Explicit and Implicit Modeling of Short Vowels for Arabic STT*

Lori Lamel, Abdelkhalek Messaoudi and Jean-Luc Gauvain
Spoken Language Processing Group
CNRS-LIMSI, BP 133
91403 Orsay cedex, France
{lamel,abdel,gauvain}@limsi.fr

Abstract

This contribution reports on research aimed at improving the pronunciation and thereby, acoustic modeling, for automatic transcription of Arabic broadcast audio data by modeling short vowels and other diacritics. Being the link between the acoustic units and lexical ones, the pronunciation dictionary is a critical component of a speech recognition system. In previous work it was demonstrated that explicit modeling of short vowels improved recognition performance, even when producing non-vocalized hypotheses. The main challenge is dealing with incomplete information since Arabic is almost exclusively written without diacritics, which are needed for accurate grapheme-to-phoneme conversion. Training on audio data with non-vocalized transcripts is greatly facilitated by use of a generic vowel. Explicitly modeling some specificities of the Arabic language also improves recognition performance.

1 Introduction

Since Arabic texts are almost always written without diacritics, the main challenge in pronunciation modeling is to deal with incomplete information. A strongly consonantal language with nominally only three vowels, Arabic is a highly inflected language. There are typically multiple possible vocalizations for a given written word. Thus one of the challenges of explicitly modeling vowels in Arabic is to obtain vocalized resources, or to develop efficient ways to use non-vocalized data (Vergyri and Kirchhoff, 2004). Given a properly vocalized form, Arabic grapheme-to-phoneme conversion is (relatively) straightforward. The Buckwalter morphological analyzer (Buckwalter, 2002) is widely used to produce possible vocalizations. However, not all words are successfully processed by the Buckwalter morphological analyzer which mainly fails on proper names, technical words, and words in Arabic dialects. In the case of a large quantity of training data, there can easily be many words without vocalizations and determining vocalizations for these words manually or even semi-automatically can be prohibitive. Even though most Arabic speech recognizers output a non-vocalized word sequence, it has been shown that better recognition performance is obtained by explicitly modeling short vowels (Afify et al., 2005) than with a grapheme-based approach (Billa et al., 2002).

Therefore in order to simplify the problem of training on non-vocalized audio data, a generic vowel model has been introduced to hold the place of a short vowels. Rules were developed to automatically generate pronunciations from the non-vocalized word form. After applying the rules, each word has multiple pronunciations represented with consonants, long vowels, and the generic vowel. The absence of a vowel (the Sukoun) is modeled by removing a generic vowel. In addition to the explicit modeling of short vowels, extending the phone set to explicitly model gemination was found to improve recognition performance. The 'tanwin' is another grammatical mark which specifies that a noun

*This work was in part supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

is non-definite. The tanwin causes short vowels in word final position to be 'doubled', which is phonetically realized as adding an 'n' after the final vowel (also called nunation). As for the short vowels, the gemination and tanwin markers may be useful for translation or other downstream processing.

While the original purpose of the generic vowel models was to simplify training, in a second phase, the generic vowel was also used in the recognition lexicon. This allowed the inclusion in the lexicon of words that were previously excluded since they were not handled by Buckwalter. Including them was found to improve lexical coverage and reduce perplexity of development data.

2 Generating pronunciations with generic vowels

One possibility considered to process words that were not able to be processed with the Buckwalter morphological analyzer, was to allow all 3 short vowels or no vowel after every consonant. This idea was quickly rejected since there are too many possible vocalized forms. For example, with words with 4 consonants generate 512 possible pronunciations, and words with 8 consonants have 8192 possible pronunciations. In order to reduce the combinatorial explosion, a generic vowel was proposed to serve as a place holder for any of the three short vowels (Lamel et al., 2007). This does not pose any problem since even though short vowels are represented internally in the system, the Arabic recognizer outputs the non-vocalized word form. Using a generic vowel offers two main advantages. First, the manual work in dealing with words that are not handled by the Buckwalter morphological analyzer. These words can be automatically processed. Second, the number of vocalizations, and hence pronunciations, per word is greatly reduced (1 vowel instead of 3).

A set of detailed rules were used to generate pronunciations with a generic vowel from the non-vocalized word form. Some rules concern the word initial Alif (support of the Hamza), which can be stable or unstable. For the former case a pronunciation is generated with a glottal attack (denoted /'/) followed by a generic vowel (denoted /@/). These rules also cover word initial letter sequences [wAl, wbAl, wkAl, fAl, fbAl, fkAl] which often correspond to a composed prefix ending in "Al". Different pronunciations are generated to represent both situations. For example, the possible pronunciations for wAl are: /w@l/, /w'@l/, and /wAl/. In word final position, short vowels can be followed by an "n" (tanwin), so two forms are proposed, the generic vowel alone and the generic vowel followed by an "n".

