Development of a TV Broadcasts Speech Recognition System for Qatari Arabic

Mohamed Elmahdy[⋆], Mark Hasegawa-Johnson[†], Eiman Mustafawi[⋆]

[⋆]Qatar University, Doha, Qatar
[†]University of Illinois at Urbana-Champaign, USA
melmahdy@ieee.org, jhasegawg@illinois.edu, eimanmust@qu.edu.qa

Abstract

A major problem with dialectal Arabic speech recognition is due to the sparsity of speech resources. In this paper, a transfer learning framework is proposed to jointly use a large amount of Modern Standard Arabic (MSA) data and little amount of dialectal Arabic data to improve acoustic and language modeling. The Qatari Arabic (QA) dialect has been chosen as a typical example for an under-resourced Arabic dialect. A wide-band speech corpus has been collected and transcribed from several Qatari TV series and talk-show programs. A large vocabulary speech recognition baseline system was built using the QA corpus. The proposed MSA-based transfer learning technique was performed by applying orthographic normalization, phone mapping, data pooling, acoustic model adaptation, and system combination. The proposed approach can achieve more than 28% relative reduction in WER.

Keywords: dialectal Arabic, transfer learning, speech recognition

1. Introduction

Arabic language is the largest still living Semitic language in terms of the number of speakers. More than 300 million people use Arabic as their first native language, and it is the sixth most widely used language based on the number of first language speakers.

Modern Standard Arabic (MSA) is currently considered the formal Arabic variety across all Arabic speakers. MSA is used in news broadcasts, newspapers, formal speech, books, movie's subtitling, and whenever the target audience or readers come from different nationalities. Practically, MSA is not the natural spoken language for native Arabic speakers. MSA is always a second language for all Arabic speakers. In fact, dialectal (or colloquial) Arabic is the naturally spoken variety of Arabic in everyday life communications.

A significant problem in Arabic automatic speech recognition (ASR) is the existence of many different Arabic dialects (Egyptian, Levantine, Iraqi, Gulf, etc.). Every country has its own dialect and usually there exist different dialects within the same country. Moreover, the different Arabic dialects are only spoken and not formally written, and significant phonological, morphological, syntactic, and lexical differences exist between the dialects and the standard form. This situation is called Diglossia (Ferguson, 1959).

Because of the diglossic nature of dialectal Arabic, relatively little research has been done in dialectal Arabic ASR, or in the use of dialect, in any natural language processing tasks. For MSA, on the other hand, much research has been conducted. The limited research done for dialectal Arabic ASR is also due to the sparsity of dialectal speech resources for training different ASR models.

To tackle the problem of data sparsity, in (Kirchhoff and Vergyri, 2005), they proposed a cross-lingual approach where a pooled MSA and dialectal speech data were jointly used to train the acoustic model. This approach resulted in around 3% relative reduction in WER.

Similarly, in (Huang and Hasegawa-Johnson, 2012), a joint cross-lingual training method based on the similarity between phonemes in MSA and dialectal speech data also showed improvements in phone classification tasks.

In (Elmahdy et al., 2010), another cross-lingual approach based on acoustic model adaptation was proposed, which resulted in about 12% relative reduction in WER. Acoustic model adaptation can perform better than data pooling when dialectal speech data are very limited compared to existing MSA data, and adaptation may avoid dialectal acoustic features masking by large MSA data as in the data pooling approach.

In the DARPA GALE project (Mangu et al., 2011), the acoustic model was trained using a large amount of speech data collected from various news channels. Evaluation was performed on news speech and conversational speech. Conversational speech is mostly spontaneous and includes a significant percentage of dialectal Arabic as well as MSA. However, the system was not evaluated or adapted with a specific under-resourced Arabic dialect. Moreover, most of the conversational data in the GALE project are coming from new broadcasts, and it was noticed that the majority of speakers tend to speak in MSA rather than in their own Arabic dialect.

