Cross-Dialectal Data Transferring for Gaussian Mixture Model Training in Arabic Speech Recognition

Po-Sen Huang, Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign
Department of Electrical and Computer Engineering
405 North Mathews Avenue, Urbana, IL 61801 USA
huang146@illinois.edu, jhasegaw@illinois.edu

***Abstract*—Dialectal Arabic speech recognition is a difficult problem and is relatively less studied. In this paper, we propose a cross-dialectal Gaussian mixture model training criteria to transfer knowledge from one domain to the other by data sharing. Specifically, phone classification experiments on West Point Modern Standard Arabic Speech corpus and Babylon Levantine Arabic Speech corpus demonstrate that the cross-dialectal training improves phone classification accuracy significantly, especially when a small amount of MSA data is transferred.**

***Keywords*—Transfer learning, Arabic automatic speech recognition, Gaussian Mixture Models, dialectal Arabic**

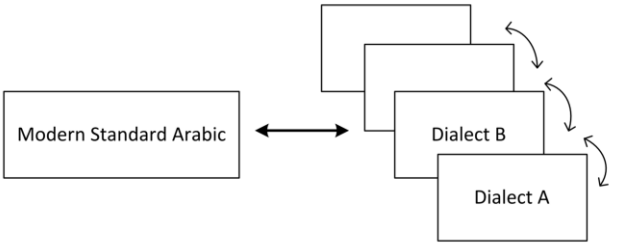


Figure 1. The idea of cross-dialectal transfer learning is to transfer knowledge between Modern Standard Arabic (MSA) and regional dialects, and also to transfer knowledge between different dialects.

I. INTRODUCTION

Dialectal variation is a difficult problem in automatic speech recognition (ASR) [1, 2]. The spoken language is affected from acoustic realization of phones to differences in syntax, vocabulary, and morphology. In some languages, such as English, dialectal variation is relatively small and dialects are mutually intelligible. In other languages, such as Arabic and Chinese, on the other hand, large numbers of dialects are different to the extent that they are not mutually intelligible. Moreover, these dialects are spoken instead of written, that is, the written dialectal material is limited and does not generally follow a writing standard. Furthermore, when people develop ASR systems, training data is acquired by recording and transcribing the data manually, which is very time-consuming. Hence, the limited amount of training data and the unavailability of written transcriptions are serious bottlenecks in dialectal ASR system development.

The Arabic language can be viewed as a family of related languages, with limited vocabulary overlap between dialects, but with a high percent overlap among the phoneme inventories [3]. Training ASR in a regional dialect is difficult because the training data is relatively limited. The idea of cross-dialectal transfer learning is to bridge the knowledge from one dialect to the other, assuming that different dialects still have knowledge in common. As shown in Fig. 1, the knowledge can be transferred between Modern Standard Arabic (MSA) and dialects, and transferred between different dialects.

In this paper, given the fact that some phones are shared across different dialects, we propose a Gaussian Mixture Model (GMM) training method, which learns model parameters by maximizing an optimality criterion composed of the following terms: (1) the maximum likelihood of in-dialect data and their labels, and (2) the maximum likelihood of cross-dialectal data and their labels, weighted by the affinity between in-dialect and cross-dialect acoustic models for phones believed to be similar. For example, if one phone P in dialect A is believed to be similar to phone Q in dialect B, then the model parameters learned in dialect B are penalized for differences between phones P and Q.

The organization of this paper is as follows. Section 2 introduces previous work in Arabic ASR and dialectal Arabic ASR. Section 3 presents the proposed GMM training criteria for transferring knowledge between MSA and Levantine Arabic by data sharing. Section 4 introduces the West Point Modern Standard Arabic Speech and BBN/AUB DARPA Babylon Levantine Arabic Speech corpora used in this paper. Section 5 presents the experiments and results. Finally, section 6 concludes our paper.

II. RELATED WORK

Most previous work on Arabic ASR has focused on developing recognizers for MSA [4, 5]. MSA is a formal linguistic standard used throughout the Arabic-speaking world, and is employed in the lectures, broadcast news,

etc. However, everyday communication is conducted in regional dialects, of which there are four main types: Egyptian, Levantine, North African and Gulf Arabic.