Specific rules handle the pronunciation of words ending in "wA" and a final letter "p" (which symbolizes the ta marbouta). Within a word, a generic vowel is added after each consonant with the exception of the semivowels "w" and "y" which can be realized as semivowels or can serve as a support for the long vowels /U/ and /I/, respectively. Similarly a word internal Alif can represent the long vowel /A/ or a glottal attack.

After applying these rules, each word has multiple pronunciations represented with consonants, long vowels, and the generic vowel. Since vowels may also be absent (written with a Sukoun), additional pronunciations are generated by removing one generic vowel at a time. Using the rules the following two generic vowel forms are generated for the word "ktb".

ktb k@t@b@ k@t@b@n

which after allowing each generic vowel to be deleted produces:

ktb k@t@b@ k@t@b@n
kt@b@ kt@b@n
k@tb@ k@tb@n k@t@b

These generic vowel rules generate a large number of pronunciations per word: on average 35, compared to 6 per word with a vocalized form.

A test was done to assess the impact of using a generic model instead of 3 short vowels. All short vowels in the vocalized lexicon were mapped to a generic vowel. Acoustic models were trained by first mapping all short vowels to a single generic vowel (@), and training context dependent (CD) models with the standard consonant set and the single generic vowel. A pronunciation lexicon was created using the standard pronunciations with short vowels for the vocalized words and automatically

generated pronunciations with the generic vowel for the non-vocalized words. The audio data was then segmented using this lexicon with a combined set of acoustic models formed by merging the CD models with short vowels and those with a generic vowel. Note that since the basic idea was to use the generic vowel only in training, and not during recognition, a number of CD models are never used.

Two model sets were built with 5k tied states (64 Gaussians per state) and covering 5k phone contexts, one modeling 37 phones and the other 35 (the 3 short vowels being replaced by a single one). These models were tested using a single pass system without acoustic model adaptation. On broadcast news data (bnat06) there was about a 5% relative performance degradation using only the generic vowel. For broadcast conversation (bcat06) data, the relative loss was less than 1%. The smaller degradation on broadcast conversations may be due to a higher proportion of dialect or less formal data, for which there is a larger variability in the way that vowels are pronounced. These results provided an indication the generic vowel could be an effective means for facilitating training on non-vocalized data. All acoustic models used in the LIMSI GALE systems were built using dictionaries for which about 15% of the words have generic vowels in their pronunciation (Lamel et al., 2007).

3 Generic vowels in the STT lexicon

Since the generic vowel was found to be helpful in acoustic model training, a natural second step is to also use it in the recognition lexicon. When typical frequency-based methods are used select a recognition word list, a number of words may be excluded if no pronunciation is readily available. For example, when selecting a 200k word list, 36k words that should have been included, were excluded since they could not be vocalized by Buckwalter. As for the training data, these words were mostly proper names, technical terms or words in various Arabic dialects. Adding these 36k words to the 200k word list reduced the OOV rate by 0.2 to 0.9% absolute for different development data sets, showing the interest of including them. A simplified method for generating a limited number of vocalized forms (and thus pronunciations) for recognition was developed. The basic idea is to put a generic vowel after each consonant and to map all semivowels to a long vowel. This corresponds to the manner in which foreign names and technical terms are spoken. The distribution of the number of pronunciations per word automatically generated in this manner is shown in Table 1. About 10% of the words have a single pronunciation, and over 50% have only one or two. Only 2% of the words have more than 10 forms, with the maximum being 16, and all of these words (1.7%) include the prefix 'Al'.

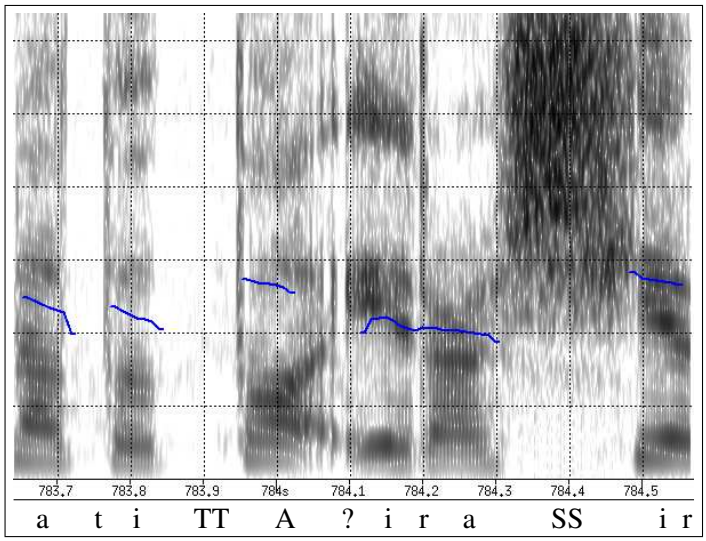
#pronunciations	% words
1	9.6%
2	42.7%
3	20.4%
$\leq$10	98.0%
16	1.7%

