Pronunciation Modeling for Dialectal Arabic Speech Recognition

Hassan Al-Haj, Roger Hsiao, Ian Lane, Alan W. Black, Alex Waibel

interACT / Language Technologies Institute
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA

{hhaj,wrhsiao,ianlane,awb,ahw}@cs.cmu.edu

***Abstract*— Short vowels in Arabic are normally omitted in written text which leads to ambiguity in the pronunciation. This is even more pronounced for dialectal Arabic where a single word can be pronounced quite differently based on the speaker's nationality, level of education, social class and religion. In this paper we focus on pronunciation modeling for Iraqi-Arabic speech. We introduce multiple pronunciations into the Iraqi speech recognition lexicon, and compare the performance, when weights computed via forced alignment are assigned to the different pronunciations of a word. Incorporating multiple pronunciations improved recognition accuracy compared to a single pronunciation baseline and introducing pronunciation weights further improved performance. Using these techniques an absolute reduction in word-error-rate of 2.4% was obtained compared to the baseline system.**

I. INTRODUCTION

The Arabic alphabet only consists of letters for long vowels and consonants. Other pronunciation phenomena, including short vowels (harakat), nunation (tanwin) and consonant doubling (shadda), are not typically written. However, they can be explicitly indicated using diacritics. Vowel diacritics represent the three short vowels: a, i, and u (fatha, kasra and damma) or the absence of a vowel (sukun). For example, the four vowel diacritics in conjunction with the Arabic letter ب /b/[1] are written as:

بَ	/ba/	(fatha)
بِ	/bi/	(kasra)
بُ	/bu/	(damma)
بْ	/b/	(sukun), absence of a vowel.

Diacritics indicating nunation (tanwin) only occur in word final positions in indefinite nominals (nouns, adjectives and adverbs). They indicate a short vowel followed by an unwritten n sound: بً /ban/, بٌ /bun/, بٍ /bin/. The diacritic shadda indicates a consonant doubling: بّ /bb/. Shadda can also combine with a vowel or a nunation as in: بُّ /bbu/ or بٌّ /bbun/. In this paper, a word fully annotated with diacritics, as it will be pronounced, will be called the "vowelized form". Diacritics are normally omitted (or appear only partially) in most written text. This leads to two main problems when developing Arabic ASR systems.

TABLE I
EXAMPLES OF ARABIC WORDS AND THEIR VOWELIZED FORMS

Written Form	Pronunciation	Meaning
مدربين /ktb/ (MSA)		
كتب	كَتَبَ /kataba/	he wrote
كتب	كُتِبَ /kutiba/	it was written
كتب	كَتَّبَ /kattab/	make/cause him write
كتب	كُتُبْ /kutub/	books
كتب	كُتُبٍ /kutubin/	books (indefinite)
مدربين /mdrbjn/ (Iraqi)		
مدربين	مُدَرَّبِيْنْ /mudarrabijn/	trained (adj. , dialect 1)
مدربين	مْدَرَّبِيْنْ /mdarrabijn/	trained (adj. , dialect 2)
مدربين	مْدَرْبِيْنْ /mdarbijn/	two trainers
مدربين	مُدَرِبِيْنْ /mudarribijn/	trainers

First, the vowelized form of a word is ambiguous, thus it must either be explicitly hypothesized, or this ambiguity must be handled in the models applied during ASR decoding. Second, the absence of short vowels increases ambiguity in the language model. Words with identical written-forms may actually be homographs that occur in very different linguistic contexts. Treating them identically will decrease the predictability of the language model. Table I shows there are six possible interpretations of the modern standard Arabic (MSA) word كتب /ktb/ showing the difficulty of this task.

There exist many dialects of Arabic which are used in daily communication. These dialects differ significantly from MSA which is primary used in written text and formal communications. As opposed to MSA, dialectal Arabic can be spoken quite differently based on a person's nationality, level of education, social class and religion. This creates numerous pronunciations of the same word, for example, the four possible pronunciations of the Iraqi word مدربين /mdrbjn/ as given in Table I.

