Transcription of Arabic Broadcast News

Abdel. Messaoudi,[*†] *Lori Lamel, Jean-Luc Gauvain*
Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{abdel,lamel,gauvain}@limsi.fr

ABSTRACT

This paper describes recent research on transcribing Modern Standard Arabic broadcast news data. The Arabic language presents a number of challenges for speech recognition, arising in part from the significant differences in the spoken and written forms, in particular the conventional form of texts being non-vowelized. Arabic is a highly inflected language where articles and affixes are added to roots in order to change the word's meaning. A corpus of 50 hours of audio data from 7 television and radio sources and 200 M words of newspaper texts were used to train the acoustic and language models. The transcription system based on these models and a vowelized dictionary obtains an average word error rate on a test set comprised of 12 hours of test data from 8 sources is about 18%.

1. INTRODUCTION

This paper describes some out work in developing a broadcast news transcription system for the Modern Standard Arabic. By Modern Standard Arabic we refer to the spoken version of the official written language, which is spoken in much of the Middle East and North Africa, as is used in major broadcast news shows. At LIMSI we have found that porting a broadcast news system developed for American English to several other languages was quite straightforward if the required resources are available. Our observation is that given a similar quantity and quality of linguistic resources (audio data, language model training texts, and a consistent pronunciation lexicon) somewhat comparable recognition accuracies results can be obtained in different languages [5].

The Arabic language poses challenges somewhat different from the other languages (mostly Indo-European Germanic or Romance) we have worked with. Modern Standard Arabic is that which is learned in school, used in most newspapers and is considered to be the official language in most Arabic speaking countries. [1] Arabic texts are written and read from right-to-left and the vowels are generally not indicated. It is a strongly consonantal language with nominally only three vowels, each of which has a long and short form. Arabic is a highly inflected language, and as a result has many different word forms for a given root, produced by appending articles at the word beginning "the, and, to, from, with, ...") and possessives ("ours, theirs, ...") at the word end. The different right-to-left nature of the Arabic texts required modification to the text processing utilities. Texts are non-vowelized, meaning the short vowels and gemination are not indicated. There are typically several possible (generally semantically linked) vowelizations for a given written word, and the word final vowel varies as a function of the word context. For most written texts it is necessary to understand the text in order to know how to vowelize and pronounce it correctly.

2. ARABIC LANGUAGE RESOURCES

Three types of Arabic resources were created for this work: an audio corpus containing over 50 hours of radio and television broadcast news data; a corpus of text materials from various newspaper sources; and a pronunciation lexicon.

2.1 Audio data

Audio data from 5 radio and 2 television sources were collected during the period from September 1999 through October 2000 for training purposes. A summary of the amount of data for each source is given in Table 1. Most of the data are from Radio Orient, an Arabic station that broadcasts directly in France. The television and other radio sources were recorded via satellite (Arabsat). The test data is comprised of 7 hours of Radio Qatar recorded during September and October 2000 and 5 hours of more recent data from 8 sources recorded in 2001 and 2002.

The audio data were manually transcribed using an Arabic version of Transcriber [1] and an Arabic keyboard. The manual transcriptions are vowelized, enabling accurate modeling of the short vowels, even though these are not usually present in written texts. This is different from the approach taken by Billa et al. [2] where only characters in the non-vowelized written form are modeled. Each Arabic character, including short vowel and geminate markers, is transliterated to a single ascii character. Transcription conventions were developed to provide guidance for marking vowels and dealing with inflections and gemination, as well as to consistently transcribe foreign words, in particular for proper names and places, which are quite common in Arabic broadcast news. The foreign words can have a variety of spoken

[*†] Visiting scientist from the Vecsys Company.

[1] In contrast many people speak in dialects for which there is only a spoken from and no recognized written form.