In this paper, The Qatari Arabic dialect has been chosen as a typical example for an under-resourced Arabic dialect. QA is the Arabic dialect spoken in Qatar, and it is a sub-variety of the Gulf dialect. Despite the huge differences between QA and MSA, it is possible to benefit from large existing MSA speech and text resources. In the proposed framework, MSA data and QA data are jointly used in training improved acoustic and language models for QA.

Since transcription conventions may be different between MSA and dialectal Arabic, phone mapping rules across MSA and dialectal Arabic are applied. In addition, we propose data pooling followed by acoustic model adaptation for cross-lingual acoustic modeling and interpolation for cross-lingual language modeling.

Our assumption is that the contribution of limited dialectal speech data in a pooled acoustic model depends on the ratio between MSA data and dialectal data. Usually, there are far more data available in MSA than in the dialect. Thus, it is expected to have little contribution of dialectal data to the final pooled acoustic model. In order to boost the weight of dialectal features, acoustic model adaptation techniques are applied on the pooled acoustic model using dialectal speech data.

All experiments have been conducted with QA in the domain of TV broadcasts. The remainder of this paper is organized as follows: Section 2 introduces the MSA and QA speech corpora. Section 3 and 4 present the speech recognition system and the baseline approach, respectively. The proposed cross-lingual language modeling and acoustic modeling are discussed in Section 5 and 6, respectively. Section 7 discusses the experimental results. Section 8 concludes this study.

2. Speech Corpora

2.1. Modern Standard Arabic

The MSA corpus has been collected from the domain of news broadcast. The corpus consists of two speech resources from the European Language Resources Association (ELRA)[1]. All resources are recorded in linear PCM format, 16 kHz, and 16 bit. The ELRA speech resources are:

- The NEMLAR Broadcast News Speech Corpus, which consists of about 40 hours from different radio stations: Medi1, Radio Orient, Radio Monte Carlo, and Radio Television Maroc.
- The NetDC Arabic Broadcast News Speech Corpus, which contains about 22.5 hours recorded from Radio Orient.

Detailed composition of the resources is shown in Table 1.

Source	Duration (hrs)
Radio Orient	34.6
Medi1	9.5
Radio Monte Carlo	9.0
Radio Tele. Maroc	9.3
Total	62.4

Table 1: Composition of the MSA speech corpus.

2.2. Qatari Arabic Corpus

The QA corpus[2] has been collected from different TV series and talk show programs. Data are selected from programs in which the majority of speech segments is in QA; segments from each program are selected after audition confirms the quality of the speech signal. The programs are: Tesaneef (popular Qatari series with almost 100% in QA), Sabah El-Doha (talk show with almost 80% in QA), and some episodes from Al-Jazeerah are selected if guest speakers are speaking Qatari dialect. The corpus is recorded in linear PCM, 16 kHz, and 16 bits. The overall length is 15 hours. Detailed composition is shown in Table 2. Unlike prior work, as in (Elmahdy et al., 2011; Kilany et al., 2002), where transcriptions were performed manually using Latin orthography, in this corpus, transcription is performed manually in traditional Arabic orthography. Five more Persian letters are used to indicate non-standard Arabic consonants. چ denotes the /tʃ/ consonant, گ denotes /g/, ڤ denotes /v/, ژ denotes /ʒ/, and پ denotes /p/. Some diacritic marks are added for ambiguous words. The following non-speech filler tags are transcribed: pause, breath, laugh, ah, noise, and music. Speech segmentation is done with a 10 second maximum for each segment delimited by filler tags. The QA corpus is divided into a training set of 13 hours, a development set of 1 hour, and an evaluation set of 2 hours. The training set is used either to train the QA baseline acoustic model or to adapt existing MSA acoustic model.

Source	Duration (hrs)
Tesaneef series	9.3
Sabah El-Doha talk show	2.0
Al-Jazeerah programs	3.7
Total	15.0

Table 2: Composition of the QCA speech corpus.

