination of more diverse systems such as the LEV and MSA PR. However the combination of the three leads to a significant additional gain, and even better results are achieved combining all the PR systems. This final PR-ALL combination is comparable to the performance of the baseline GMM model as shown in Table 2.

In the second part of the Table 3, the cepstral GMM system is used as our baseline and starting point for combinations. Each individual PR system performs much worse than the GMM system, but each of them yields improvement when combined with the GMM. The table only shows the results when combining with the best PR system for each set of models. We observe that combining each of the LEV, PA and ML systems to the cepstral system gives a significant relative error reduction, and the results for these three are all very close. This is surprising for the case of the PR-ML model, which by itself performed worse than the LEV and PA, but it may be explained by the fact that both features and training/modeling paradigms are very different in this system.

Finally we note that combining all the PR systems with the GMM system results in around 40% of the EER of the GMM system alone, as shown in the last row of Table 3.

Figure 1 shows individual DET curves for all dialect pairs, as well as the average DET curve. It is clear that our system does much better recognizing Egyptian compared to the other three dialects, which we can attribute to likely differences in channel or recording characteristics, as discussed earlier.

5. Conclusions and Future Work

Starting from standard cepstral and PRLM systems, we have investigated a number of improved modeling techniques for Arabic dialect recognition, on a four-dialect task for which the best reported average EER is 3.96% [13]. We confirmed results

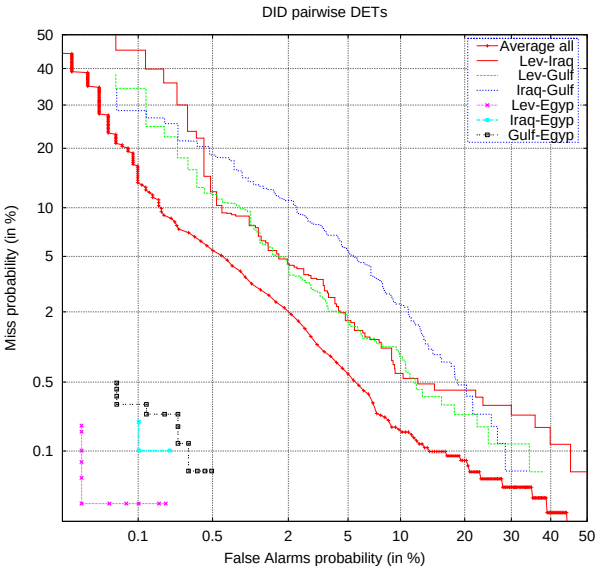

Figure 1: DET curves for pairwise dialect classification.

from language and speaker recognition showing that phonotactic SVM modeling can greatly improve over LM-based approaches, and can be combined with the latter for additional gains. Inspired by prior success in multilingual phonotactic models for language recognition, we explored multidialectal models for tokenization. When combining PRLM and PRSVM scores with the cepstral baseline, such a single multidialectal model gives better results than a single-dialect approach. Surprisingly, even the multilingual phonotactic model makes for excellent combination with the cepstral model, as well as with other phonotactic models. Overall, we find that all modeling parameters provide some complementarity, such that fusing single-dialect, multidialect, and multilingual systems with the baseline is highly effective, lowering average pairwise EER to 2.0%, a 36% reduction relative to the cepstral baseline. The best 4-way avgEER is 2.47%.

Additional single-dialect phonotactic models could probably give improvements, however we find the single multidialect modeling approach more appealing and practical since dialectal training data is typically limited. Based on results here, the multidialectal model can probably be improved, e.g., by excluding MSA training data which seems to be badly mismatched to the data targeted here. Also, the PA phone recognizer could benefit from (labor intensive) standardization of pronunciation encodings in these different dialects, possibly employing automatic full vowelization. Future work also needs to address the question whether similar results hold for dialects in other languages, and whether a generalized recipe for effective dialect recognition can be developed. Additional improvements can be obtained by using other features, such as adaptation transforms estimated by maximum likelihood linear regression (MLLR), which have been shown to give good results for nonnative accent detection [20] and for language recognition [11].

