A Baseline Speech Recognition System for Levantine Colloquial Arabic

Mohamed Elmahdy[*1], Mark Hasegawa-Johnson[**2], Eiman Mustafawi[*3]

[*]*Qatar University, Qatar*

[1]mohamed.elmahdy@qu.edu.qa

[3]eimanmust@qu.edu.qa

[**]*University of Illinois, USA*

[2]jhasegaw@illinois.edu

***Abstract*— The Arabic language is characterized by the existence of many different colloquial varieties that significantly differ from the standard Arabic form. In this paper, we propose a state-of-the-art speech recognition system for Levantine Colloquial Arabic (LCA). A fully continuous context dependent acoustic model was trained using 50 hours of speech from the BBN DARPA Babylon corpus. Pronunciation modeling was initially grapheme-based due to the absence of diacritic marks in transcriptions. Acoustic model parameters have been optimized including number of senones and Gaussians. In order to improve speech recognition accuracy, a cross-lingual hybrid acoustic and pronunciation modeling approach is proposed, where a MSA phoneme-based acoustic model is adapted using a small amount of LCA speech data. The adapted AM was then combined with the initial grapheme-based model to create a hybrid acoustic model.**

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 6th 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.

The majority of previous work in Arabic Automatic Speech Recognition (ASR) has focused on MSA whilst relatively little work has focused on dialectal Arabic. A significant problem in Arabic speech recognition is the existence of many different Arabic dialects. Every country has its own dialect and usually 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.

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 and it has been discussed in [1, 2].

In this work, we are proposing an ASR system for Levantine Colloquial Arabic (LCA). LCA Arabic is the dialect of Arabic spoken by people in Lebanon, Syria, Jordan, and Palestine. Since the majority of dialectal Arabic speech resources are usually provided with graphemic transcriptions lacking diacritic marks, we have initially created a grapheme-based ASR system, where the phonetic transcription is approximated to be the word letters rather than the actual pronunciation.

In order to improve speech recognition accuracy, a cross-lingual hybrid acoustic and pronunciation modeling approach is proposed. First, a MSA phoneme-based acoustic model was trained. The MSA AM was then adapted using Maximum Likelihood Linear Regression (MLLR) followed by Maximum A-Posteriori (MAP) with a little amount of LCA speech data that has been phonetically transcribed. Afterwards, the adapted phonemic AM and the initial graphemic AM are fused together in order to create the hybrid model.

2 LEVANTINE COLLOQUIAL ARABIC SPEECH CORPUS

The BBN/AUB DARPA Babylon Levantine Arabic Corpus [6] is a corpus of controlled spontaneous speech. The corpus is recorded from subjects having Levantine colloquial Arabic as their native language. The subjects were from Lebanon, Syria, Jordan, and Palestine.

The corpus was recorded using a close-talking, noise-cancelling, headset microphone at 16kHz sampling rate. The subjects in the corpus were responding to refugee/medical questions like: (Where is your pain?, How old are you?, etc.), and were playing the part of refugees. Each subject was given a part to play, that prescribed what information they were to give in response to the questions. However, they were advised to express themselves naturally, in their own way, in Arabic. To avoid priming subjects to give their answer with a particular Arabic wording, the parts were given in English rather than Arabic.

The total number of recorded speakers is 164 with a vocabulary size of 14.7K unique words. The lexicon consists of only words in the graphemic form without phonetic transcription. Furthermore, we could not find any accurate pronunciation lexicon to map words to corresponding phonemes. Actually, this case is normal for all the Arabic dialects since they are only spoken and not written. That is why it is difficult to find a standard way to estimate the correct phonemes sequence for a given dialectal word.

The LCA corpus was sub-divided into a training set of ~50 hours (136 speakers) and a testing set of ~10 hours (28 speakers).

3 Language Modeling

The language model is a statistical backoff tri-gram model with Kneser-Ney smoothing. The language model has been trained with the transcriptions of the LCA corpus training set (~230K words). The vocabulary size of the train set was found to be 13.6K unique words. The evaluation of the language model against the transcriptions of the testing set resulted in an OOV rate of 1.8%, tri-grams hits of 74.3%, and perplexity of 34 (entropy of ~5.1 bits) as shown in Table I. Language modeling in this research was carried out using the CMU-Cambridge Statistical Language Modeling Toolkit [3, 10].

