CONTEXT-DEPENDENT PRONUNCIATION MODELING FOR IRAQI ASR

Stavros Tsakalidis, Rohit Prasad, Prem Natarajan

BBN Technologies
10 Moulton St., Cambridge, MA, USA
{stavros,rprasad,pnataraj}@bbn.com

ABSTRACT

In this paper, we introduce a novel pronunciation modeling technique that in contrast to existing techniques uses *word context* information. This *context-dependent* pronunciation modeling is designed to overcome the challenges posed by absence of diacritics in transcripts for training acoustic models for Arabic dialects. To demonstrate the efficacy of the proposed pronunciation modeling, we present experimental results with both manually created and automatically generated vowelized lexicons on the DARPA TRANSTAC colloquial Iraqi corpus.

Index Terms— Pronunciation modeling, Arabic ASR

1. INTRODUCTION

One of the prominent problems in developing automatic speech recognition (ASR) systems for Arabic dialects is the absence of short vowel information from acoustic transcripts and in other text for language modeling. Typically, there are two approaches to address this problem. The first is to use a "grapheme" system where the pronunciation of a word is based on just the orthographic form [1]. The alternative approach is to use a manual or automatic diacritization method to derive the phonetic transcription of each word that explicitly includes the short vowels. Previous work [2, 3] showed that such systems achieve significant word error rate (WER) reduction over grapheme systems.

In principle manual diacritization should be the most accurate. However, in practice it is a difficult and time consuming task, especially for colloquial dialects of Arabic. Furthermore, most ASR systems use a language model which is typically trained on both the acoustic training and a potentially large corpus of text from similar or different domains. Majority of such relevant text resources do not have the short vowel information and could not be used as is, if the ASR system was trained with transcripts containing short vowels.

Automatic diacritization methods also have several shortcomings, most of which result due to the creation of significantly more vowelization variants than that are practically useful. For example, the Buckwalter morphological analyzer (BMA) outputs all possible vowelizations of an unvowelized word. The increased pronunciation variations for any given word has several undesired effects. First, it increases the confusability with other words because the difference in pronunciation between words usually becomes smaller. Second, it increases the search space during decoding since the decoder has to consider all possible pronunciations for each word.

A solution for trading off the benefit of multiple pronunciation variants to the aforementioned undesired effects is to use a pronunciation model [4]. In this paper, we introduce a novel pronunciation model that incorporates word context information to better model pronunciation variability in absence of any diacritics. The design of our proposed *context-dependent* pronunciation model is motivated by the fact that while reading Arabic text written without diacritics, an Arabic speaker chooses the pronunciation for each word based on neighboring words. We demonstrate the efficacy of this context-dependent pronunciation modeling with a series of experiments on the DARPA TRANSTAC Iraqi [5] corpus.

2. PRONUNCIATION MODELING

In order to formally introduce the pronunciation model for ASR, the formulation of the speech recognition problem is being presented first. The goal of speech recognition is to find the word sequence W^* that has the highest posterior probability, given the sequence of observations $X = \{x_1, \ldots, x_T\}$:

$$\begin{aligned} W^* &= \operatorname*{argmax}_W P(X|W)\,P(W) && (1)\\ &= \operatorname*{argmax}_W \sum_U P(X|U,W)\,P(U,W) && (2)\\ &\approx \operatorname*{argmax}_W \sum_U P(X|U)\,P(U|W)\,P(W) && (3) \end{aligned}$$

Here, $U = \{u_1, \ldots, u_M\}$ denotes the sequence of sub-word units, usually phonemes represented by a hidden Markov model (HMM). Equation 3 makes the assumption that the acoustic likelihood is independent of the word sequence given the phoneme sequence. To compute the product $P(X|U)P(U|W)P(W)$ we employ stochastic models of the acoustic and linguistic properties. Hence, the values

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

$P(X|U)$, $P(U|W)$ and $P(W)$ are provided from the acoustic model, the pronunciation model, and the language model (LM) respectively.

The pronunciation model provides a mapping between acoustic models U and the words W. Let $B = \{b_1, \ldots, b_N\}$ represent the sequence of phonemic representations (baseforms) for the word sequence $W = \{w_1, \ldots, w_N\}$. A pronunciation dictionary determines how the sequence of phonemes U are concatenated to form the baseform b_i of each word w_i. In case b_i has more than one phonemic representation then the word has multiple pronunciations. In general, the baseform pronunciations of a word are assumed to be independent of the word context, that is,

$$P(U|W) = P(B|W) \approx \prod_{i=1}^{N} P(b_i|w_i) \tag{4}$$

where b_i is the pronunciation given to the word w_i in the pronunciation sequence B.