Table 1: Distribution of the number of pronunciations with generic vowels for the 36k words in the extended 236k word list.

Using the extended 236k wordlist reduced the word error rate by 0.2% absolute using a single decoding pass without cluster-based adaptation on the eval06 data set. In the reference transcripts, there were 162 occurrences of one of the 36k word forms, accounting for 0.6% of all word tokens. Thus one third of the possible gain was obtained.

4 Modeling Geminates and Tanwin

The initial phone set for Arabic contained 37 symbols: 28 Arabic consonants, 3 foreign consonants, and 6 vowels (i,a,u short and long) (Messaoudi et al., 2005). When pronunciations were generated with this phone set, all consonants with a gemination mark were simply doubled. While this may be a reasonable approximation for some sounds, such as fricatives, it is clearly not well adapted to plosives where gemination does not result in multiple bursts. Another sources of gemination arises is a contextual variant of the definite article 'Al' ('the'). When the 'Al' precedes a lunar consonant it is usually pronounced as /al/. When the 'Al' precedes a solar consonant it is usually silent, but transforms the following consonant into a geminate. Generally speaking all of the Arabic consonants can oc-



Solar	
y	50.5%
$	26.9%
S	20.1%
p	19.1%
v	19.0%
Z	18.4%
d	15.3%
s	14.1%
t	9.2%
n	8.9%
T	8.7%
r	8.3%
z	7.4%
l	6.4%
D	4.9%
g	2.6%

Lunar	
k	6.6%
w	4.7%
m	4.2%
q	3.1%
j	2.7%
G	2.5%
b	2.1%
x	1.8%
H	1.2%
c	0.2%
J	1.1%
f	1.7%
h	0.5%
’	0.0%
V	0.0%

Figure 1: **Left:** Spectrogram illustrating gemination. **Right:** Occurrences (%) of geminates for Solar and Lunar consonants.

cur as singletons or geminates. The left part of Figure 1 illustrates a portion of the phrase “(kaAn)ati AlT∼aA}irap Al$∼ir(aAEiyap)”. An aligned approximate phone transcription is shown on the bottom. The sukoun indicating that the vowel following the first t is not pronounced is transformed to a short /i/ because it precedes an ’Al’. The ’Al’ in turn precedes a solar consonant so it is not pronounced but causes the T (emphatic t) to be geminated. The short /i/ (around time 783.8) is realized as a schwa-like vowel. Another geminate ’sh’ (SS) is centered at time 784.4.

In order to more precisely model such phenomena, an alternative phone set was explored in which an additional 30 phone symbols were added to represent the geminate phones. The frequencies of the consonants in single and geminate form were counted in a 100 hour corpus of manually transcribed and vocalized Arabic broadcast news data (Messaoudi et al., 2005). The right part of Figure 1 lists the solar and lunar consonants, along with the percentage of occurrences as geminates. It can be observed that the solar consonants generally have a higher proportion of geminates than the lunar ones.

Figure 2 shows how the geminates are represented in the original pronunciation dictionary (top) and the dictionary with specific geminate symbols. The left column gives the transliterated lexical entry. Associated with each entry are several pairs of vocalized forms and their pronunciations.

To compare the original phone set and the extended one with geminates, acoustic models were trained on a corpus of about 1000 hours Arabic broadcast data using the both. They were tested on the broadcast news (bnat06) and broadcast conversation (bcat06) data, each containing about 3-hours of audio data (Lamel et al., 2007). For both data types using specific phones to model geminates gave a small performance improvement, and a further gain was obtained by combining the two models. Increasing the phone set also has the advantage of increasing the number of context-dependent phones that are modeled.