TABLE I
LANGUAGE MODEL PROPERTIES AND EVALUATION AGAINST THE TRANSCRIPTIONS OF THE TESTING SET

Training words	230K
Vocabulary	13.6K
OOV	1.8%
Perplexity	34
Tri-gram hits	74.3%

4 System Description

Our system is a GMM-HMM architecture based on the CMU Sphinx engine [4, 5]. Acoustic models are all fully continuous density context-dependent tri-phones with 3 states per HMM trained with MLE. The feature vector consists of the standard 39 MFCC coefficients. During acoustic model training, linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) were applied to reduce dimensionality to 29 dimensions. This was found to improve accuracy as well as recognition speed. Decoding is performed in multi-pass, a fast forward Viterbi search using a lexical tree, followed by a flat-lexicon search and a best-path search over the resulting word lattice.

5 Grapheme-based Acoustic Modeling

Grapheme-based acoustic modeling (also known as graphemic modeling) is an acoustic modeling approach where the phonetic transcription is approximated to be the word letters rather than the exact phonemes sequence. Short vowels and germinations are assumed to be implicitly modeled in the acoustic model. Each letter was mapped to a unique model resulting in a total number of 36 base units (letters in the Arabic alphabet). In this case, pronunciation modeling is a straightforward process, so for any given word, pronunciation modeling is done by splitting the word into letters. In this case, each word is associated with only one graphemic pronunciation variant.

The 50 hours LCA training set was used for acoustic modeling. The acoustic model consists of both context-independent (CI) and context-dependent (CD) phones. During decoding, CI models are used to compute likelihood for tri-phones that have never been seen in the training set. The graphemic acoustic model consists of 108 CI states. In order to determine the optimized number of CD tied-states (senones) and the number of Gaussians per state, several acoustic models have been created with varying number of senones (500 to 4500) and varying number of Gaussians (4 to 128). For each setting, decoding was performed over the testing set as shown in Table II.

The optimized number of senones and Gaussians per state were found to be 2000 and 64 respectively, resulting in an absolute WER of 30.5 % as shown in Table II.

TABLE III
WORD ERROR RATE (WER) % ON THE LCA TEST SET, OPTIMIZING THE TOTAL NUMBER OF TIED-STATES (TS) AND GAUSSIAN DENSITIES

Gaussians	**Tied states (TS)**								
	500	**1000**	**1500**	**2000**	**2500**	**3000**	**3500**	**4000**	**4500**
4	44.1	41.8	41.7	40.1	38.6	39.7	38.8	40.2	38.5
8	39.8	38.1	37.3	36.0	36.2	36.4	34.7	36.1	33.3
16	38.2	35.7	34.7	34.8	33.5	32.4	32.6	33.2	32.9
32	36.3	34.4	33.1	33.4	32.4	33.6	33.6	32.3	34.5
64	34.5	33.1	32.8	**30.5**	32.4	33.1	32.4	32.8	34.1
128	32.8	32.1	31.3	32.0	33.3	34.3	35.5	37.8	40.2

In order to improve accuracy, we have applied linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) to reduce dimensionality from 39 to 29 dimensions. This was found to decrease WER with 1.3% relative as shown in Table III. Since we are training a graphemic model, cross-phone training was applied and this was found to further decrease WER to 28.8% absolute (-5.6% relative to the baseline) as shown in Table III.

TABLE IIIII
WER (%) ON THE LCA TEST SET, COMPARING FEATURE TRANSFORMATION AND CROSS-PHONE TRAINING RELATIVE TO THE BASELINE AM

	WER	**Relative**
Baseline	30.5%	-
Feature transform (FT)	30.1%	-1.3%
FT + Cross-phone	28.8%	-5.6%

6 HYBRID ACOUSTIC MODELING

Hybrid acoustic modeling is performed by two independent acoustic model trainings (phonemic model and another graphemic model). Afterwards, the two models are fused together into one hybrid model [8].

Since we could not obtain large amounts of phonetically transcribed speech data for Levantine Arabic, we have adopted a cross-lingual approach to train the phonemic AM using MSA speech data in a similar way as described in [9]. The MSA phonemic AM was trained using 62 hours of speech data from two corpora provided by ELRA:

- The NEMLAR Broadcast News Speech Corpus (~40 hours) [11].
- The NetDC Arabic BNSC (Broadcast News Speech Corpus) (~22 hours).