Relatively few works have concentrated on dialectal Arabic. Kirchhoff et al. [6] reported the results in dialectal Arabic speech recognition from the 2002 John Hopkins Summer Workshop. They proposed automatic romanization for vowel restoration, morphology-based language modeling, and using out-of-corpus language model data. Kirchhoff and Vergyri proposed the idea of cross-dialectal data sharing using automatic diacritization for MSA and Egyptian Colloquial Arabic (ECA) [1]. Elmahdy et. al. proposed using model adaptation techniques like Maximum Likelihood Linear Regression (MLLR) and Maximum A-Posteriori (MAP) to adapt existing phonemic MSA acoustic models with a small amount of dialectal ECA speech data [2]. Our approach, on the other hand, assumes the diacritization information is given and focuses on the cross-dialectal GMM training using data from MSA and Levantine Arabic.

III. CROSS-DIALECTAL TRAINING OBJECTIVE

We first formulate our problem setting. Suppose we are given a set of points $X_{D_p} = \{x_1, \ldots, x_n\}$ with labels $Y_{D_p} = \{y_1, \ldots, y_n\}$ and another set of points $X_{D_q} = \{x_{n+1}, \ldots, x_{n+k}\}$ with labels $Y_{D_q} = \{y_{n+1}, \ldots, y_{n+k}\}$, where D_p and D_q represent two different dialect data. In our case, $x_i \in \mathbb{R}^n$ is the n-dimensional spectral feature vector with a phone occurrence i, and $y_i \in \{1, \ldots, C\}$ is the phone class label, assuming that there are totally C phonetic classes.

The classification rule $f : \mathbb{R}^n \to \{1 \ldots C\}$ is based on Bayes rule,

$$\hat{y} = f(x) = \max_{y \in \{1 \ldots C\}} p(x \mid y) p(y),$$

where $p(y)$ is the phone class prior estimated from the training set, and the conditional distribution $p(x \mid y)$, $y \in \{1 \ldots C\}$ is modeled using a Gaussian mixture model,

$$p(x \mid y) = \sum_{i=1}^{M} w_i N(x; \mu_i, \Sigma_i)$$

where w_i is the weight for component i and satisfies the constraints $\sum_{i=1}^{M} w_i = 1$, $w_i \geq 0, i = 1, \ldots, M$.

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation $\lambda = \{w_i, \mu_i, \Sigma_i\}$, $i = 1, \ldots, M$. For classification, each class is represented by a GMM parameterized by λ. Our goal is to learn GMM parameters λ for phone classification.

We propose estimating class c GMM parameters λ_c based on maximizing the following objective,

$$\begin{aligned} J &= L(\lambda_c \mid X_{D_p,c}) + \alpha \sum_{j=1}^{C} P(Y_{D_p,c} = Y_{D_q,j}) L(\lambda_c \mid X_{D_q,j}) \\ &= \sum_{i=1}^{n} \log p_{\lambda_c}(x_i \mid y_i = c) + \alpha \sum_{i=n+1}^{n+k} p(y_i = c) \log p_{\lambda_c}(x_i \mid y_i) \end{aligned}$$

where the function $L(\lambda \mid X)$ is the likelihood function and the term $P(Y_{D_p,c} = Y_{D_q,j})$ is the similarity measure between phoneme c and j in dialect D_p and D_q, respectively. For simplicity, we set parameter $\alpha = 1$ and the similarity measure as a delta function, that is, if there exists a phoneme c in D_p and D_q, then the probability is one; otherwise, the probability is zero. Maximum likelihood model parameters are estimated using the Expectation-Maximization (EM) algorithm [7].

IV. CORPORA

In this section, we introduce the West Point Modern Standard Arabic Speech corpus [4, 8] and BBN/AUB DARPA Babylon Levantine Arabic Speech corpus [9, 10], used in our experiments for transfer learning.

A. West Point Modern Standard Arabic corpus

The West Point Modern Standard Arabic corpus, provided by Language Data Consortium (LDC) [8], consists of collections of four main Arabic scripts, Collection Script 1, 2, 3, and 4. There are 8,516 speech files, totaling 11.42 hours or 1.7 GB of speech data. Each speech file is recorded by one subject reciting one prompt from one of four prompt scripts. Approximately 7,200 files are from native speakers and 1,200 files are from nonnative speakers. There are totally 1,131 distinct Arabic words. All scripts were written with MSA and were diacritized. As given by the LDC West Point catalog, the system was designed by considering all 37 MSA phones, which has three more phonemes than the number of MSA phonemes mentioned in linguistic literature [4, 11], that is, /g/ ``voiced velar stop", /aw/ ``back upgliding diphthong", and /ey/ ``upper mid front vowel."

