Cross-Lingual Acoustic modeling for Dialectal Arabic Speech Recognition

Mohamed Elmahdy[1,2], *Rainer Gruhn*[3], *Wolfgang Minker*[1], *Slim Abdennadher*[2]

[1]Faculty of Engineering & Computer Science, University of Ulm, Ulm, Germany
[2]Faculty of Media Engineering & Technology, German University in Cairo, Cairo, Egypt
[3]SVOX AG, Ulm, Germany

mohamed.elmahdy@guc.edu.eg, rainer.gruhn@alumni.uni-ulm.de, wolfgang.minker@uni-ulm.de, slim.abdennadher@guc.edu.eg

Abstract

A major problem with dialectal Arabic acoustic modeling is due to the very sparse available speech resources. In this paper, we have chosen Egyptian Colloquial Arabic (ECA) as a typical dialect. In order to benefit from existing Modern Standard Arabic (MSA) resources, a cross-lingual acoustic modeling approach is proposed that is based on supervised model adaptation. MSA acoustic models were adapted using MLLR and MAP with an in-house collected ECA corpus. Phoneme-based and grapheme-based acoustic modeling were investigated. To make phoneme-based adaptation feasible, we have normalized the phoneme sets of MSA and ECA. Since dialectal Arabic is mainly spoken, graphemic form usually does not match actual spelling as in MSA, a graphemic MSA acoustic model was used to force align and to choose the correct ECA spelling from a set of automatically generated spelling variants lexicon. Results show that the adapted MSA acoustic models outperformed acoustic models trained with only ECA data.

Index Terms: Dialectal Arabic, speech recognition, adaptation, acoustic modeling, cross-lingual

1. Introduction

Arabic language is the largest still living Semitic language in terms of number of speakers. Around 250 million persons are using Arabic as their first native language and it is the 6^{th} 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 broadcast, newspapers, formal speech, books, movies 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 natural spoken variety of Arabic in everyday life.

A major problem in Arabic speech recognition is the existence of quite many different Arabic dialects. Every country has its own dialect and sometimes there exist different dialects within the same country. Dialects can be classified into two groups: Western Arabic and Eastern Arabic. Western Arabic can be sub-classified into Moroccan, Tunisian, Algerian, and Libyan dialects, while Eastern Arabic can be sub-classified into Egyptian, Gulf, Damascus, and Levantine. In most cases, Arabic dialects differ drastically compared to MSA to the extent considering dialects as totally different languages. There are many speech data resources for MSA, but unfortunately the available resources for dialectal Arabic are very finite. That is why there are limited researches done in the area of dialectal Arabic speech recognition.

Basically, we are proposing a cross-lingual acoustic modeling approach for dialectal Arabic, where we can benefit from existing MSA speech resources, in order to improve dialectal Arabic recognition rate. Our assumption in this work is that MSA is always a second language for any Arabic speaker, and in most cases we can identify the original dialect of a speaker even though he is speaking MSA. For instant, in MSA news broadcast, speakers from Egyptian origins usually use the phoneme /g/ instead of /ʤ/ and this is a clue that most likely the speaker is originally Egyptian (we use IPA for phonetic transcription). So, we assume that an acoustic model trained with a sufficient number of MSA speakers will implicitly model the acoustic features of the different Arabic dialects, so we can call it dialect-independent. In order to fit the MSA acoustic model with a specific Arabic dialect (to make it dialect-dependent), we can use any of the acoustic modeling adaptation techniques like Maximum Likelihood Linear Regression (MLLR) [1], Maximum A-Posteriori (MAP) re-estimation [2], or a combination of different techniques.

In MLLR adaptation, we compute a set of linear transformations for the Gaussian mixture model (GMM) parameters (mainly Gaussian means) to map them in order to maximize the likelihood of HMM to generate the adaptation data. MAP re-estimation (also known as Bayesian adaptation) benefits from the prior knowledge about the initial model. Limited adaptation data would modify the initial model parameters with the guidance of the prior knowledge, in order to prevent large modifications of the initial parameters unless the adaptation data are enough to make large changes in parameters.

Acoustic model adaptation techniques are usually applied where the phoneme set of the adaptation data matches the phoneme set of the initial model. Since, usually Arabic dialects have different phoneme sets compared to MSA, then in order to make adaptation feasible, we have to normalize both phoneme sets of MSA and the Arabic dialect as shown later in this paper.