[1] We use the Buckwalter transliteration to romanize Arabic examples (Buckwalter, 2002).

Resources such as morphological analysers, which are generally used when building MSA speech recognition systems, are not available for resource deficient Arabic dialects. Moreover, it is non-trivial to automatically generate diacritics using basic grammatical rules due to the large variability of pronunciations within Arabic dialects. In this work we investigate explicitly modeling diacritics in the acoustic model, and handling pronunciation variants within the recognition dictionary. We built and compared three vowelized ASR systems: a baseline system, where each lexical entry has only a single pronunciation; a multi-pronunciation system, where each lexical entry could contain more than one pronunciation, and a second multi-pronunciation system where entries in the recognition dictionary had non-uniform pronunciation weights.

In previous works [1]-[3], MSA speech recognition systems that explicitly modeled diacritics in the acoustic model and considered multiple pronunciations during decoding were shown to outperform grapheme-based systems. In these works, the Buckwalter Morphological Analyzer [4] was used to generate all possible pronunciations of a word, and acoustic models were then trained using either manually vowelized transcripts [1], or by automatically generating labels when performing force alignment of the training data. Both approaches obtained higher recognition accuracy than comparable grapheme-based systems. When using multi-pronunciation dictionaries in recognition, further improvement was gained by assigning non-uniform weights to pronunciations, for example, based on word counts obtained during forced alignment.

In this work we focus on one specific dialect of Arabic, Iraqi-Arabic. For this dialect we cannot rely on off-the-shelf morphological analyzers, as Iraqi is typically not written, furthermore, in Iraqi pronunciation variants cannot simply be generated via grammatical rules. Therefore, we leverage pronunciation dictionaries, including the LDC Iraqi Arabic Morphological Lexicon to obtain pronunciations of Iraqi words. For words that did not appear in the manually compiled lexicon we automatically generate a pronunciation using the method described in Section II-A.

We compared two variants of a multi-pronunciation system to a baseline system which uses a single pronunciation per word. The first multi-pronunciation system has uniform weights assigned to all pronunciations. The second, uses weights computed via forced alignment. We show that both multi-pronunciation systems outperform the baseline and that using non-uniform pronunciation weights further improves recognition accuracy. Further gains were obtained by retraining the acoustic model using phonetic labels generated by a multi-pronunciation system.

The remainder of the paper is organized as follows: In Section II we describe the multi-pronunciation dictionary, and introduce a method to estimate pronunciation weights. In Section III we describe the baseline system, and compare the performance of the uniform and weighted multi-pronunciation systems. Finally, conclusion and future works are described in Section IV.

II. Pronunciation Modelling for Iraqi-Arabic

We built and compared three vowelized Iraqi speech recognition systems. The first uses a single pronunciation per word, while the other two use multiple pronunciations. The two multi-pronunciations systems use the same recognition lexicon but differ in the probabilities (weights) assigned to the pronunciations variants of a word.

A. Vowelized Pronunciation Dictionaries

The single pronunciation dictionary contains 130k words. The pronunciations for 95k of the words were obtained from manually compiled pronunciation dictionaries, including the LDC Iraqi-Arabic Morphological Lexicon, as well as, wordlists, and name entity lexicons provided to groups within the DARPA TransTAC project. When a word appeared in these lists with more than one pronunciation, only the most frequent was selected. Pronunciations for the remaining 45k words were generated using a statistical method trained using the dictionaries listed above. We used an extension to the CART-based method described in [5], in which we first predict probability densities of possible phones and then perform Viterbi decoding applying a phoneme-based trigram model. This approach is described in [6]. The single best hypothesis was used as the resulting pronunciation.

The multi-pronunciation dictionary was built in a similar manner, except all pronunciations from manually compiled dictionaries were included. The resulting dictionary contained 1.7 pronunciations per word on average.

