Improved Acoustic Modeling for Transcribing Arabic Broadcast Data [*]

Lori Lamel, Abdel. Messaoudi and Jean-Luc Gauvain

Spoken Language Processing Group
CNRS-LIMSI, BP 133
91403 Orsay cedex, France
{lamel,abdel,gauvain}@limsi.fr

10.21437/Interspeech.2007-562

ABSTRACT

This paper summarizes our recent progress in improving the automatic transcription of Arabic broadcast audio data, and some efforts to address the challenges of the broadcast conversational speech. Our efforts are aimed at improving the acoustic, pronunciation and language models taking into account specificities of the Arabic language. In previous work we demonstrated that explicit modeling of short vowels improved recognition performance, even when producing non-vocalized hypotheses. In addition to modeling short vowels, consonant gemination and nunation are now explicitly modeled, alternative pronunciations have been introduced to better represent dialectical variants, and a duration model has been integrated. In order to facilitate training on Arabic audio data with non-vocalized transcripts a generic vowel model has been introduced. Compared with the previous system (used in the 2006 GALE evaluation) the relative word error rate has been reduced by over 10%.

Index Terms – Speech recognition, Arabic, broadcast news, broadcast conversations

1. INTRODUCTION

This paper summarizes recent progress at LIMSI carried out in the context of the DARPA GALE program to improve the automatic transcription of Arabic broadcast audio data. Speech recognition and machine translation are key supporting technologies for the GALE program (www.darpa.mil/ipto/Programs/gale). In addition to the automatic processing of Broadcast News data, a new challenge is the transcription of broadcast conversational speech, such as radio and television talk shows, debates, and interactive programs where the general public are invited to participate in the discussion by telephone. This type of data requires the explicit modeling of spontaneous speech effects, much more common than in broadcast news, and also the ability to deal with speech from a variety of Arabic dialects.

While research is underway to improve all aspects of the models used in the LIMSI Arabic system, this paper focuses on improvements in the acoustic and pronunciation models. One primary consideration is to explicitly model more of specificities of the Arabic language. It has been recently shown that even when producing a non-vocalized transcript, explicitly modeling short vowels improves recognition performance [2] over a grapheme-based approach where only characters in the non-vocalized written form are modeled [3]. We previously demonstrated that by building a very large vocalized vocabulary of more than 1.2 million words, and by using a language model including a vocalized component, the word error rate can be significantly reduced [11]. While pronunciation generation in Arabic from vocalized texts is often considered straightforward there are several rules that modify the pronunciations. One of the frequent variants is the pronunciation 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 (the consonant is 'doubled'). Generally speaking all of the Arabic consonants can occur as singletons or geminates. The 'tanwin' is another grammatical mark which specifies that that noun is to be intended 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 referred to as nunation). These studies aim to improve the acoustic and lexical models by explicitly representing the gemination and tanwin. One obvious way to improve the acoustic models is to train them on more data. Since untranscribed data are easily available, we make use of unsupervised methods to train with these [7].

2. TRAINING WITH GENERIC VOWELS

Generally speaking, extending the pronunciation dictionary to include entries for additional training data entails some manual intervention or verification. For Arabic, the difficulty lies in determining the vocalized forms for the new words, after which grapheme-to-phoneme conversion is (relatively) straightforward. In the case of a large quantity of training data with non-vocalized transcripts there can be too many words without vocalizations to add these manually or even semi-automatically. One possibility that we considered was to generate all possible vocalized forms, allowing 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.

In order to simplify the problem, we investigated the use of a generic vowel to replace the three short vowels. 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

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

with words that are not handled by the Buckwalter morphological analyzer (typically proper names, technical words, words in Arabic dialects) is reduced. With this approach these 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 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”.

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 added by removing one generic vowel at a time. For example, the rules generate the following two generic vowel forms 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

It should be noted that the diacritic for gemination has not been taken into account when generating pronunciations with generic vowels. These decision was taken to limit the number of pronunciations even though the gemination is explicitly represented for most words in the lexicon. In the current system words with generic vowels are not included in the recognition word list, and are only used during training.