Both corpora are provided with fully vocalized transcriptions, and therefore grapheme-to-phoneme is almost a one-to-one mapping. The phonemic MSA AM was then adapted using MLLR adaption followed by MAP along with 100 utterances from the LCA train set. Phonetic transcriptions for these 100 utterances were manually prepared.

Phonemic pronunciation variants were generated using the LDC Standard Arabic Morphological Analyser (SAMA) [7]. Pronunciation modeling was hybrid by generating a graphemic variant along with all possible phoneme variants generated by the morphological analyzer. Decoding results of the LCA test set shows that we can achieve an improved accuracy of 28.4% absolute WER with a 1.3% relative reduction compared to the baseline. It was noticed that many pronunciation variants that are generated from the morphological analyzer differ from the actual LCA pronunciation only in vowels. e.g. /i/ realized as /e/. That is why our next step is to normalize vowels in the output from the morphological analyzer as well as the MSA phonemic AM.

TABLE IVV
WER (%) ON THE LCA TEST SET, COMPARING GRAPHEMIC AND HYBRID ACOUSTIC MODELING

	WER	**Relative**
Graphemic AM	28.8%	-
Hybrid AM	28.4%	-1.3%

7 CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed an ASR system for Levantine Colloquial Arabic (LCA). First, a grapheme-based acoustic model was trained using 50 hours of LCA speech data. The optimized number of senones and Gaussians densities were found to be 2000 and 64 respectively. Batch decoding of the test set (10 hours) resulted in a WER of 30.5%. Feature transformation using LDA and MLLT was applied to reduce dimensionality resulting in WER reduction of 1.3% relative. Cross-phone training was found to add further reduction in WER achieving 5.6% relative to the baseline.

A cross-lingual phonemic acoustic model for LCA was prepared by adapting existing MSA acoustic model with little LCA speech data along with manually prepared phonetic transcription. A hybrid acoustic model was created by combining the graphemic and the cross-lingual phonemic models. Hybrid acoustic modeling resulted in a further reduction in WER of 1.3% relative (28.4% absolute).

For future work, the proposed hybrid approach will be extended and evaluated with the other Arabic colloquials (e.g. Egyptian, Iraqi, Gulf, etc.).

ACKNOWLEDGMENT

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 Linguistic Data Consortium (LDC) and the European Language Resources Association (ELRA) for providing us with the required speech and text resources to conduct this research.

REFERENCES

[1] A.S. Kaye, "Modern Standard Arabic and the Colloquials", Lingua 24, pp. 374-391, 1970.

[2] C. Ferguson, "Diglossia", Word 15, pp. 325-340, 1959.

[3] Carnegie Mellon University-Cambridge, CMU-Cambridge Statistical Language Modeling toolkit, http://www.speech.cs.cmu.edu/SLM/toolkit.html

[4] Carnegie Mellon University Sphinx, Speech Recognition Toolkit, http://cmusphinx. sourceforge.net/

[5] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A I. Rudnicky, "Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices", In Proceedings of ICASSP, vol. 1, pp. 185-188, 2006.

[6] J. Makhoul, B. Zawaydeh, F. Choi, and D. Stallard, "BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts", Linguistic Data Consortium(LDC), LDC Catalog No.: LDC2005S08, 2005.

[7] M. Maamouri, D. Graff, B. Bouziri, S. Krouna, A. Bies, and S. Kulick, "LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1", Linguistic Data Consortium(LDC), LDC Catalog No.: LDC2010L01, 2010.

[8] M. Elmahdy, M. Hasegawa-Johnson, and E. Mustafawi, "Hybrid Pronunciation Modeling for Arabic Large Vocabulary Speech Recognition", Qatar Foundation Annual Research Forum, 2012.

[9] M. Elmahdy, R. Gruhn, and W. Minker, "Cross-Lingual Acoustic Modeling for Dialectal Arabic Speech Recognition", In Proceedings of Interspeech, pp. 873-876, 2010.

[10] P. Clarkson, and R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit", In Proceedings of ISCA Eurospeech, 1997.

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