The ’tanwin’ is a grammatical mark which specifies that a noun is non-definite. The tanwin causes short vowels in word final position to be ’doubled’, which is phonetically realized as adding an ’n’ after the final vowel (also called nunation). The tanwin can be realized as a vowel-n sequence or a nasalized vowel, or some combination. In an attempt to better capture this variability, three phones were added to the phone set representing the three tanwin phones (in, an, un) with a single unit. Acoustic models were built using this new phone set, and tested on the de-

Lexical Entry	*Pronunciation = Transliteration*			
ktAb	kitAb=**kitaAb**	kitAba=**kitaAba**	kuttAb=**kut˜aAb**	kuttAba=**kut˜aAba**
	kitAbi=**kitaAbi**	kitAbin=**kitaAbK**	kuttAbi=**kut˜aAbi**	kuttAbin=**kut˜aAbK**
	kitAbu=**kitaAbu**	kitAbun=**kitaAbN**	kuttAbu=**kut˜aAbu**	kuttAbun=**kut˜aAbN**
ktAb	kitAb	kitAba	ku+Ab	ku+Aba
(with	kitAbi	kitAbin	ku+Abi	ku+Abin
geminates)	kitAbu	kitAbun	ku+Abu	ku+Abun

Figure 2: Sample pronunciations for ktb in the dictionary without (top) and with (bottom) geminate symbols. The lexical entry is the non-vocalized word class encompassing all possible vocalized forms. On the left of the equal sign is the phonemic form and with the corresponding transliterated graphemic form on the right. The transliterated form is not repeated in the bottom part of the figure.

velopment data sets. Comparable word error rates were obtained with these models and the non-tanwin models, and an absolute gain of 0.4% was achieved with system combination. Contrastive models were built permitting multiple forms for tanwin, both as a single phone and as a short vowel-n sequence. These models had comparable performance as the previous model sets. Given the large variability in the realization of tanwin, these results are not surprising and suggest that any reasonable, coherent representation for tanwin will suffice.

5 Conclusions

This paper has summarized recent advances pronunciation and acoustic modeling to explicitly take into account some of the specificities of the Arabic language. One of the challenges is training with incomplete information since most Arabic texts are written without diacritics, yet the diacritics provide useful information for pronunciation modeling and higher level processing (**?**). Many vocalized word forms can be derived using the Buckwalter morphological analyzer and modifications thereof. In order to enable training for words that Buckwalter is not able to process, rules to generate pronunciations with a generic vowel have been proposed. This method has been used to significantly facilitate training on non-vocalized data, and all of the acoustic models used in the LIMSI GALE systems were trained with them. Pronunciations for 16% of the words in the training dictionary and 11% in the recognition dictionary have a generic vowel. Performance gains were also obtained by explicit modeling of geminate consonants and pronunciation variants of the definite article 'Al'. Some other variants have also improved performance, such as modeling the g-sound in Egyptian Arabic, the tendency for a final /a/ to be realized as a front vowel /i/ in Lebanese data, and the trend to have a Sukoun as the final vowel in spoken Arabic, in particular for dialectal speech.

References

Mohamed Afify, Long Nguyen, Bing Xiang, Sherif Abdou, John Makhoul 2005. Recent Progress in Arabic Broadcast News Transcription at BBN *InterSpeech Eurospeech'05*, 1637-1640, Lisbon.

Jayadev Billa, Mohammed Noamany, Amit Srivastava, Daben Liu, Rebecca Stone, Jinxi Xu, John Makhoul, Francis Kubala. 2002. Audio Indexing of Arabic Broadcast News. *IEEE International Conference on Acoustics, Speech, & Signal Processing*, **1**:5-8.

Tim Buckwalter. 2002. Buckwalter Arabic Morphological Analyzer (LDC2002L49, LDC2004L02). http://www.qamus.org/morphology.htm Linguistic Data Consortium, Philadelphia.

Lori Lamel, Abdelkhalek Messaoudi, Jean-Luc Gauvain. 2007. Improved Acoustic Modeling for Transcribing Arabic Broadcast Data *InterSpeech Eurospeech'07*, 2077-2080, Antwerp.

Abdelkhalek Messaoudi, Jean-Luc Gauvain, Lori Lamel. 2005. Modeling Vowels for Arabic BN Transcription *InterSpeech Eurospeech'05*, 1633-1636, Lisbon.

Abdelkhalek Messaoudi, Jean-Luc Gauvain, Lori Lamel. 2006. Arabic Broadcast News Transcription using a One Million Word Vocalized Vocabulary. *IEEE International Conference on Acoustics, Speech, & Signal Processing*, **I**-1093-1096, Toulouse.

Dimitra Vergyri and Katrin Kirchhoff. 2004. Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition *COLING Workshop on Arabic-script Based Languages,* 66-73, Geneva.