B. Estimating Pronunciations Probabilities

We built two speech recognition systems that used the multi-pronunciation dictionary described above. The first, assigned uniform probabilities to pronunciation variants; the second assigned weights estimated using 450 hours of acoustic model (AM) training data. Pronunciation weights were estimated using the following algorithm:

1. Perform forced alignment, using a multi-pronunciation dictionary, and generate labels.
2. Use the labels generated to train an AM.
3. Perform forced alignment using the AM trained above.

For a given word W with n pronunciations $pr_i \quad 1 \le i \le n$ the pronunciation probability is:

$$P_i = \begin{Bmatrix} \frac{\#FA(pr_i)}{\#C(w)} & \#C(w) \neq 0 \\ 1/n & \#C(w) = 0 \end{Bmatrix} \qquad (1)$$

$\#FA(pr_i)$: The count that pr_i appears in the forced alignment.

$$\#C(w) = \sum_{i=1}^{n} \#FA(pr_i) \qquad (2)$$

If for some pronunciation $\#FA(pr_i)=0$ but $\#C(w)\neq 0$ then this pronunciation is eliminated from the dictionary i.e. it will not be used during decoding.

III. Experimental Evaluation

We evaluated the three systems described in Section II using the DARPA TransTAC 2008 evaluation sets. The June08 set, which comprises of 7.5k words and 58 minutes of speech, was used for development, and the Nov08 set, consisting of 6.5k words and 54 minutes of speech, was used as unseen test data.

A. System Architecture

Our Iraqi ASR system consists of a 3-state, sub-phonetically tied, semi-continuous, HMM acoustic model and is composed of 7000 context dependent triphone/quintphone models. Each model consists of a mixture of up to 64 Gaussians, where the exact number is determined by merge-and-split training. Input speech features consist of 13 Mel Frequency Cepstral Coefficients (MFCC), power, and approximations of the first and second derivatives. Linear discriminant analysis is applied to reduce the dimensionality to 42 coefficients. The acoustic model was trained using 450 hours of Iraqi-Arabic speech data provided within the TransTAC project. A trigram language model using modified Kneser-Ney smoothing was applied during decoding with a recognition vocabulary of 62k words. The language model was trained using approximately 4M words. ASR decoding was performed using the Ibis decoder [7], which was developed as part of our Janus Recognition Toolkit (JRTk) [8].

B. Evaluation

We evaluated the effectiveness of using multiple pronunciations in the Iraqi ASR component of our Iraqi-English TransTAC system [9]. First, we evaluated the performance of the single pronunciation (baseline) system. This system obtained a word-error-rate (WER) of 37.0% on the June08 development set and 35.7% on the Nov08 test set. Next, the two multi-pronunciation systems were evaluated applying the original acoustic model trained using the single pronunciation lexicon (AM0). When uniform pronunciation weights were applied, WERs of 36.5% and 35.0% were obtained on the June08 and Nov08 sets. When pronunciation weights were introduced (as described in step 3, section II-B), the WERs obtained were 36.5% and 34.8%, for June08 and Nov08, respectively. In both cases, the multi-pronunciation systems improved recognition accuracy compared to the single pronunciation baseline. This indicates that a single pronunciation is inadequate to model all the variants present in Iraqi speech. Introducing pronunciation weights further improved performance showing the importance of weighting competing variants during recognition.

To evaluate the effect of the acoustic model, we trained a new model AM1 using the multi-pronunciation lexicon (steps 1 and 2, section II-B). Using AM1, decoding with uniform pronunciation weights obtained WERs of 36.7% and 34.5% on June08 and Nov08, while the system with pronunciation weights (re-estimated using AM1) obtained WERs of 36.1% and 33.8%. Finally, an additional iteration of AM training was performed. After this second iteration (AM2) the system with uniform weights had a WER of 36.4% on the June08 and 34.4% on Nov08, while the weighted system had WERs of 35.8% and 33.3%. Results are summarized in Table II.