[1]http://www.elra.info/

[2]The QA corpus can be downloaded from: http://sprosig.isle.illinois.edu/corpora/1

3. System Description

The ASR system is a GMM-HMM architecture based on Kaldi speech recognition engine (Povey et al., 2011). Acoustic models are all fully continuous density context-dependent tri-phones with three states per HMM trained with Maximum Mutual Information Estimation (MMIE). The feature vector consists of the standard 39-dimensional MFCC coefficients. During acoustic model training, linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) are applied to reduce dimensionality, which improves accuracy as well as recognition speed. Feature-space MLLR (fMLLR) was used for Speaker Adaptive Training (SAT) of the acoustic models. The first decoding pass uses a relatively smaller language model of around 800K n-grams. Then in the second pass, the generated trigram lattices are rescored against a larger trigram model of around 10M n-grams.

4. Baseline System

4.1. Acoustic Modeling

Grapheme-based acoustic modeling (also known as graphemic modeling) is adopted. Graphemic modeling is an acoustic modeling approach where the phonetic transcription is approximated to be the word graphemes rather than the exact phoneme sequence. Short vowels and geminations are assumed to be implicitly modeled in the acoustic model (Vergyri et al., 2005; Billa et al., 2002).

The baseline acoustic model is trained with the QA training set. The optimized number of tied-states and Gaussians mixture per state are found to be 1000 and 8, respectively. Each grapheme letter is mapped to a unique model resulting in a total number of 41 base units (36 letters in the standard Arabic alphabet and 5 Persian letters).

4.2. Language Modeling

The language model is a backoff tri-gram model with Modified Kneser-Ney smoothing. The baseline language model has been trained with the transcriptions of the QA training set (65K words). The vocabulary size is about 15.5K unique words. LM training parameters have been optimized to minimize the perplexity of the QA development set.

The evaluation of the language model against the transcriptions of the evaluation set results in an OOV rate of 22.2% and a perplexity of 315.5 whilst on the development set, it resulted in an OOV rate of 18.4% and a perplexity of 399.4 the as shown in Table 4. We could not observe any improvement in speech recognition accuracy by increasing the order to 4-grams, apparently because of the limited amount of QA training text that can result in more sparsity in higher order n-grams.

4.3. Evaluation Settings

For the QA baseline system, batch decoding resulted in WER of 61.7% on the QA development set and 80.8% on the evaluation set as shown in Table 3. By examining results, it was found that about 1.0% of the errors are caused by either: the different forms of Alef (e.g. أ instead of ا), final Teh Marbuta (ة instead of ه or vice versa), or final Alef Maksura (ى instead of ي or vice versa). Since there is no standard orthographic form for dialectal Arabic and these kinds of errors are already common orthographic variants in dialectal Arabic, it was decided to ignore these types of errors by normalizing both hypothesis and reference, before alignment, as follows:

- Normalizing all forms of Alef (آ إ أ ا) to ا .
- Normalizing final Yeh ي to Alef Maksura ى .
- Normalizing Teh Marbuta ة to Heh ه .

After applying orthographic normalization, absolute WER decreases to 60.9% on the dev. set with 1.3% relative reduction and 79.9% on the eval. set with 1.1% relative reduction as shown in Table 3.

	dev.	eval.
QA baseline	61.7%	80.8%
+Orthographic norm.	60.9%	79.9%

Table 3: Word Error Rate (WER) (%) evaluation of the QA baseline system with and without orthographic normalization on the development set and the evaluation set.

5. Cross-Lingual Language Modeling

In the baseline system, a significant percentage of errors is mainly due to the high OOV rate that exceeds 18%. In an attempt to improve the LM, a MSA tri-gram LM is trained using the LDC Gigaword corpus (Parker et al., 2009) that consists of more than 800M words. The MSA vocabulary consists of the top 256K words in the corpus. The evaluation of the MSA LM resulted in a perplexity of 1366.7 and 1199.2 on the dev. and eval. sets respectively as shown in Table 4. The OOV rate was found to be 22.3% and 22.1% on the dev. and eval. sets respectively as shown in Table 4. In order to decrease OOV, the QA LM was linearly interpolated with the MSA LM. Interpolation weights were optimized on the dev. set. The cross-lingual interpolation resulted in a vocabulary size of 265.7K words. OOV rate is significantly decreased to 8.9% and 9.2% on the dev. and eval. sets respectively as shown in Table 4. Perplexity test resulted in 1147.0 and 1262.7 on the dev. and eval. sets respectively. In Figure 1, a block diagram for the proposed cross-lingual language modeling approach is shown. Using the cross-lingual MSA/QA LM, batch decoding resulted in absolute WER of 56.0% and 64.4% on the dev. and eval. sets respectively with significant relative reduction of 3.6% and 16.3% compared to the baseline as shown in Table 4.