In this research, we have selected Egyptian Colloquial Arabic (ECA) as a typical Arabic dialect. ECA is the first ranked Arabic dialect in terms of number of speakers. ECA is also well known across all the Arab countries because of the popularity of the Egyptian movies and series.

Previous work [3], a cross-lingual approach was proposed where they tried to use a pooled MSA and ECA data in training the acoustic model. The result was a reduction in WER from 42.7% to 41.4% (relative: -3%). We think that the problem with the proposed approach in [3] is that the more the data of MSA,

the less contribution of ECA. Statistically, a small ECA corpus will not be enough to change the acoustic model parameters (means, variances, mixture weights, and transition weights) in conjunction with a huge amount of MSA data which will have the dominant effect. Hence, the model will be always biased to MSA and adding more MSA data will decrease the effect of ECA data.

2. MSA and ECA phonetics

We have collected the main differences (relevant to our work) between MSA and ECA: there are three short vowels in MSA /a/, /i/, and /u/ besides their long forms /aː/, /iː/, and /uː/ respectively. The /a/ vowel is sometimes pronounced as /ɑ/ (emphatic /a/) but usually it is miss-considered in the phonetic transcription of MSA. In ECA, there are three more vowels /e/, /o/, and /ɑ/ and their long forms as well. The letter Jeem should be formally pronounced /ʤ/ in MSA while it is inverted to /g/ in ECA. The letter Theh is pronounced /θ/ in MSA but it is inverted in ECA to /s/ or /t/. The letter Qaf is always pronounced /q/ in MSA but in ECA it is sometimes inverted to the glottal stop /ʔ/. The letter Thal /ð/ in MSA is inverted in ECA to /d/ or /z/. The letter Zah /ẓ/ is pronounced in ECA without being interdental, it is just a /ẓ/ (emphatic /z/). There are two diphthongs in MSA /ay/ and /aw/ that do not exist in ECA, and they are usually inverted in ECA to /eː/ and /oː/ respectively. ECA text transcriptions tend to use MSA spelling instead of the exact pronunciation, for example, the word دقيقة (minute) is always written with the letter Qaf /q/ though it is pronounced as a glottal stop /ʔ/.

3. Speech corpora

3.1. Modern Standard Arabic corpus

The Nemlar news broadcast speech corpus was chosen in training MSA acoustic models [4]. The corpus consists of 40 hours of MSA news broadcast speech. The broadcasts were recorded from different radio stations. All files were recorded in linear PCM format, 16 kHz, and 16 bit. The total number of speakers is 259 and the lexicon size is 62K distinct words with a phoneme set of 34 phonemes. This corpus was mainly selected because the transcriptions are fully diacritized and manually reviewed, and hence we have accurate phonetic transcription. We have processed the Nemlar corpus to exclude speech segments with music or noise in the background. Cross-talks and segments with truncated words were excluded as well. After filtration the 40 hours have been reduced to 33 hours with a tri-phones coverage of 57K distinct tri-phones.

3.2. Egyptian Colloquial Arabic corpus

We have collected the ECA corpus in-house. A database of the most frequently used words and utterances was created. The database includes utterances from different speech domains like: greetings, time and dates, words spelling, restaurants, train reservation, Egyptian proverbs...etc. Some samples are shown in Fig.1. The diversity of speech domains ensures good coverage of acoustic features. A lexicon of 700 words was created with accurate phonetic transcription using the dictionaries [5] and [6]. The phoneme set consists of 41 phonemes including the long and short forms of the vowels /a/, /i/, /u/, /e/, /o/, and /ɑ/. The foreign phonemes /v/, /p/, and /ʒ/ were also added to the phoneme set. The total number of speakers is 22 native Egyptian speakers with a tri-phones coverage of 15K distinct tri-phones. Every speaker was prompted to read 50 utterances chosen randomly from the database. All recordings were performed using the Sennheiser ME 3-N super-cardioid microphone connected with the InSync Buddy USB 6G digitizer, based on the Micronas UAC3556b microchip. All recordings were performed in linear PCM, 16 kHz, and 16 bits. The ECA corpus was divided into a training set of 65% of the speakers and a testing set of 35% of the speakers. The training set is used either to train the ECA baseline acoustic model or in adapting exiting MSA acoustic model.