TABLE II
Comparison of ASR Performance of Different Systems on June08 (Development) and Nov08 (Test) Evaluation Sets

AM	System	WER on June08	WER on Nov08
AM0	SP	37.00	35.70
AM0	UP	36.53	35.00
AM0	WP	36.49	34.80
AM1	UP	36.69	34.50
AM1	WP	36.07	33.80
AM2	UP	36.38	34.40
AM2	WP	35.84	33.30

SP= single pronunciation (baseline).
UP= multi-pronunciation with uniform weights.
WP= multi-pronunciation with estimated weights

On our Iraqi-Arabic system it was observed that pronunciation weights, which are estimated based on acoustic evidence always improved recognition performance. The weighted multi-pronunciation system using AM2 improved WER by 1.2% absolute on the development set, and by 2.4% absolute on unseen test data. Even with uniform weights, the multi-pronunciation system outperformed the baseline system, which used a single pronunciation per word. Additional iterations of AM training which involved relabeling the training corpora and then retraining the AM further improved recognition performance.

IV. Conclusion and Future Work

In this paper we investigated pronunciation modeling for dialectal Arabic speech, focusing on speech recognition of Iraqi-Arabic. Due to the ambiguity of pronunciations inherent in MSA and dialectal Arabic text, we investigated approaches to introduce multiple pronunciations into our Iraqi speech recognition system. This was done by incorporating alternative pronunciations, acquired from manually compiled pronunciation dictionaries. A significant improvement in recognition accuracy was obtained compared to a baseline system in which only a single pronunciation was used. Further improvement was gained by applying pronunciation probabilities estimated via forced alignment of the training corpora. Finally, by iteratively retraining the acoustic model an absolute reduction in word-error-rate of 2.4% was obtained compared to the single pronunciation baseline.

In future work we intend to investigate pronunciation modeling in combination with discriminative training of acoustic models. We will also investigate methods to optimally handle pronunciation variants within the language model. Specifically, pronunciation variants of the same word should be treated identically within the language model, whereas, it may be best to handle homographs independently.

ACKNOWLEDGMENT

This work is in part supported by the DARPA TransTAC (Spoken Language Communication and Translation System for Tactical Use) program. Any opinions, findings, and conclusions or recommendations references expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

REFERENCES

[1] Messaoudi A., Gauvain J. L., and Lamel L., "Arabic broadcast news transcription using a one million word vocalized vocabulary," in ICASSP-2006, Toulouse, France.

[2] Soltau H., Saon G., Kingsbury B., Kuo J., Mangu L., Povey D. And Zweig G., "The IBM 2006 GALE Arabic ASR System", International Conference on Acoustics, Speech and Signal Processing, Hawaii, 2007.

[3] Noamany M., Schaaf T., Schultz T., "Advances in the CMU-InterACT" Arabic Gale Transcription System, Proceedings of the HLT/NAACL 2007, Rochester, NY, US, April 22-27, 2007.

[4] Buckwalter T., "BuckWalter Arabic morphological analyzer version 2.0," in LDC2004L02. 2004, Linguistic Data Consortium.

[5] Black, A., Lenzo, K. and Pagel, V. (1998) Issues in Building General Letter to Sound Rules 3rd ESCA Workshop on Speech Synthesis, pp. 77-80, Jenolan Caves, Australia.

[6] Chotimongkol, A. and Black, A. (2000) Statistically trained orthographicto sound models for Thai, ICSLP2000, Beijing, China.

[7] Soltau, H., Metze, F., Fugen, and Waibel A., "A onepass decoder based on polymorphic linguistic context assignment," in Proc. ASRU, 2001.

[8] Finke. M., Geutner. P., Hild , H., Kemp T., Ries, K. and M.Westphal, "The Karlsruhe-Verbmobil Speech Recognition Engine," in Proc. ICASSP, 1997.

[9] Bach, N., Eck, M., Charoenpornsawat, P., Köhler, T., Stüker, S.,Nguyen, T., Hsiao, R., Waibel, A., Vogel, S., Schultz, T., and Black A., "The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System", IWSLT 2007, Trento, Italy.