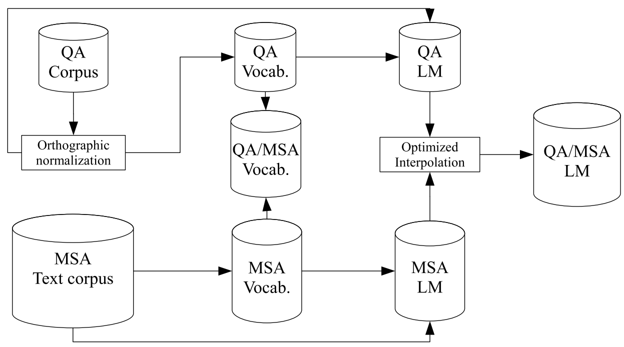

Figure 1: A Block diagram for the proposed cross-lingual language modeling approach.

LM	Vocab.	Perp.		OOV(%)	
		dev.	eval.	dev.	eval.
QA	15.5K	399.4	315.5	18.4	22.2
MSA	256K	1366.7	1199.2	22.3	22.1
QA/MSA	265.7K	1147.0	1262.7	8.9	9.2

Table 4: Language models evaluation with development set and evaluation set.

6. Cross-Lingual Acoustic Modeling

6.1. MSA Acoustic Model

In this section, a MSA acoustic model is used to decode QA speech data. Initially, that is not possible because of the mismatch between the phone sets of MSA and QA. This mismatch is solved by applying phone mapping. Consonants that do not exist in MSA have been mapped to the closest ones in MSA as follows:

- /g/ and /ʒ/ are mapped to /ʤ/.

- /tʃ/ is mapped to /t/ followed by /ʃ/.
- /v/ is mapped to /f/.
- /p/ is mapped to /b/.

After applying QA phone mapping, an MSA graphemic acoustic model is trained using the MSA 62.4 hours corpus. Decoding results are an absolute WER of 61.9% and 81.3% on the dev. and eval. sets respectively with 1.6% and 1.8% relative increase compared to the QA baseline as shown in Table 5. This relative increase is expected as the MSA acoustic model does not yet cover all QA dialect specific features.

6.2. Data Pooling

In data pooling acoustic modeling, the acoustic model is jointly trained using both QA and MSA data. Decoding results are an absolute WER of 56.6% and 64.4% on the dev. and eval. sets respectively outperforming the baseline by a relative decrease of 7.1% and 19.4% as shown in Table 5.

6.3. Acoustic Model Adaptation

In this section, state-of-the-art acoustic model adaptation techniques are applied on the MSA model using QA speech Data. Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995) followed by Maximum A-Posteriori (MAP) re-estimation (Lee and Gauvain, 1993) is applied. Decoding results are an absolute WER of 57.3% and 65.9% on the dev. and eval. sets respectively outperforming the baseline by a relative decrease of 5.9% and 17.5% as shown in Table 5.

6.4. Combined Data Pooling and Acoustic Model Adaptation

Data pooling and acoustic model adaptation have been combined in this section. Acoustic model adaptation is applied on the MSA/QA pooled model rather than the MSA model. Decoding results are an absolute WER of 55.6% and 62.5% on the dev. and eval. sets respectively outperforming the baseline by a significant relative decrease of 8.7% and 21.8% as shown in Table 5.

6.5. System Combination

In this section, different systems are combined to further improve the accuracy using Minimum Bayes-Risk (MBR) decoding (Goel and Byrne, 2000). As shown in Figure 2, MBR is applied on the generated lattices from the two systems: :

1. QA AM (sys. 1 in Table 5).
2. QA/MSA pool/adapt AM. (sys. 5 in Table 5).

In both systems, the QA/MSA interpolated LM is used. System combination using lattice MBR resulted in an absolute WER of 47.9% and 56.8% on the dev. and eval. sets respectively outperforming the baseline system by a relative decrease of 21.3% and 28.9% as shown in Table 5.