B. BBN/AUB DARPA Babylon Levantine Arabic Speech corpus

Levantine Arabic is an Arabic dialect of people in Lebanon, Jordan, Palestine, and Syria. It is a spoken language instead of a written language. There are different word pronunciations and different words from MSA. The corpus was developed with funding from DARPA, as part of the Babylon program. Approximately 80\% of the corpus was recorded by the students and staff of American University of Beirut (AUB), and remaining 20\% was recorded by paid subjects recruited in the Boston area. The BBN/AUB dataset consists of 164 speakers, 101 males and 63 females. It is a set of

spontaneous speech sentences, recorded in Levantine colloquial Arabic. The duration of recorded speech is 45 hours distributed among 79,500 audio clips [12].

V. EXPERIMENTS

A. Experiment Setting

In our experiments, since there is no ground truth transcription at the phone level, we first use forced alignment to generate phone boundary information. Then, for each phone occurrence, frame-level spectral features are calculated (12 dimensional PLP coefficients) with 25 ms Hamming window and 10 ms frame rate. Finally, we use segmental features to represent frame-level PLP features [13]. Specifically, frame-level PLP features of each phone segment are divided into three regions with a 3-4-3 ratio, plus two 40 ms regions centered at the start and end time of the segment. Each phone occurrence is represented by the five averages plus the log duration of the phone, totally a 61 (12 × 5+1) dimensional measurement vector.

For classification task, we omit the glottal stop phone in both corpuses, since it is not included in the standard TIMIT phone classification task [14]. There are 36 phone classes in West Point Modern Standard Arabic Speech corpus and 38 phone classes in BBN/AUB DARPA Babylon Levantine Arabic Speech corpus. For Levantine corpus, we randomly divide 60% of the data as training set, 10% of the data as development set, and the remaining 30% of the data as test set. We use the whole set of MSA data for transfer learning. We use Gaussian mixture models with diagonal covariance matrices and select mixture number according to the development set.

B. Transfer MSA data to Levantine Arabic

MSA is a formal language and is easier to be acquired from broadcast news or lectures, and dialectal Arabic speech is more difficult to obtain. We examine our proposed cross-dialectal GMM training technique for transferring MSA data to Levantine Arabic model. In our experiments, we examine different percentages of MSA data and Levantine Arabic data.

As shown in Figure 2 and 3, we compare the classification results of transferring different amounts (0% (*Levantine only* case), 2%, 4%, 6%, 8%, 10%) of MSA data to Levantine data, and compare the classification results using different amounts (10%, 20%, 40%, 60%, 80%, 100%) of Levantine training data.

From the experimental results, we can observe that the classification accuracy is relatively low, compared with the standard TIMIT phone classification task [14]. One possible reason is that the forced aligned phone boundaries are not precise. Hence, the feature of a phone occurrence might also include features from its adjacent phones. Furthermore, from Figure 3, we can observe that the classification accuracy is relevant to the ratio between the length of Levantine training data and MSA data. The results suggest that when proper ratio of MSA data is transferred to Levantine data, we enhance the phonetic coverage of GMMs and achieve higher accuracies.

Figure 2. Phone classification accuracy using different percentages of MSA data and Levantine training data. The horizontal axis represents different percentages of Levantine training data. The *Levantine only* case is the baseline.

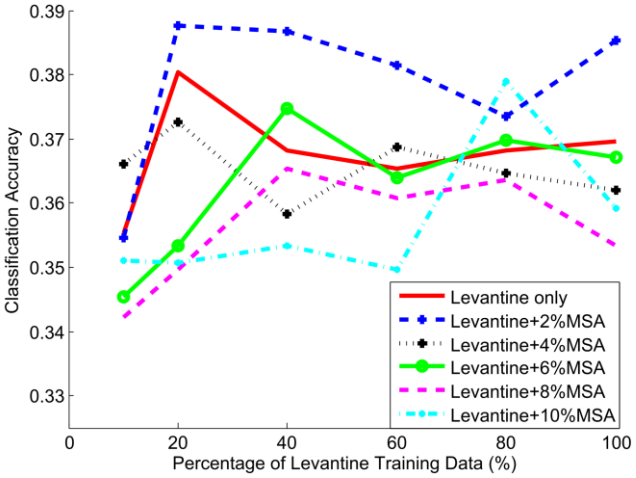
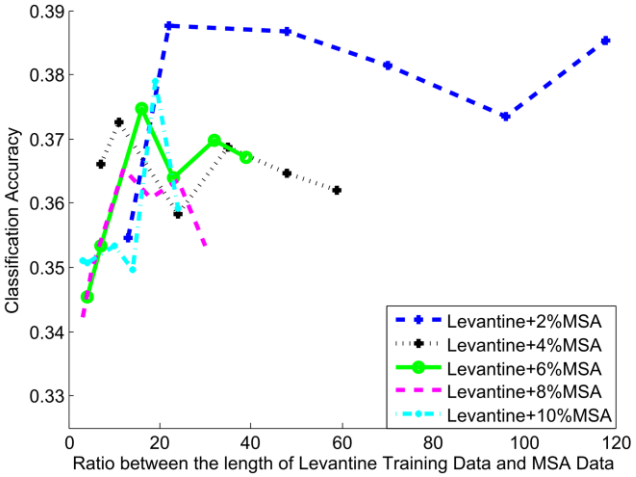


