Analysis of Dialectal Influence in Pan-Arabic ASR

Udhyakumar Nallasamy[1], *Michael Garbus*[1], *Florian Metze*[1], *Qin Jin*[1], *Thomas Schaaf*[2] *and Tanja Schultz*[1]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2]M*Modal Technologies, Pittsburgh, PA, USA

{unallasa, mgarbus, fmetze, qjin, tanja}@cs.cmu.edu, tschaaf@mmodal.com

Abstract

In this paper, we analyze the impact of five Arabic dialects on the front-end and pronunciation dictionary components of an Automatic Speech Recognition (ASR) system. We use ASR's phonetic decision tree as a diagnostic tool to compare the robustness of MFCC and MLP front-ends to dialectal variations in the speech data and found that MLP Bottle-Neck features are less robust to such variations. We also perform a rule-based analysis of the pronunciation dictionary, which enables us to identify dialectal words in the vocabulary and automatically generate pronunciations for unseen words. We show that our technique produces pronunciations with an average phone error rate 9.2%.

Index Terms: automatic speech recognition, dialect analysis, front-end evaluation

1. Introduction

Arabic language is characterized by its multitude of dialects. Although Modern Standard Arabic (MSA) is used in writing, TV/radio broadcasts and for formal communication, all informal communication is typically carried out in one of the regional dialects of Arabic. Dialectal variations influence the pronunciation dictionary, acoustic and language models in an ASR. Previous works on dialectal Arabic ASR include cross-dialectal data sharing [1], improved pronunciation and language modeling [2, 3], etc. In this paper, we describe our experiments on a dialectal Arabic speech database, where we focus on analyzing the behavior of different front-ends and pronunciation dictionary due to dialectal variations between speakers. We evaluate Mel-Frequency Cepstral Coefficients (MFCC) and Multi-Layer Perceptrons (MLP), on their ability to handle these variations that arise due to different dialects. Extending our previous work on gender normalization [4], we use phonetic decision trees as a diagnostic tool to analyze the influence of dialect in the clustered models. We introduce questions pertaining to dialect in addition to context in the building of the decision tree. We then build the tree to cluster the contexts and calculate the number of leaves that belong to branches with dialectal questions. The ratio of such 'dialectal' models to the total model size is used as a measure for dialect normalization. The higher the ratio, the more models are affected by the dialect, hence less normalization and vice versa.

We further extend our analysis to the pronunciation dictionary, where we investigate ways to generate rule-based pronunciations for unseen words in a dialect with minimum manual effort. Our setup features a 'Pan-Arabic' dictionary, which contains pronunciations typically found in five Arabic dialects. We analyze the pronunciation variants in our common dictionary using acoustic model alignments to derive the dialect-specific pronunciations for each word. This forms the source of our rule-learning algorithm which maps word pronunciations from one dialect to another. These rules are then used to generate pronunciations for unseen words and the accuracy is estimated.

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

2. Pan-Arabic Database

All our experiments are carried out on the Pan-Arabic dataset provided by AFRL. The database consists of Arabic speech collected from regional Arabic speakers, corresponding transcriptions and lexicons for 5 different dialects – United Arab Emirates (UAE), Egyptian, Syrian, Palestinian and Iraqi. It is a balanced data set with approximately 50 recording sessions for each dialect, with each session comprising of 2 speakers. The amount of data broken down according to dialect is shown in Table 1 below.

Table 1. *Amount of audio data in the Pan-Arabic database*

Dialect	No. of Hours
UAE (AE)	29.61
Egyptian (EG)	28.49
Syrian (SY)	28.51
Palestinian (PS)	29.29
Iraqi (IQ)	24.92
Total	140.82

Each speaker is recorded in separate channels, including long silences between speaker-turns. Hence the actual conversational speech in the dataset amounts to around 60 hours. The transcriptions of the speech are fully diacritized and included both UTF8 and Buckwalter representations. The first 5 sessions in each dialect are held out and used as test data, while the remaining form the training set. The database also contains dialect-specific pronunciation dictionaries.