An experiment was carried out to assess the quality of acoustic models with generic vowels by mapping all short vowels in the vocalized lexicon to a generic vowel. Acoustic models were retrained by first mapping all short vowels to a single generic vowel (@), and training context dependent models with the standard consonant set and the single generic vowel. A pronunciation lexicon was then created that used the standard pronunciations with short vowels for the vocalized words and automatically generated pronunciations with the generic vowel for the non-vocalized words. We then segmented all of the audio data 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 the basic idea was to use the generic vowel only in training, but not during recognition so a number of CD models are never used. In the future we may consider also extending the recognition lexicon in an analogous manner. In order to assess the feasibility of this, several model sets were built and tested in decoding using only a generic vowel.

Recognition word error rates with a single pass system (corresponding to the first pass of the evaluation system described below) are given in Table 1 with the standard phone set including 3 short vowels, and with models trained with only one generic short vowel. Both model sets have 5k tied states (64 Gaussians per state) and covering 5k phone contexts. It can be seen that there is only a slight degradation in performance when using a generic vowel. Therefore it was decided that the generic vowels provide an effective means to facilite training on non-vocalized data.

	bnat06	bcat06
Standard model	24.4%	35.2%
Generic model	25.7%	35.5%

Table 1: Word error rates on GALE broadcast news (bnat06) and broadcast conversation (bcat06) development data with scoring with small acoustic models, representing 3 short vowels or 1 generic vowel.

3. MODELING GEMINATES AND TANWIN

The component Arabic STT system used in the GALE’06 evaluation has three decoding passes, where each decoding pass generates a word lattice with cross-word, position-dependent, gender-dependent AMs, followed by consensus decoding with 4-gram and pronunciation probabilities [5, 11]. Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR [8] techniques prior to each decoding pass. The lattices of the last two decoding pass are rescored by the neural network LM interpolated with a 4-gram backoff LM [12] trained on the audio transcripts and the Arabic gigaword corpus [1].

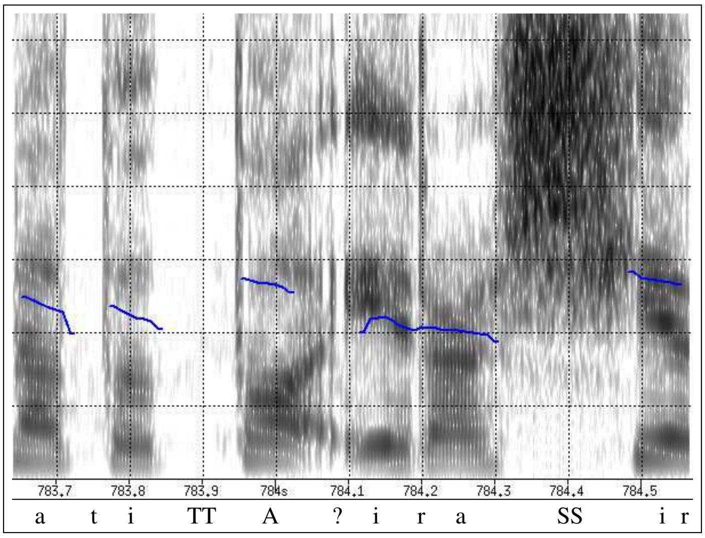
The original phone set contains 37 symbols: 28 Arabic consonants, 3 foreign consonants, 6 vowels (i,a,u short and long). When pronunciations were determined with this phone set, all consonants with a gemination mark were simply doubled. While this may be a reasonable approximation for some sounds, such as fricatives, if is clearly not well adapted to plosives where gemination does not result in multiple bursts.

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 is preceding 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 reduced to a schwa-like vowel. Another geminate ’sh’ is centered at time 784.4.

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 [10]. 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 new dictionary with specific geminate symbols.

Acoustic models were trained on a large corpus of about 1000 hours Arabic broadcast data mostly from the GALE program using the both the original phone set and the extended one which includes geminates. Recognition results are given in Table 2 on two 3-hour sets of development data used in the GALE community. It can be seen that modeling geminates improves



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:** Percentage occurrences of geminates for Solar and Lunar consonants.