Figure 3. Phone classification accuracy using different percentages of MSA data and Levantine training data. The horizontal axis represents the ratio between the length of Levantine training data and MSA data.



Moreover, when a small amount of data is transferred (2% of MSA data), it achieves the best accuracies.

VI. CONCLUSION

In this paper, we present a cross-dialectal GMM training scheme for West Point Modern Standard Arabic Speech corpus and Babylon Levantine Arabic corpus for phone classification task. From our experiments, we demonstrate that the cross-dialectal GMM training helps phone classification tasks significantly, especially when a small amount of MSA data is transferred.

There are many possibilities to extend current work. For example, we can extend current phone classification tasks to word recognition tasks. Also, we can extend current cross-dialectal maximum likelihood training to discriminative training using other objective functions

such as maximum mutual information and conditional entropy minimization [14]. Moreover, the idea of cross-dialectal transfer learning can be further extended from data sharing to other contexts. For example, instead of transferring cross-dialectal data directly, we can do domain adaptation and transfer the adapted features to target domain/dialects [15]. Besides, we can transfer knowledge at different levels such as model parameter space [16].

VII. ACKNOWLEDGMENT

This research was supported by Qatar National Research Fund under grant number QNRF NPRP 410-1-069. We thank Jui-Ting Huang for helpful discussions.

VIII. REFERENCES

[1] K. Kirchhoff and D. Vergyri, "Cross-dialectal acoustic data sharing for Arabic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004, vol. 1, pp. 765–768.

[2] Mohamed Elmahdy, Rainer Gruhn, and Wolfgang Minker, *Novel Techniques for Dialectal Arabic Speech Recognition*, Springer, 2012.

[3] Mary Catherine Bateson, *Arabic Language Handbook*, Georgetown University Press, 2003.

[4] Yousef Ajami Alotaibi, Sid-Ahmed Selouani, and Douglas O'Shaughnessy, "Experiments on automatic recognition of nonnative Arabic speech," *EURASIP J. Audio Speech Music Process.*, vol. 2008, pp. 1:1–1:9, January 2008.

[5] S.-A. Selouani and Y.A. Alotaibi, "Investigating automatic recognition of non-native Arabic speech," in *International Conference on Innovations in Information Technology*, Nov. 2007, pp. 451–455.

[6] K. Kirchhoff, J. Bilmes, S. Das, N. Duta, M. Egan, Gang Ji, Feng He, J. Henderson, Daben Liu, M. Noamany, P. Schone, R. Schwartz, and D. Vergyri, "Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003, vol. 1, pp. I–344– I–347.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[8] Linguistic Data Consortium (LDC) Catalog Number LDC2005S02, "West Point Arabic Speech," 2002.

[9] Linguistic Data Consortium (LDC) Catalog Number LDC2005S08, "BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts," 2005.

[10] Y.A. Alotaibi and A.H. Meftah, "Comparative evaluation of two Arabic speech corpora," in *International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Aug. 2010, pp. 1–5.

[11] M. Elshafei Ahmad, "Toward an Arabic text-to-speech system," *Arabian Journal for Science and Engineering AJSE*, vol. 16, no. 4B, pp. 565–583, 1991.

[12] Mansour Alsulaiman, Youssef Alotaibi, Muhammad Ghulam, Mohamed A. Bencherif, and Awais Mahmoud, "Arabic Speaker Recognition: Babylon Levantine Subset Case Study," *Journal of Computer Science*, vol. 6, pp. 381–385, 2010.

[13] A. K. Halberstadt, *Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition*, Ph.D. thesis, MIT, 1998.

[14] Jui-Ting Huang and Mark Hasegawa-Johnson, "Semisupervised training of Gaussian mixture models by conditional entropy minimization," in *Interspeech*, 2010, pp. 1353–1356.

[15] Xin Lei, Wen Wang, and A. Stolcke, "Unsupervised domain adaptation with multiple acoustic models," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2010, pp. 247–252.

[16] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.