3. CONTEXT-DEPENDENT PRONUNCIATION MODELING

While reading text without diacritics, an Arabic speaker chooses a pronunciation based on neighboring words. Figure 1 illustrates the process of inferring the short vowels from Arabic script that does not include diacritics. The top box shows an Arabic text that contains the words "*hw ktb*" (in Buckwalter romanization) without diacritics. The absence of short vowels in the example makes the pronunciation of the words ambiguous. The second word ("*ktb*") in the text can be diacritized in several ways (e.g. "*kataba*", "*kutub*", "*kutubu*"). For brevity, in this example we give two possible choices. However, when the word "*ktb*" follows the word "*hw*" then the vowelization in the bottom right box is incorrect whereas the vowelization in the bottom left box is correct. Therefore, the pronunciation of the words is really *context-dependent* [4, 6, 7].

Therefore, instead of using context-independent pronunciation probabilities, as shown in Equation 4, we use a n-gram word context pronunciation model:

$$P(U|W) = \prod_{i=1}^{N} P(b_i|w_i, \ldots, w_{i-n}) \tag{5}$$

In this paper we only consider a bigram word context, i.e. $P(b_i|w_i, w_{i-1})$. Since the pronunciation model uses bigram word context information we encounter the problem of data sparseness. To overcome data sparseness we smooth the context-dependent pronunciation probabilities with the context-independent pronunciation probabilities using the Witten-Bell smoothing [8]:

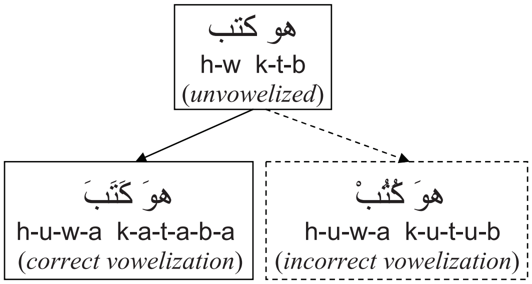

Fig. 1. Schematic example of vowelization. *Top*: Arabic script with no diacritics; *Bottom left*: Arabic script with the correct diacritics; *Bottom right*: Arabic script with incorrect diacritics. The pronunciations are shown below the Arabic script.

$$\begin{aligned} P(b_i|w_i, w_{i-1}) &= \lambda_{w_i,w_{i-1}} * P_{ML}(b_i|w_i, w_{i-1}) \\ &\quad + (1 - \lambda_{w_i,w_{i-1}}) * P(b_i|w_i) \end{aligned} \tag{6}$$

where $P_{ML}(b_i|w_i, w_{i-1})$ is the Maximum Likelihood (ML) estimate of the bigram pronunciation probability and

$$\lambda_{w_i,w_{i-1}} = \frac{N_{w_i,w_{i-1}}}{N_{w_i,w_{i-1}} + \sum_{b_i} c(b_i, w_i, w_{i-1})} \tag{7}$$

where $N_{w_i,w_{i-1}} = |\{b_i : c(b_i, w_i, w_{i-1}) > 0\}|$ is the number of unique pronunciations that follow the history (w_i, w_{i-1}). The term $P(b_i|w_i)$ in Equation 6 is the smoothed unigram pronunciation probability estimated by interpolating the ML estimate of the unigram pronunciation probability with the uniform pronunciation probability.

4. EXPERIMENTAL SETUP

4.1. Training and Test Data

The acoustic training consisted of 405 hours of Iraqi Arabic speech collected under the TRANSTAC effort. These include 1.5-way (simple answers to questions) and 2-way (full dialog) data collections for the force protection domain. Recognition experiments for reporting WER were performed on a held out validation set (Val), consisting of 16 hours (115K words). An additional test set was used for development (Dev), consisting of 16 hours (117K words). The Dev and Val sets are held-out sets randomly selected from the TRANSTAC training data.

4.2. System Architecture

We used a perceptual linear prediction (PLP) front-end, that computes 14 cepstral coefficients and normalized energy for each frame of speech. Phonetic word pronunciations were created using a set of 39 phonemes derived from graphemes [1]. The acoustic models were estimated in the ML framework. Bigram and trigram LMs were estimated using 2.8 million words of text. The decoding lexicon was restricted to 65K most frequent words in the acoustic training data.