All the dialects have a common phone set, except for one minor variation. UAE, Egyptian and Iraqi have the voiced postalveolar affricate, /dZ/ phone. Palestinian and Syrian have the voiced postalveolar fricative, the /Z/ phone instead. These phones are merged into one, while designing the ASR phone set. The final phone set contains 41 phones, including, 6 vowels, 33 consonants in SAMPA representation [5, 6] plus a noise and a silence phone.

3. Baseline ASR

The baseline ASR is trained on speech data from all five dialects, with no dialect adaptation. The individual, dialect-specific dictionaries are merged to form a single ASR dictionary which contains pronunciation variants derived from each dialect. The total vocabulary size is 75046 words with an average of 1.6 pronunciations per word. The language model is a 3-gram model trained on the training transcriptions and Arabic background text, mainly consisting of broadcast news and conversations. The OOV rate of the LM on the test data is 1.8%. The perplexity of LM on the test set is 112.3.

3.1. Acoustic models

We trained two sets of acoustic models based on MFCC and MLP features. For MFCC features, we extract the power

spectrum using an FFT with a 10 ms frame-shift and a 16 ms Hamming window from the 16 kHz audio signal. We compute 13 MFCC features per frame and perform cepstral mean subtraction and variance normalization on a per-speaker basis. To incorporate dynamic features, we concatenate 15 adjacent MFCC frames (±7) and project the 195 dimensional features into a 42-dimensional space using a Linear Discriminant Analysis (LDA) transform. After LDA, we apply a globally pooled ML-trained semi-tied covariance matrix. For the development of our context dependent (CD) acoustic models, we applied an entropy-based, poly-phone decision tree clustering process using context questions of maximum width ±2, resulting in quinphones. The system uses 2000 states with a total of 62K Gaussians with diagonal covariance matrices assigned using merge and split training. The total number of parameters in the acoustic model amounted to 7.8M.

In addition to MFCC system, we trained another set of acoustic models using MLP Bottle-neck features [7, 8, 9]. A multi-layer perceptron is trained using ICSI's QuickNet MLP package [10]. We stack ±7 MFCC frames, which serve as input to the MLP. The context-independent (CI) state labels are used as targets. The MLP has a 4-layer architecture – input (195), 2 intermediate (1000, 42) and output (125) layers, with a total of 243,292 parameters. The training data for the MLP is derived from the ASR training set, 90% of the training speaker list is used for training MLP while the remainder 10% of the speakers is used as a development set. For each training iteration MLP's accuracy on the development set is calculated. The training is stopped when the accuracy saturates on the development set. In our case, MLP training took 5 epochs and reached a frame-level accuracy of 63.86% on the training data and 63.56% on the development data. The activations in the third layer, also called the bottle-neck layer [11] are used as inputs to build GMM-based HMM acoustic models. Apart from MLP parameters, the MFCC and MLP acoustic models used same number of parameters. The baseline Word Error Rate (WER) for the MFCC and MLP system is given in Table 2 below. The WER of MLP ASR system is 0.6% (absolute) lower than the MFCC system. The speaker adapted system produces a WER of 26.8%

Table 2. *Word Error Rate of MFCC and MLP ASR systems*

Dialect	Pan-Arabic ASR (WER)	
	MFCC	MLP
UAE	28.7	28.2
Egyptian	30.0	29.5
Syrian	27.9	27.2
Palestinian	29.4	28.6
Iraqi	27.7	27.0
Average	**28.7%**	**28.1%**

4. Decision Tree based Dialect Analysis

Phonetic decision trees have been traditionally used in ASR to cluster context-dependent acoustic models based on the available training data. The number of leaves in a phonetic decision tree refers to the size of the acoustic model. In our training process, the decision tree building is initialized by cloning the CI models to each available context in the training data. Two iterations of Viterbi training is performed to update the distributions while the codebooks remain tied to their respective CI models. Several phonetic classes of the underlying phones such as voiced/unvoiced, vowels/consonants, rounded/unrounded, etc are presented as questions, to the decision tree algorithm. The algorithm then greedily chooses the best question at each step which maximizes the information gain in a top-down clustering of CD distributions. The clustering is stopped once the desired model size is reached or when the number of training samples in the leaves has reached the minimum threshold [12]. A threshold of 2500 training samples is enforced for all clustered models in the experiments reported in this paper.