6. Acknowledgments

We thank our SRI colleagues W. Wang, L. Ferrer, H. Bratt, and C. Richey for useful suggestions and valuable comments on the work presented here. Special thanks are due to Fadi Biadsy for providing data set definitions and score files from his own work. This work was supported in part by the Defense Advanced Research Projects Agency, Program Grants No. HR0011-06-1-0003 and HR0011-08-0004. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

7. References

[1] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models", *in Proceedings IEEE Workshop Automatic Speech Recognition and Understanding*, Madonna di Campiglio, Italy, Dec. 2001.

[2] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S. youn Yoon, "Accent detection and speech recognition for Shanghai-accented Mandarin", *in Proc. Interspeech*, pp. 217–220, Lisbon, Sep. 2005.

[3] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features", in J. H. L. Hansen and B. Pellom, editors, *Proc. ICSLP*, pp. 89–92, Denver, Sep. 2002.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified", *in Proc. ICASSP*, vol. 1, pp. 637–640, Philadelphia, Mar. 2005.

[5] N. Brümmer, A. Strasheim, V. Hubeika, P. Matějka, L. Burget, and O. Glembek, "Discriminative acoustic language recognition via channel-compensated GMM statistics", *in Proc. Interspeech*, pp. 2187–2190, Brighton, U.K., Sep. 2009.

[6] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling", *in Proc. ICASSP*, vol. 1, pp. 305–308, Adelaide, Australia, 1994.

[7] P. A. Torres-Carrasquillo, E. Singer, W. M. Campbell, T. Gleason, A. McCree, D. A. Reynolds, F. Richardson, W. Shen, and D. E. Sturim, "The MITLL NIST LRE 2007 language recognition system", *in Proc. Interspeech*, pp. 719–722, Brisbane, Australia, Sep. 2008.

[8] P. Matějka, L. Burget, O. Glembek, P. Schwarz, V. Hubeika, M. Fapšo, T. Mikolov, O. Plchot, and J. Černocký, "BUT language recognition system for NIST 2007 evaluations", *in Proc. Interspeech*, pp. 739–742, Brisbane, Australia, Sep. 2008.

[9] M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech", *in Proc. ICASSP*, vol. 2, pp. 777–780, Atlanta, May 1996.

[10] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based PPRLM language identification", *in Proceedings IEEE Odyssey-06 Speaker and Language Recognition Workshop*, pp. 1–6, San Juan, Puerto Rico, June 2006.

[11] A. Stolcke, M. Akbacak, L. Ferrer, S. Kajarekar, C. Richey, N. Scheffer, and E. Shriberg, "Improving language recognition with multilingual phone recognition and speaker adaptation transforms", *in Proceedings Odyssey Speaker and Language Recognition Workshop*, pp. 256–262, Brno, Czech Republic, June 2010.

[12] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken arabic dialect identification using phonotactic modeling", *in Proceedings EACL Workshop on Computational Approaches to Semitic Languages*, pp. 53–61, Athens, Mar. 2009.

[13] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors", *in Proc. Interspeech*, Florence, Italy, Aug. 2011.

[14] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification", *in Proc. Interspeech*, pp. 1242–1245, Antwerp, Aug. 2007.

[15] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis", *in Proc. ICASSP*, pp. 4057–4060, Taipei, Apr. 2009.

[16] M. F. BenZeghiba, J.-L. Gauvain, and L. Lamel, "Context-dependent phone models and models adaptation for phonotactic language recognition", *in Proc. Interspeech*, pp. 313–316, Brisbane, Australia, Sep. 2008.

[17] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines", in S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pp. 1377–1384, Cambridge, MA, 2004. MIT Press.

[18] N. Brümmer, "Focal multi-class—tools for evaluation, calibration and fusion of, and decision-making with, multi-class statistical pattern recognition scores", http://sites.google.com/site/nikobrummer/focalmulticlass, June 2007.

[19] A. O. Hatch, B. Peskin, and A. Stolcke, "Improved phonetic speaker recognition using lattice decoding", *in Proc. ICASSP*, vol. 1, pp. 169–172, Philadelphia, Mar. 2005.

[20] E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak, "Detecting nonnative speech using speaker recognition approaches", *in Proceedings IEEE Odyssey-08 Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, Jan. 2008.