Training data		
Source	Origin, Date	Duration
Radio Elsharq	Syria, 2000	1h30
Radio Kuwait	Kuwait, 2000	2h15
Radio Orient	Paris, 1999-2000	30h
Radio Qatar	Qatar, 2000	5h
Radio Syria	Syria, 2000	6h15
TV Aljazeera	Qatar, 2000	5h15
TV Syria	Syria, 2000	2h

Test data		
Source	Origin, Date	Duration
Radio Qatar	Qatar, 2000	7h
Radio Qatar	Qatar, 2001	32mn
Radio Kuwait	Kuwait, 2001	25mn
Radio BBCA2	London, 2001	47mn
Radio Medi1	Morocco, 2002	43mn
TV Syria	Syria, 2001	22mn
TV Aljazeera	Qatar, 2002	43mn
TV ESC	Egypt, 2002	57mn
TV 7	Tunisia, 2002	48mn

Table 1: Composition of the Arabic broadcast news audio corpus. The training data date from Sep'99 through Oct'00. The test data were recorded in Sep-Oct'00, and Apr'01 to Dec'02.

realizations depending upon the speaker's knowledge of the language of origin and how well-known the particular word is to the target audience. The transcripts contain a total of 320k words, of which 57k are distinct.

2.2 Text data

The written resources consist of over 200 million words of texts from six newspaper sources, with most of the data coming from the years 1998-2000, and early 2001: Addustour, Ahram, Albayan, Alhayat, Al-Watan, Raya. With the exception of the 1998 Alhayat texts which were available on CDROM, the texts were obtained from the Internet. The texts were preprocessed to remove undesirable material (tables, lists, punctuation markers) and transliterated using an slightly extended version of Buckwalter transliteration[2] from the original Arabic script form to improve readability.

The texts were then further processed for use in language model training. First the texts were segmented into sentences, and then normalized in order to better approximate a spoken form. Common typographical errors were also corrected. The main normalization steps are similar to those used for processing texts in the other languages [3, 5]. They consist primarily of rules to expand numerical expressions and abbreviations (*km, kg, m2*), and the treatment of acronyms (*A. F. B.* $\rightarrow$ *A_F_B*). A frequent problem when processing numbers is the use of an incorrect (but very similar) character in place of the comma (*20r3* $\rightarrow$ *20,3*). The most frequent errors that were corrected were: a missing Hamza above or below an Alif; missing (or extra diacritic marks) at word ends:

below y (eg. Alif maksoura)
above h (eg. t marbouta),

and missing or erroneous interword spacing, where either two words were glued together or the final letter of a word was glued to the next word. After processing there were a total of 204 million words, of which 1.4 M are distinct.

[2] T. Buckwalter, http://www.qamus.org/transliteration.htm

Vowelized lexicon	
kitaAb	kitAb
kitaAba	kitAba
kitaAbi	kitAbi
kut˜aAbi	kuttAbi

Non-Vowelized lexicon	
ktAb	kitAb=kitaAb
	kitAba=kitaAba
	kitAbi=kitaAbi
	kuttAbi=kut˜aAbi
sbEyn	sabEIna=saboEiyna
	sabEIn=saboEiyn

Figure 1: Example lexical entries for the vowelized and non-vowelized pronunciation lexicons. In the non-vowelized lexicon, the pronunciation is on the left of the equal sign and the written form on the right.

2.3 Pronunciation lexicon

Letter to sound conversion is quite straightforward when starting from vowelized texts. A grapheme-to-phoneme conversion tool was developed using a set of 37 phonemes and three non-linguistic units (silence/noise, hesitation, breath). The phonemes include the 28 Arabic consonants (including the emphatic consonants and the hamza), 3 foreign consonants (/p,v,g/), and 6 vowels (short and long /i/, /a/, /u/). In the 57k word pronunciation lexicon each vowelized orthographic form of a word is treated as a distinct lexical entry. The example entries for the word "kitaAb" are shown in the top part of Figure 1. An alternative representation uses the non-vowelized orthographic form as the entry, allowing multiple pronunciations, each being associated with a particular written form. Each entry can be thought of as a word class, containing all observed (or even all possible) vowelized forms of the word. The pronunciation is on the left of the equal sign and the vowelized written form is on the right. If this representation is chosen, then 33k distinct orthographic forms cover the original 57k lexicon.