Arabic: يوم الجمعة ٢ أكتوبر ٢٠١٠
English: Friday 2nd of October 2010
Phonemes: /yoːm iggumʕa ʔitneːn ʔuktoːbɑr ʔalfeːn wʕɑʃɑːrɑ/

Arabic: تذكرة درجة أولى
English: First class ticket
Phonemes: /tɑzkɑrɑ dɑrɑgɑ ʔuːla/

Arabic: بيض مسلوق
English: Boiled eggs
Phonemes: /beːḍ masluːʔ/

Figure 1: *Samples from the ECA corpus.*

4. System description

Our system is a GMM-HMM architecture based on the CMU Sphinx engine. Acoustic models are all fully continuous density context-dependent tri-phones with 3 states per HMM. The transcriptions of the ECA training set were used to build an open vocabulary bi-gram language model. In order to evaluate the language model, test-set perplexity test was performed using the ECA testing set transcriptions. The test results were as follows: perplexity of 19.6 (i.e. entropy of 4.29 bits), OOV rate of 6.54%, and bi-gram hits of 81.78%. All language modeling parameters were fixed during all the steps of the experiment, so that any change in recognition rate is mainly due to the acoustic model.

5. Phoneme-based acoustic modeling

5.1. Baseline

A baseline ECA phoneme-based acoustic model was trained with the ECA training set. The optimized number of tied-states and Gaussians were found to be 250 and 4 respectively (search was done in the range from 50 to 1000 tied-states and from 1 to 32 Gaussians). No approximations were applied on the phoneme set that consists of 41 phones. The result of decoding the ECA test set using the baseline acoustic model was an absolute WER of 35.1% as shown in Table 1.

5.2. Phonetic transcription normalization

Normalization is performed in order to have the same phoneme set and the same phonetic transcription convention across MSA and ECA. Actually, the majority of changes were done on the ECA side. The vowels /e/, /o/, and /ɑ/ in ECA were approximated to /i/, /u/, and /a/ respectively. We noticed that usually there are errors with the transcription of foreign phonemes in

MSA, for example, the word باريس (Paris) has the foreign phoneme /p/, however it is sometimes miss-interpreted as /b/. That is why we decided to normalize foreign phonemes. Foreign phonemes were normalized to the nearest standard ones, so /v/, /p/, and /ʒ/ were approximated to /f/, /b/, and /ʤ/ respectively. The consonant /g/ in ECA was approximated to /ʤ/. The consonant /ẓ/ in ECA was approximated to /ð̣/. On the other hand, in MSA, very minor changes were done: the diphthongs /aw/ was decomposed to /a/ followed by /w/ and /ay/ was decomposed to /a/ followed by /y/. All ECA and MSA transcriptions were updated with the above modifications. Finally, we have the same phoneme set across MSA and ECA.

5.3. MSA acoustic model

After normalization, the whole amount of the MSA corpus was used to train the MSA acoustic model with a typical number of tied-states and Gaussians of 4000 and 8 respectively. Decoding results of the ECA testing set were an absolute WER of 48.4% with a +37.9% relative increase compared to the ECA phoneme-based baseline. The increase in WER was expected, as according to our assumption, the MSA model is considered as a dialect-independent model.

5.4. Acoustic model adaptation

Basically, we are trying now to adapt the MSA acoustic model, in order to make it dialect-dependent and hence improve recognition rate. The MSA acoustic model was adapted using the ECA training set along with the normalized transcriptions. Three adaptation techniques were evaluated:

5.4.1. MLLR adaptation

Two iterations of MLLR were applied on the Gaussian means. In each iteration, Gaussian means were offline transformed using the MLLR matrix. The adapted model resulted in 37.9% absolute WER where we have -10.5% absolute reduction compared to MSA alone, and we were able to get closer to the baseline accuracy by +8.0% relative increase in WER.

5.4.2. MAP adaptation

MAP adaptation was found to give best results when adapting all acoustic model parameters: Gaussian means, variances, mixture weights, and transition weights. The absolute WER was 31.6% and it outperformed the baseline by -10.0% relative reduction in WER.

5.4.3. MLLR and MAP combined adaptation

In order to further improve results, we have combined both MLLR and MAP, starting by two iterations of MLLR followed by MAP. The absolute WER was 29.1% and it outperformed the baseline by -17.1% relative reduction in WER.