ktAb	kitAb=**kitaAb** kitAba=**kitaAba**
	kitAbi=**kitaAbi** kitAbin=**kitaAbK**
	kitAbu=**kitaAbu** kitAbun=**kitaAbN**
	kuttAb=**kut˜aAb** kuttAba=**kut˜aAba**
	kuttAbi=**kut˜aAbi** kuttAbin=**kut˜aAbN**
	kuttAbu=**kut˜aAbu** kuttAbun=**kut˜aAbN**

ktAb	kitAb=**kitaAb** kitAba=**kitaAba**
	kitAbi=**kitaAbi** kitAbin=**kitaAbK**
	kitAbu=**kitaAbu** kitAbun=**kitaAbN**
	ku+Ab=**kut˜aAb** ku+Aba=**kut˜aAba**
	ku+Abi=**kut˜aAbi** ku+Abin=**kut˜aAbK**
	ku+Abu=**kut˜aAbu** ku+Abun=**kut˜aAbN**

Figure 2: Sample pronunciations for ktb in the original dictionary (top) and with geminate symbols (bottom). Each lexical entry is the non-vocalized word class encompassing all possible vocalized forms.

	bnat06	bcat06
Standard model	22.0%	32.6%
Geminate model	21.7%	32.3%
Combination	21.5%	31.9%

Table 2: Word error rates without and with explicit modeling of geminates on the GALE 2000 development data sets. bnat06: broadcast news, bcat06: broadcast conversations.

performance for both the broadcast news (bnat06) and broadcast conversation (bcat06) data types, and that a further gain is obtained by combining the two models. Increasing the phone set also has the added advantage of increasing the number of context-dependent phones that are modeled.

As mentioned above, final short vowels are followed by /n/ for indefinite word forms. These can be realized as a vowel-n sequence or a nasalized vowel. In order to better capture this variability three additional phones were added to the phone set to represent the three tanwin phones (in, an, un) with a single unit. Acoustic models were built using this new phone set, and tested on the development data sets. These models obtained word error rates comparable to that of the non-tanwin models, and when used in system combination gave a gain of 0.4% absolute. Given the large variability in the realization of tanwin, these results are not surprising. Further model refinement is underway and an a contrastive experiment permitting multiple forms for tanwin as both a single phone or a sequence of a short vowel-n will be carried out.

4. PRONUNCIATION VARIANTS

An error analysis on the bnat06 data was carried out for data the Arabic system evaluated in the GALE 2006 evaluation. The main source of errors involves the insertion or deletion of a prefix or a suffix, such as the confusion of ktAb/wktAb or ktAbh/ktAb. The article ’Al’ is found in 37% of the prefix errors (many prefixes end in Al), and contributes an absolute error of 1%. In examining the errors a number of dialectal pronunciation variants were observed, that were not represented in the lexicon. Figure 3 shows two spectrograms of the word ’aljalsa (meeting). The final short vowel in the example on the left is an /a/. The right example is the same word, but the final vowel is not produced in the same manner. Arabic speakers consider this to be an /i/, whereas it appears more like an /e/ in the spectrogram.

Alternative variants were systematically added to the pronunciation lexicon, and an absolute performance improvement of 0.3% was obtained on the broadcast news dev data and 0.6% on the broadcast conversation dev data.

5. DURATION MODELING

It is well known that HMMs are not properly modeling the phone and the word durations. The segment duration being implicitly encoded in the model topology, the transition probabilities, and the derivative features, none of these model parameters can properly capture segment duration when considering a wider context than a triphone. Even though phone and word durations do not appear to be discriminative features in English, it may be worth looking to this issue for Arabic in combination with the geminate models.

The adopted strategy is to add duration information as a post-processing of the decoding process, but instead of applying such