In our previous work, we used decision trees as a diagnostic tool to measure the capability of MLP and MFCC features on gender normalization by building decision trees with gender questions, and to measure model sharing in English and Arabic ASR systems [4]. In this paper, we use the decision tree as a diagnostic tool to analyze the influence of dialects in different front-ends. We combine dialectal questions with contextual questions and let the entropy-based search algorithm to choose the best question at each stage. The resulting decision tree will have a combination of dialectal and contextual phonetic questions. The earlier the question is asked, the greater its influence on the ensuing models. For each leaf node, we traverse the tree back to the root node. If we encounter a dialectal question in a node, then that leaf is assigned as a dialectal model. The ratio of dialectal to total leaves is used as an estimate of dialectal influence. The calculation is repeated for different model sizes.

4.1. Analysis of dialectal models

In the first experiment, we examine the influence of dialect in MFCC front-end. Table 3 summarizes the dialectal analysis for different model sizes.

Table 3. *Ratio of dialect nodes in MFCC decision tree*

Model Size	Dialect nodes	Non-Dialect nodes	Ratio	Dialect Nodes	Non-Dialect nodes	Ratio
	MFCC			MFCC (VTLN + FSA)		
1000	13	987	1.3%	9	991	0.9%
2000	82	1918	4.1%	72	1928	3.6%
3000	224	2776	7.5%	226	2774	7.5%
4000	483	3517	12.1%	465	3535	11.6%

We observe that speaker adaptation, including vocal tract length normalization (VTLN) and feature space adaptation (FSA) training, only marginally reduce the influence of dialect (~0.5% absolute) in the acoustic models. In the resulting decision trees, we observe that the /Z/ appears very early in the split. This is the phone we merged from /dZ/ and /Z/ that belongs to two different dialect classes. Dialect questions in the decision tree allowed the phone to split into its dialectal counterparts. The distribution of different dialects for each model size is shown in Figure 2.

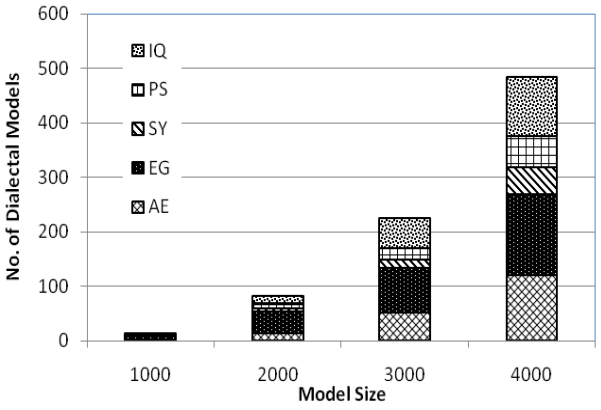

Figure 2: *Distribution of different dialects among the dialectal models*

We noticed that most dialectal models belong to Egyptian across different model sizes. This behavior is consistent with the results found in the literature, where Egyptian is found to be most distinguishable from other dialects [13, 14]. We also observed that vowels are more influenced by dialect than consonants. Table 4 shows the ratio of dialectal models to all clustered models for vowels and consonants. Except for the case of model size 1000, vowels have more dialectal models and hence more dialectal influence, than consonants. This result is in line with the fact that the majority of differences between Arabic dialects are characterized by vowels. These observations indicate that decision trees can be used as an effective analytic tool to evaluate the effect of different dialects in acoustic models.