Table 1: *Decoding results of the ECA test set using MSA and different phoneme-based acoustic model adaptation techniques.*

acoustic model	WER	relative
ECA	35.1%	baseline
MSA	48.4%	+37.9%
MSA/ECA MLLR	37.9%	+8.0%
MSA/ECA MAP	31.6%	-10.0%
MSA/ECA MLLR+MAP	29.1%	-17.1%

6. Grapheme-based acoustic modeling

Grapheme-based acoustic modeling (also known as graphemic modeling) is an acoustic modeling approach for Arabic where the phonetic transcription is approximated to be the word letters instead of the exact phoneme sequence. Short vowels and gemination can be only estimated from a fully diacritized Arabic script. In Arabic phoneme-based acoustic modeling, we cannot just use a simple lookup table for phonetic transcription because of the morphological complexity and the high homograph rate. We found that in MSA we have in average ∼1.6 pronunciation variants for each word. Arabic script does not usually include diacritic marks. That is why in graphemic modeling, we rely on the acoustic model to implicitly model the missing diacritics and all pronunciation variants. Grapheme-based modeling was introduced in [7], and it was noticed that it works with an acceptable accuracy but it is less accurate than phoneme-based acoustic modeling. The main advantage of graphemic modeling is the rapid transcription development and to avoid the need for manual or automatic diacritization, as in [3] and [8]. The graphemic convention in our work is assigning one distinct phoneme to each letter except all forms of Alef and Hamza, they were all assigned the same phoneme. Overall we have 30 phonemes in our grapheme-based transcription convention.

6.1. Baseline

The grapheme-based baseline was built using only the ECA training set in a similar way to the phoneme-based baseline but the lexicon in this case is only graphemic. The result of decoding the ECA test set using the baseline acoustic model was an absolute WER of 40.2% as shown in Table 2 with a +5.1% absolute increase compared to the phoneme-based baseline.

6.2. MSA acoustic model

The graphemic MSA acoustic model was trained in an analogous way to the phoneme-based experiment except that the graphemic transcription convention was applied. The absolute WER was found to be 64.8% with a +61.2% relative increase compared to the baseline. The high WER increase is interpreted mainly due to the fact that ECA transcription does not follow the correct spelling as in MSA.

6.3. Adaptation

The same adaptation techniques were applied as shown in Table 2. It was found that with MLLR adaptation, the absolute WER was 50.3% which is lower by -40.5% absolute compared to MSA alone but it is still higher than the baseline by +25.1% relative. With MAP adaptation, the absolute WER was 39.1% and it slightly outperformed the baseline by -1.2% relative reduction in WER. When combining MLLR and MAP, the absolute WER was 36.1% and it outperformed the baseline by -10.2% relative reduction in WER and we were even able to approach the same accuracy of the ECA phoneme-based baseline.

6.3.1. Adaptation with spelling variants

Dialectal Arabic is mainly spoken and not formally written. That is why there is no single standard graphemic form for the same word, and in many cases we have different graphemic forms for the same word. Furthermore, ECA text is highly affected by MSA, and writers tend to use the MSA graphemic forms instead of the form that matches pronunciation. On the

Table 2: *Decoding results of the ECA test set using grapheme-based MSA acoustic model and different adaptation techniques.*

acoustic model	WER	relative
ECA	40.2%	baseline
MSA	64.8%	+61.2%
MSA/ECA MLLR	50.3%	+25.1%
MSA/ECA MAP	39.7%	-1.2%
MSA/ECA MLLR+MAP	36.1%	-10.2%

other hand, in MSA, the transcription always matches pronunciation. So our idea was to improve our ECA graphemic lexicon by adding automatically generated variants guided by the knowledge of which letters are usually miss written in ECA. In this case, it is more convenient to name the variants: *spelling variants* rather than pronunciation variants because the pronunciation in this case is the same but we have different graphemic forms. In ECA the majority of words with the letter Qaf ق are pronounced with the glottal stop /ʔ/ instead of /q/, while it is pronounced /q/ in other words. In the grapheme-based approach, we rely only on the graphemic transcription. That is why we do not have information about whether it is actually pronounced /q/ or /ʔ/. So we have added two spelling variants for words with the letter Qaf: one variant with /q/ and the other one with /ʔ/. Some words with the letter Thal ذ are pronounced /z/ and some other are pronounced /d/. So we have added two more spelling variants for words with the letter Thal: one variant with /z/ and the other one with /d/. Words with the letter Theh ث were added 3 spelling variants with /s/, /t/, and /θ/ respectively. Some samples are shown in Table 3.