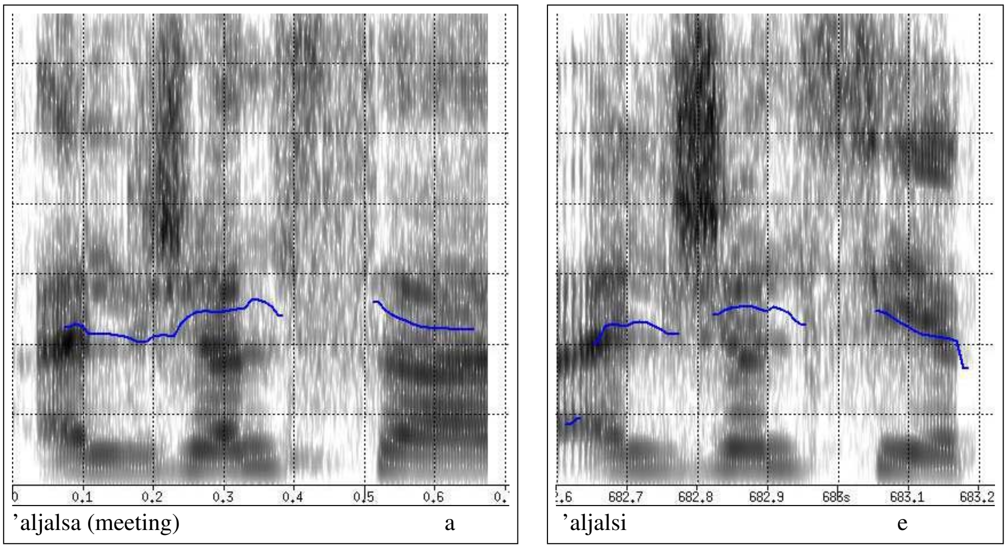


Figure 3: Example spectrogram of the word 'aljalsa (meeting) illustrating the Lebanese realization of the final vowel.

	bnat06	bcat06
Standard model	19.7%	28.5%
+ Duration model	19.5%	28.1%

Table 3: Word error rates on GALE broadcast news (bnat06) and broadcast conversation (bcat06) development data with and without the duration model.

post-processing to an N-best list as it often done, here a word lattice representation which also includes the phone segmentation for each word edge is used. For each hypothesis, the augmented likelihood is the product of the HMM likelihood and the duration likelihood properly scaled. As proposed by SRI [4], phone and word durations are modeled with Gaussian mixtures, using word duration (seen as a vector of phones) when enough data is available to properly estimate it, and backing off to phone durations if it is not the case. This approach allows duration information to be used in conjunction with consensus decoding [9] as proposed in [6]. Recognition results with the best Arabic system configuration are given without and with the duration model in Table 3. The duration model is seen to give 0.2-0.4 absolute gain on the GALE development data.

6. CONCLUSIONS

In this paper we have summarized recent advances in acoustic and pronunciation modeling for an Arabic broadcast data transcription system. One of the new challenges addressed is the recognition of broadcast conversational speech, for which word error rates are about 50% higher than for broadcast news. Overall the word error rates have been reduced by over 10% from the LIMSI component used in the GALE 2006 evaluaion. The improvements arise from several factors only some of which were discussed here, in particular the explcit modeling of specificities of the Arabic language. Other improvements include a revised training procedure, the use of partial supervision during training, the use of multiple data partitioners and the incorporation of a connectionist language model. Modeling of spontaneous speech and improved‘ lexical modeling for dialectal Arabic is an ongoing activity.

REFERENCES

[1] Linguistic Data Consortium. The Arabic Gigaword corpus (LDC2003T12), 2003.

[2] M. Afify, L. Nguyen, B. Xiang, S. Abdou, J. Makhoul, "Recent Progress in Arabic Broadcast News Transcription at BBN" *InterSpeech Eurospeech'05*, 1637-1640, Lisbon, September 2005.

[3] J. Billa, N. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, F. Kubala, "Audio Indexing of Arabic Broadcast News," *ICASSP'02*, **1**:5-8, Apr 2002.

[4] V.R.R. Gadde, "Modeling Word Duration," *Proc. 6th International Conference on Spoken Language Processing (ICSLP)*, **1**:601-604, 2000.

[5] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**(1-2):89-108, May 2002.

[6] N. Jennequin and J.L. Gauvain, "Modeling Duration Via Lattice Rescoring," *ICASSP-07*, Honolulu, April 2007.

[7] L. Lamel, J.L. Gauvain, G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer, Speech and Language*, **16**(1):115-229, January 2002.

[8] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2):171-185, 1995.

[9] L. Mangu, E. Brill, A. Stolcke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *ISCA EuroSpeech'99*, 495-498, Budapest, September 1999.

[10] A. Messaoudi, J.L. Gauvain, L. Lamel, "Modeling Vowels for Arabic BN Transcription," *Eurospeech'05*, 1633-1636, Lisbon, September 2005.

[11] A. Messaoudi, J.L. Gauvain, L. Lamel, "Arabic Broadcast News Transcription using a One Million Word Vocalized Vocabulary," *IEEE ICASSP'06*, **I**-1093-1096, Toulouse, May 2006.

[12] H. Schwenk, J.L. Gauvain, "Training Neural Network Language Models On Very Large Corpora," *HLT/EMNLP*, 201-208, Vancouver, October 2005.