Recognition was performed using our two-pass decoder. The forward pass uses a State Tied Mixture (STM) model, and an approximate bigram LM to produce a word lattice. The backward pass then uses the word lattice and associated scores from the forward pass to perform a detailed search a using within-word state-clustered tied-mixture (SCTM) quinphone acoustic model and a trigram language LM. The top scoring hypothesis represents the recognition output and an N-best or a lattice can also be produced.

Ideally, the context-dependent pronunciation model should be incorporated in the decoder itself. However, for our initial exploration in this paper, we employed the following N-best rescoring procedure:

1. For each hypothesis in the N-best list create a word lattice. The lattice has only one path and the number of arcs is the same as the number of the words in the hypothesis (Figure 2, *Top*).

2. Expand the lattice for multiple pronunciations of a word in the pronunciation dictionary. Each word in the expanded lattice is also tagged by its corresponding pronunciation (Figure 2, *Bottom*).

3. Rescore the lattice by relaxing the time boundaries by +/- 10 frames.

4. Extract the top-10 hypotheses from the lattice along with their acoustic score. The words are tagged by their pronunciation. Now, each original N-best hypothesis can create a maximum of 10 new hypotheses due to the multiple pronunciations of its words.

5. Compute the pronunciation score of each expanded N-best via Equation 5 using the pronunciation information that is attached to the words.

6. Compute the total score by summing the individual acoustic, pronunciation and language model scores. Note, that since every expanded N-best list has the same word sequence as the original hypothesis the language model scores remain the same. The top scoring hypothesis represents the recognition output.

Since the rescoring is performed on each arc separately, all expanded arcs of a word are being constrained by the time boundaries of the original hypothesis in the N-best list. While relaxing the time boundaries alleviates this constraint, the optimal approach for evaluating context-dependent pronunciation model is to incorporate it in the full search.

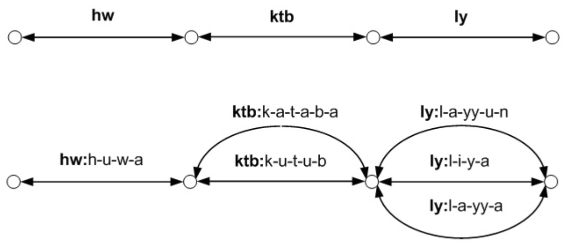

Fig. 2. Schematic example of lattice rescoring. *Top*: Lattice created from N-best hypothesis; *Bottom*: Expanded lattice for multiple pronunciations. Words in the arcs are being tagged by their corresponding pronunciation.

5. EXPERIMENTAL RESULTS AND CONCLUSIONS

We evaluated the context-dependent pronunciation model with two dictionaries that had the same vocabulary but different number of pronunciations per word. The vowelization method and creation of the pronunciations for the two dictionaries is discussed next.

The missing short vowels are added to the words in the vocabulary using two resources: the manually vowelized dictionary from Appen, Pty Ltd. (APPEN) and the BMA. The APPEN dictionary has only about 1.1 pronunciations per word. The Buckwalter morphological analyzer outputs all possible vowelizations of a word that is in its dictionary and produces an average of 5 pronunciations for each word.

Table 1 summarizes the characteristics of the dictionaries. The first dictionary (Dict1) primarily uses manually vowelized words from the APPEN dictionary. If a word is not found in the APPEN dictionary we use the automatically vowelized form from Buckwalter. The second (Dict2) uses the reverse procedure. Finally, several phonological rules are applied to the pronunciations in both dictionaries to create the final vowelized dictionaries [9]. Dict1 has an average of 2.5 pronunciations per word whereas Dict2 has 4.8.

The phonetic set consists of 62 speech phonemes (56 consonants, 6 vowels), in addition to “silence” and two other non-speech phones. This is to be compared to 39 phones for the grapheme system. Note that our phonetic system uses almost twice the number of consonants than other phonetic systems for Arabic dialects [2]. This is due to the treatment of the

Source	Dict1		Dict2	
	#wrds	#prons/wrd	#wrds	#prons/wrd
APPEN	30K	1.1	6K	1.1
Buckwalter	27K	4.9	51K	5.2
Total	57K	2.5	57K	4.8

Table 1. Number of words and average pronunciations per word from each source and for each dictionary.

Shadda diacritic that marks the gemination (doubling) of a consonant. When a consonant is marked by Shadda we use a different phoneme for the specific consonant.

The phonetic system was trained via the same procedure used for the baseline system as described in Section 4.2. No pronunciation model was used during training. Similarly, the baseline decoding experiments were carried out using the same decoding procedure previously described in Section 4.2. No pronunciation model was used during decoding for the baseline configuration. The output of the decoding was an N-best list with N=100.