Table 4: *Ratio of dialectal models for vowels and consonants*

Size	Dialectal models	Ratio of Dialectal models	
		Vowels	Consonants
1000	13	1.1%	1.4%
2000	82	6.2%	2.9%
3000	224	10.8%	5.4%
4000	483	17.1%	8.8%

4.2. MFCC Vs MLP dialect normalization

In this section, we examine the influence of dialect in MLP and MFCC front-ends. The number of dialectal models for MLP and MFCC systems is shown in Figure 3. From the graph, it can be seen that speaker adaptation marginally reduces the influence of dialect in the final models, in both MFCC and MLP. Comparing, the two front-ends, MFCC has less dialectal models than MLP for all cases.

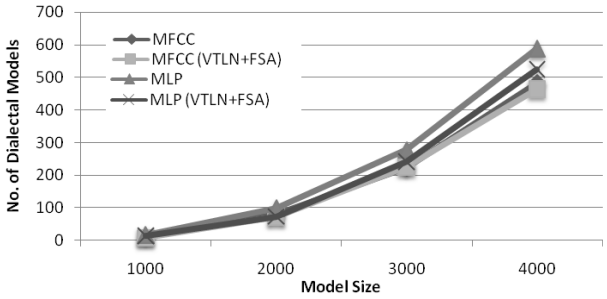

Figure 3: *Dialectal models in MFCC and MLP trees*

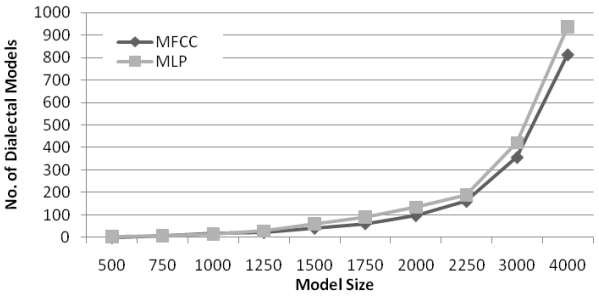

Figure 4: *Dialectal models for MFCC and MLP with single pronunciation dictionary and combination of dialect questions*

To confirm the hypothesis that MLP features are more sensitive to dialect, we created a more rigorous setup. The pilot experiment used a combined dictionary obtained by composing individual, dialect-specific dictionaries. The use of different "dialectal" pronunciation variants can render the models to be insensitive to dialectal variations. Hence, in our next experiment, we constrained the dictionary to have only one pronunciation for each word. The training data is force-aligned with the combined dictionary and the most frequent pronunciation variant is selected for each word, which is the only variant used in the experiment. Also, in the previous experiment only singleton dialect questions (eg. Is current phone IRAQI?) were used. We experimented with combinations of dialect questions in the following setup (eg. Is current phone IRAQI OR EGYPTIAN?). This would allow more dialectal questions to be available for clustering. Figure 4 shows the results of the new setup. It can be observed that more MLP models are influenced with dialect than in the case of MFCC. These results show that MLP features are more sensitive to linguistic variations, i.e. dialect.. This contradicts our findings with respect to gender in this database (Figure 5) and in our previous work [4], where we found that both MLP and speaker adaptation greatly reduce the influence of gender in the clustered models.

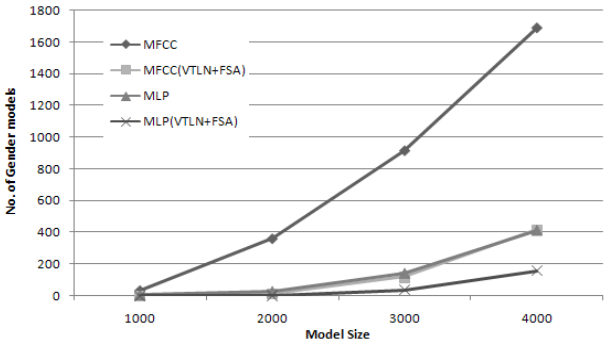

Figure 5: *Gender models for MFCC and MLP with and without speaker adaptation*

To analyze the dialect sensitive behavior of MLP, we calculated the frame-level accuracy of vowels and consonants in the MLP outputs on the development set. The average accuracy for vowels and consonants is shown inTable 5.