Using the initial MSA graphemic acoustic model, ECA training set transcriptions were force-aligned against our multi-spelling variants lexicon. Force alignment should reduce spelling errors within ECA graphemic transcriptions. The force-aligned transcriptions along with the ECA training set were used to adapt the MSA acoustic model. Decoding results show that there is a reduction in WER as shown in Table 4. The initial MSA with the multi-spelling lexicon resulted in 64.4% absolute WER with -0.4% absolute reduction compared to not using spelling variants. MLLR adaptation resulted in 51.5% absolute WER. MAP and MLLR+MAP outperformed the baseline by -5.2% and -12.9% relative reduction in WER respectively, and by -4.0% and -3.0% relative reduction in WER compared to the same settings but without using spelling variants. It was noticed that when combining MLLR and MAP in the graphemic adaptation, the WER is almost the same as the ECA phoneme-based baseline with 35.0% absolute WER.

7. Conclusions and future work

In order to improve recognition rate of dialectal Arabic, we showed that we can benefit from existing MSA speech resources using supervised acoustic model adaptation techniques. The proposed cross-lingual acoustic modeling approach showed best results when combining MLLR and MAP adaptations. In the case of phoneme-based acoustic modeling, the adapted MSA model outperformed the ECA baseline by -17.1% relative reduction in WER. With the grapheme-based approach, we showed that by adding automatically generated spelling variants to the graphemic lexicon, we can force align ECA transcriptions using an initial MSA model to choose the correct spelling that matches the actual pronunciation. Finally, with the graphemic adaptation, we were able to outperform the ECA baseline by -12.9% relative reduction in WER. With graphemic

Table 3: *Samples from the ECA graphemic lexicon after adding spelling variants, showing which variant is commonly written and which one matches pronunciation*

Arabic	English	spelling variants	common written	match pronunciation
ذرة	corn	ذ ر ة	✓	
		ز ر ة		
		د ر ة		✓
قهوة	coffee	ق ه و ة	✓	
		أ ه و ة		✓
قرية	village	ق ر ي ة	✓	✓
		أ ر ي ة		
ثوم	garlic	ث و م	✓	
		ت و م		✓
		س و م		

Table 4: *Effect of adding spelling variants on decoding results of the ECA test set using MSA and different grapheme-based acoustic model adaptation techniques.*

acoustic model	WER	relative
ECA	40.2%	baseline
MSA	64.4%	+60.2%
MSA/ECA MLLR	51.5%	+28.1%
MSA/ECA MAP	38.1%	-5.2%
MSA/ECA MLLR+MAP	35.0%	-12.9%

modeling adaptation, it was noticed that the combination of MLLR, MAP, and spelling variants performed as good as the phoneme-based ECA baseline. For future work, we will investigate more in cross-lingual acoustic modeling via unsupervised adaptation where in this case the adaptation data are not labeled.

8. References

[1] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models", Computer Speech and Language, vol. 9, pp. 171-185, 1995.

[2] Chin-Hui Lee and Jean-Luc Gauvain, "Speaker Adaptation Based on MAP Estimation of HMM Parameters", Proceedings of ICASSP, p. II-558, 1993.

[3] Katrin Kirchhoff and Dimitra Vergyri, "Cross-Dialectal Data Sharing For Acoustic Modeling in Arabic Speech Recognition", Speech Communication, 46(1), pp. 37-51, 2005.

[4] The Nemlar project, http://www.nemlar.org/.

[5] Martin Hinds and El-Said Badawi, "A Dictionary of Egyptian Arabic", Librairie du Liban, Reprinted 2009.

[6] Virginia Stevens and Maurice Salib, "A Pocket Dictionary of the Spoken Arabic of Cairo", The American University in Cairo Press, Second printing 2005.

[7] J. Billa, M. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, and F. Kubala, "Audio Indexing of Arabic Broadcast News", Proceedings of ICASSP, pp. 1:5-8, 2002.

[8] Ruhi Sarikaya, Ossama Emam, Imed Zitouni, and Yuqing Gao, "Maximum Entropy Modeling for Diacritization of Arabic Text", Proceedings of INTERSPEECH, pp. 145-148, 2006.