The pronunciation model was trained over word sequences along with their phonetic sequence. The word and phoneme sequences were obtained by force-alignment of the reference transcripts of the training as defined in Section 4.1. The proposed pronunciation model was evaluated on the validation set defined in Section 4.1 using both Dict1 and Dict2.

Before we present the performance of the pronunciation model we return to the example of Figure 1. For brevity we only provide the pronunciations and probabilities for two choices. The context-independent or unigram pronunciation probabilities for the two pronunciations of the word “*ktb*” are:

$$\begin{aligned} P(b_i = \text{“k-a-t-a-b-a”} \,|\, w_i = \text{“ktb”}) &= 0.002 \\ P(b_i = \text{“k-u-t-u-b”} \,|\, w_i = \text{“ktb”}) &= 0.980 \end{aligned}$$

whereas the bigram context-dependent pronunciation probabilities are:

$$\begin{aligned} P(b_i = \text{“k-a-t-a-b-a”} \,|\, w_i = \text{“ktb”}, w_{i-1} = \text{“hw”}) &= 0.57 \\ P(b_i = \text{“k-u-t-u-b”} \,|\, w_i = \text{“ktb”}, w_{i-1} = \text{“hw”}) &= 0.28 \end{aligned}$$

As shown above, context-independent pronunciation probabilities fail to assign a low score to the incorrect pronunciation of the example shown in Figure 1 in the context of the previous word. On the other hand, the bigram pronunciation probabilities give higher score to the correct pronunciation. Note that the context-dependent probability for the incorrect pronunciation (i.e. “*kutub*”) is high relative to the probability of the correct pronunciation. This is a direct result of the smoothing mechanism (see Equation 6) since the prior probability of the context-independent pronunciation probability is almost 1. Nevertheless, the correct context-dependent pronunciation has higher probability than the incorrect one.

Table 2 compares the performance of the proposed context-dependent pronunciation model to the performance of the context-independent pronunciation model. For completeness we also report results with no pronunciation model. Furthermore, we report results from “best case” decoding experiments that used unigram and bigram pronunciation models trained on the “oracle” hypotheses. This last condition demonstrates the power of a well-trained pronunciation model. Finally, we include the performance of the grapheme system where the pronunciation of a word is based on its orthographic form.

Pronunciation Model	System		
	Grapheme	Phonetic Dict1	Phonetic Dict2
None	39.7	33.8	41.5
Unigram	-	33.6	41.0
Bigram	-	33.4	40.8
Unigram (oracle)	-	33.3	40.5
Bigram (oracle)	-	33.0	38.7

Table 2. Word Error Rate (%) results for Grapheme system and Phonetic systems using decoding dictionaries Dict1 and Dict2 on the Val set.

As shown in Table 2, when the decoding dictionary contains a large number of pronunciation variants, the context-dependent pronunciation model outperforms both context-independent pronunciation modeling and decoding without any pronunciation model. As one would expect the gains are modest when there are fewer pronunciation variants. Also, the oracle results indicate that the context-dependent modeling can benefit from additional training data. Given manual vowelization is time consuming, the likely concept of operations for acoustic modeling for Arabic dialects is the use of automatic vowelization. In such cases, context-dependent pronunciation modeling is clearly superior to context-independent pronunciation modeling.

6. REFERENCES

[1] J. Billa et al., “Audio indexing of Arabic broadcast news,” in *ICASSP*, 2002.

[2] M. Afify et al., “Recent progress in Arabic broadcast news transcription at BBN,” in *Eurospeech*, 2005.

[3] K. Kirchhoff et al., “Novel approaches to Arabic speech recognition: Report from the 2002 Johns Hopkins summer workshop,” in *ICASSP*, 2003.

[4] H. Strik and C. Cucchiarini, “Modeling pronunciation variation for ASR: A survey of the literature,” *Spch. Comm.*, vol. 29, pp. 225–246, 1999.

[5] D. Stallard et al., “Recent improvements and performance analysis of ASR and MT in a speech-to-speech translation system,” in *ICASSP*, 2008.

[6] E. Fosler-Lussier, “Contextual word and syllable pronunciation models,” in *IEEE Wkshp. Spch. Recog. & Und.*, 1999.

[7] D. Jurafsky et al., “The effect of language model probability on pronunciation reduction,” in *ICASSP*, 2001.

[8] I.H. Witten and T.C. Bell, “The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression,” *IEEE Trans. Inf. Thry.*, vol. 37, no. 4, pp. 1085–1094, 1991.

[9] B. Xiang et al., “Morphological decomposition for Arabic broadcast news transcription,” in *ICASSP*, 2006.