Table 5. *MLP frame accuracy for Vowels and Consonants*

Phone class	MLP Frame Accuracy
Vowels	26.41%
Consonants	40.80%
Noise/Silence	85.78%

It is clear from Table 5 that MLP frame level accuracy is higher for vowels than consonants. We already observed that dialectal models are dominated by vowels, which indicates that most dialectal variations occur in vowels. Hence, we suspect that the low MLP frame accuracy for vowels rendered MLP to be more sensitive to dialectal variations.

5. Pronunciation based Dialect Analysis

In this section, we present a technique to derive rules for converting dialect-specific pronunciations from one dialect to another. These automatically learnt pronunciation rules are applied to a source dialect to produce pronunciations for new words in a target dialect. Previous work on automatic pronunciation generation includes using various machine learning techniques on grapheme-to-phoneme tasks [15, 16]. Our approach differs from these methods as we make use of already available pronunciations from a different dialect, and learn the transformation rules from a limited set of parallel pronunciations.

For this experiment, we selected the words in the Pan-Arabic dictionary that had dialect-specific pronunciation variants. To identify these 'dialectal' words, we first restricted our search to words that are very common in source and target dialect. Frequency statistics of the pronunciation variants for these words are obtained from the forced aligned labels, using a trained speech recognizer. An example of the statistics can

be seen in Figure 6, which clearly shows that there are dialectal preferences in pronunciation for this example. Other words may not exhibit such apparent pronunciation preferences between dialects. Hence the Kullback-Leibler (KL) divergence is used to identify words that exhibit dialectal pronunciations. If the KL-Divergence over distribution of pronunciation variants for the source and target dialects is sufficiently large, the word is considered to have a dialectal pronunciation. The specific pronunciation variants that are most used by the source or target are identified using a simple heuristic that involves identifying pronunciation variants that possess greater than some amount of the probability mass.

Once pronunciations that are believed to exhibit dialectal differences are identified, we use the Levenshtein distance algorithm to identify the transformation rules between the source and target pronunciations. When a substitution, deletion, or insertion is detected in the alignment, the source and target phones were recorded. Additionally, the right and left phonetic context were also recorded. The transformation rules were aggregated across all selected words. These edit rules can be used as regular expressions to generate pronunciations in the target dialect given a pronunciation in the source dialect based on the context.

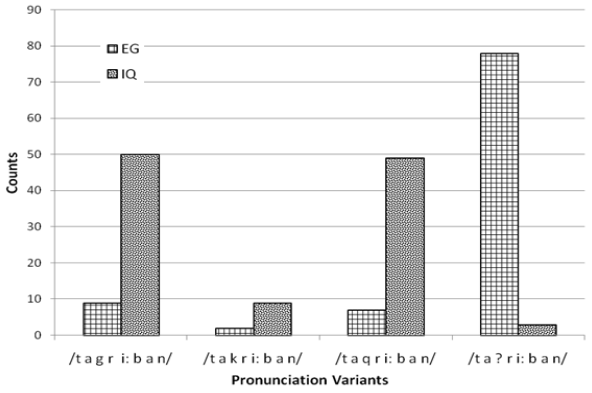

Figure 6: *Pronunciation statistics for "taqoriybaAF" Egyptian dialect prefers variant 4, while the Iraqi dialect prefers the use of variants 1 and 3*

This method was used to regenerate pronunciations for words in the unseen test set for all permutations of dialects. The Nelder-Mead algorithm [17] was used to locally maximize the number of pronunciation variants that agreed with the original pronunciations by optimizing the initial word count threshold, the KL-Divergence threshold, and the proportion of the probability mass used to determine the preferred dialectal pronunciations. For each dialect, around 19 dialect words were chosen to produce the conversion rules, and the rules were applied to an unseen test set of 1047 words across the dialectal pairs. The transformation resulted in an average of 16 variants per word.