The strategy of data pooling, followed by MLLR+MAP adaptation, is equivalent to a type of iterative transformation and adaptive re-weighting of the QCA relative to the

sys.	AM	dev.	eval.
1	QA	58.7	66.9
2	MSA	61.9	81.3
3	QA/MSA pool	56.0	64.4
4	QA/MSA adapt	57.3	65.9
5	QA/MSA pool adapt	55.6	62.5
6	1+5 MBR	47.9	56.8

Table 5: WER (%) on QA dev. and eval. sets using QA/MSA LM and various acoustic models configurations.

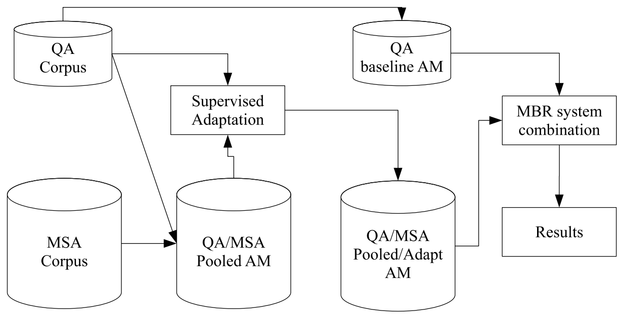

Figure 2: A Block diagram for the proposed cross-lingual acoustic modeling approach.

MSA data. For example, the mean vector of the k^{th} Gaussian, computed by the final stage of MAP adaptation, is given by

$$\mu_k = \frac{\sum_{t=1}^{T} \gamma_t(k) x_t + \tau A_k \bar{\mu}_k}{\sum_{t=1}^{T} \gamma_t(k) + \tau}, \tag{1}$$

where x_t, $1 \leq t \leq T$, is a dialectal feature vector, $\gamma_t(k)$ is the posterior probability of the k^{th} Gaussian given x_t, τ is the weight of the prior, $\bar{\mu}_k$ is the k^{th} mean prior to adaptation, and A_k is the corresponding MLLR transformation. But notice that $\bar{\mu}_k$, in turn, is given by

$$\bar{\mu}_k = \frac{1}{N_k} \sum_{t=1}^{T+S} \bar{\gamma}_t(k) x_t, \quad N_k = \sum_{t=1}^{T+S} \bar{\gamma}_t(k) x_t, \tag{2}$$

where x_t, for $T+1 \leq t \leq T+S$, is an MSA feature vector, and $\bar{\gamma}_t(k)$ is the weighting coefficient computed during the last round of maximum-likelihood EM training applied to the pooled MSA and QCA datasets. By combining Eq. (1) and (2), we discover that MAP adaptation is similar to an adaptive re-weighting scheme, such that QCA feature vectors are weighted comparably to MSA feature vectors during the initial EM training, then transformed by A_k, and then re-weighted to an increased final weight of $N_k \gamma_t(k) + \tau \bar{\gamma}_t(k)$. The effective weight of each MSA datum is similarly decreased, during MAP adaptation, to only $\tau \bar{\gamma}_t(k)$. The effect of this iterative strategy is to give greater weight to MSA data during the initial training of the model, when the MSA data may be useful to help the learning algorithm avoid spurious local optima in the likelihood function; after the model parameters have converged to a solution that is optimal for the pooled MSA+QCA data, then

MLLR improves the representation of QCA data, and, finally, MAP is used to increase the relative importance of QCA data in the final training criterion.