Table 6: *Phone Error Rate for Forced Aligned Pronunciation Variants.*

Source / Target	AE	EG	IQ	PS	SY
AE	-	16.2%	9.4%	7.1%	4.3%
EG	12.4%	-	9.8%	14.3%	12.9%
IQ	5.9%	13.6%	-	6.5%	4.8%
PS	8.6%	7.1%	7.5%	-	7.6%
SY	10.4%	10.1%	7.6%	7.6%	-

On an average, this technique created 68% more pronunciation variants than existed in the original Pan-Arabic pronunciation dictionary. To obtain the final pronunciation variant in the target dialect, the newly produced variants were used during a second forced alignment. The pronunciation variant chosen via initial forced-alignment is used as the reference for the target dialect. The final hypothesized variant is compared against its reference to obtain the phone-error rate. The procedure is repeated for each pair of dialects, one being the source and the other the target. Table 6 shows the phone error rate between the reference pronunciation and the final hypothesized variant after second forced-alignment.

Our experiments with the Pan-Arabic corpus showed that our technique produced pronunciations with an average phone error rate 9.2%. This technique uses less training data and can be helpful in situations where we need to extend a small pronunciation dictionary in target dialect, using pronunciation dictionaries from other dialects of the language.

6. Conclusions

In this paper, we compared MFCC and MLP front-ends under the influence of dialect using a phonetic decision tree. While we show that MLP front-end produces lower WER, there is no evidence that MLP 'Bottle-Neck' features are robust to dialectal variations. We also presented a method for generating pronunciations for different dialects of the same language using partially existing information. The preliminary experiments show promising results. We plan to extend this analysis to other languages/dialects. We also plan to examine the behavior of other front-ends including phonetic features under the influence of dialectal variations.

7. References

[1] K. Kirchhoff and D. Vergyri, "Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition", Speech Communication 46, pp. 37-51, 2005.

[2] K. Kirchhoff et al., "Novel approaches to Arabic speech recognition - final report from the JHU summer workshop 2002, Tech. Rep., John-Hopkins University, 2002.

[3] L. Lamel, A. Messaoudi, J-L Gauvain, "Automatic speech-to-text transcription in Arabic", ACM Transactions on Asian Language Information Processing, Vol. 8, No. 4, 2009.

[4] T. Schaaf and F. Metze, "Analysis of gender normalization using MLP and VTLN features", Proc. Interspeech, 2010.

[5] SAMPA for Arabic – phon.ucl.ac.uk/home/sampa/arabic.htm

[6] U. Nallasamy, M. Noamany, T. Schaaf, M. Fuhs, and T. Schultz. Chap. "CMU/InterACT Arabic speech recognition system for GALE", in J. Olive (ed.): GALE Book, 2010.

[7] H. Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems", Proc. ICASSP, 2000.

[8] F. Grézl and P. Fousek, "Optimizing bottle-neck features for LVCSR", Proc. ICASSP, 2008.

[9] F. Metze, R. Hsiao, Q. Jin, U. Nallasamy and T. Schultz, "The 2010 CMU GALE Speech-to-Text System", Proc. Interspeech 2010.

[10] ICSI QuickNet - www.icsi.berkeley.edu/Speech/qn.html

[11] F. Grézl, M. Karafiát and L. Burget. "Investigation into bottleneck features for meeting speech recognition", Proc. Interspeech, 2009.

[12] M. Finke and I. Rogina. "Wide context acoustic modeling in read vs. spontaneous speech", Proc. ICASSP, 1997.

[13] F. Biadsy, N. Habash, and J. Hirschberg, "Improving the Arabic Pronunciation Dictionary for phone and word recognition with linguistically-based pronunciation rules", Proc. NAACL, 2009.

[14] F. Biadsy, H. Soltau, L. Mangu, J. Navratil and J. Hirschberg, "Discriminative Phonotactics for Dialect Recognition Using Context-Dependent Phone Classifiers", Proc. Odyssey 2010.

[15] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules", Proc. ESCA WSS, 1998.

[16] P. Charoenpornsawat and T. Schultz, "Example based grapheme-to-phoneme conversion for Thai", Proc. Interspeech 2006.

[17] J. A. Nelder and R. Mead, "A simplex method for function minimization", Computer Journal 7 (1965), 308-313.