7. Discussion

Even though the differences between MSA and Arabic dialects are large, to the extent that we can consider Arabic dialects as totally different languages (Ferguson, 1959), we can still benefit from MSA speech resources to improve dialectal Arabic speech recognition. The performance of the data pooling approach may be affected by the ratio of dialectal data amount to MSA data amount. In our case, the data pooling approach results in an absolute WER of 56.0% on dev. set and 64.4% on eval. set. MSA data amount is about five times the amount of dialectal data. In order to boost the contribution of dialectal data, MLLR and MAP adaptations are then applied on the pooled acoustic model, effectively increasing the weight of dialectal acoustic features in the final cross-lingual model. The combination of data pooling followed by acoustic model adaptation results in a lower absolute WER of 55.6% on dev. set and 62.5% on eval. set. Lattice MBR decoding contributes in a further reduction in WER achieving 47.9% on dev. set and 56.8% on eval. set.

8. Conclusions and Future Work

In this paper, a speech recognition system for Qatari Colloquial Arabic (QA) is proposed. Due to the limitation of dialectal speech resources, by utilizing MSA data, cross-dialectal phone mapping, data pooling, acoustic model adaptation and system combination methods, has achieved 21.3% and 28.9% relative WER reduction on QA development set and evaluation set respectively.

For future work, it is possible to extend the current framework to other dialect speech recognition systems. Moreover, some future directions are to incorporate recent achievements in transfer learning and domain adaptation to further improve the system performance (Pan and Yang, 2010). In addition, the cross-lingual training and adaptation can be bidirectional; a multi-task framework of Arabic speech recognition can be formulated so that both MSA and dialectal recognition performance can be enhanced simultaneously (Caruana, 1997).

9. Acknowledgments

This publication was made possible by a grant from the Qatar National Research Fund under its National Priorities Research Program (NPRP) award number NPRP 09-410-1-069. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Qatar National Research Fund. We would like also to acknowledge the European Language Resources Association (ELRA) and the Linguistic Data Consortium (LDC) for providing us with data resources.

10. References

Billa, J., Noamany, M., Srivastava, A., Liu, D., Stone, R., Xu, J., Makhoul, J., and Kubala, F. (2002). Audio indexing of Arabic broadcast news. In *Proceedings of ICASSP*, volume 1, pages 5–8.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75, Jul.

Elmahdy, M., Gruhn, R., Minker, W., and Abdennadher, S. (2010). Cross-lingual acoustic modeling for dialectal Arabic speech recognition. In *Proceedings of INTERSPEECH*, pages 873–876.

Elmahdy, M., Gruhn, R., Abdennadher, S., and Minker, W. (2011). Rapid phonetic transcription using everyday life natural chat alphabet orthography for dialectal Arabic speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4936–4939.

Ferguson, C. A. (1959). Diglossia. *Word*, 15:325–340.

Goel, V. and Byrne, W. (2000). Minimum Bayes-risk automatic speech recognition. *Computer Speech and Language*, 14(2):115–135.

Huang, P.-S. and Hasegawa-Johnson, M. (2012). Cross-dialectal data transferring for Gaussian mixture model training in Arabic speech recognition. In *International Conference on Arabic Language Processing*.

Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., and McLemore, C. (2002). *Egyptian Colloquial Arabic Lexicon, LDC Catalog No.: LDC99L22*. Linguistic Data Consortium, Philadelphia.

Kirchhoff, K. and Vergyri, D. (2005). Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. *Speech Communication*, 46(1):37–51.

Lee, C.-H. and Gauvain, J.-L. (1993). Speaker adaptation based on MAP estimation of HMM parameters. In *Proceedings of ICASSP*, pages II–558.

Leggetter, C.J. and Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185.

Mangu, L., Kuo, H.-K, Chu, S., Kingsbury, B., Saon, G., Soltau, H., and Biadsy, F. (2011). The IBM 2011 GALE Arabic speech transcription system. In *Proceedings of IEEE ASRU*, pages 272–277.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct.

Parker, R., Graff, D., Chen, K., Kong, J., and Maeda, K. (2009). *Arabic Gigaword, LDC Catalog No.: LDC2009T30*. Linguistic Data Consortium, Philadelphia, fourth edition.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of IEEE ASRU*.

Vergyri, D., Kirchhoff, K., Gadde, R., Stolcke, A., and Zheng, J. (2005). Development of a conversational telephone speech recognizer for Levantine Arabic. In *Proceedings of INTERSPEECH*, pages 1613–1616.