The out-of-vocabulary (OOV) rate is very high with the 57k word lexicon, on the order of 15%. This is reduced to about 8% in the 33k word lexicon since all possible vowelized forms are matched for each entry. The latter lexical representation is also needed in order to be able to increase the lexical coverage by using the text materials which are unvowelized. The lexicon was extended to 60k words, by choosing the most frequent words in the normalized texts. The OOV rate of the 60k word list is 4.1%. Since multiple vowelized forms are associated with each non-vowelized word entry, an online morphological analyzer was used to propose possible forms that were then manually verified. The morphological analyzer was also applied to the original 33k word list in order to propose forms that did not occur in the training data. A subset of the words, mostly proper names and technical terms, were manually vowelized. The 60k non-vowelized words result in about 140k distinct vowelized forms.

3. RECOGNITION SYSTEM OVERVIEW

The LIMSI broadcast news transcription system has two main components, an audio partitioner and a word recognizer. Data partitioning serves to divide the continuous stream of acoustic data into homogeneous segments, associating appropriate labels with the segments. The segmentation and labeling process [3] first detects and rejects non-speech segments, and then applies an iterative maximum likelihood segmentation/clustering procedure on the speech segments. The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels.

For each speech segment, the word recognizer determines the sequence of words in the segment, associating start and end times and an optional confidence measure with each word. The speech recognizer makes use of continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on large text corpora for language modeling. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained by means of a decision tree. The acoustic feature vector has 39-components comprised of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives [3].

Word recognition is performed in two steps [4]. The first decoding pass generates initial hypotheses, which are used to carry out cluster-based acoustic model adaptation of both the means and variances using the MLLR technique [6]. Acoustic model adaptation is quite important for reducing the word error rate, with relative gains on the order of 20%. Experiments indicate that the word error rate of the first pass is not critical for adaptation. Then word lattices are generated using a 2-gram LM and rescored with a 3-gram or a 4-gram LM after conversion to a consensus network. The decoder was modified to handle the new style lexicon in order to produce the vowelized orthographic form associated with each word hypothesis (instead of the non-vowelized word class). Decoding is carried out in about 5 times real-time.

3.1 Acoustic models

The acoustic models are context-dependent, 3-state left-to-right hidden Markov models with Gaussian mixture. Gender-dependent, position-dependent triphones are estimated using MAP adaptation of SI seed models for wideband and telephone band speech. The first decoding pass uses a small set of acoustic models with about 5400 contexts and tied states. The second pass acoustic models cover about 11000 phone contexts represented with a total of 10000 states, and 16 Gaussians per state. State-tying is carried out via divisive decision tree clustering, constructing one tree for each state of each phone so as to maximize the likelihood of the training data using single Gaussian state models, penalized by the number of tied-states [3]. A set of 152 questions concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones.

3.2 Language models

Fourgram language models are obtained by interpolation of backoff n-gram language models trained different subsets of text corpora. Three different trigram language models were estimated corresponding to the three recognition lexicons. The first LM was estimated on the 320k words in the vowelized transcriptions of the audio data. The trigram LM is quite small, containing 57k 1-grams, 45k 2-grams and 40k 3-grams. In order to assess the influence of explicitly representing the vowelized forms, a contrast LM was estimated on the non-vowelized form of the audio transcriptions (that is removing short vowel and gemination markers).

The third LM results from the interpolation of the non-vowelized LM with a LM estimated on the normalized texts. This LM has 60k 1-grams, 6.1 M 2-grams and 10.7 M 3-grams and results in a test set perplexity of about 150. The interpolation coefficients for the text-based language model is 0.75. A fourgram language model was also estimated in the same manner.

4. EXPERIMENTAL RESULTS

The first recognition tests were carried out using a test corpus consisting of 7 hours of Radio Qatar data recorded in the fall of 2000. Word error rates are reported in Table 2 for the three language models, measuring all errors even on short vowels, excluding errors on the final vowel, and excluding all errors on short vowels or gemination markers. For these experiments, only one of of gender-independent, wideband acoustic models were used. For the 60k language model, with word error rate is about 40% relative to the vowelized reference. If the errors on short vowels and gemination markers that are not usually written are excluded, the word error rate is reduce to 20%.

A larger set of test data, from a wider variety of sources, was also used to assess the performance of the recognizer. Table 3 gives the word error rates using four different acoustic model sets: with gender-independent and gender-dependent, wideband and telephone models. A small gain is observed by using either gender- or bandwidth-dependent models, and the gain is additive when they are used conjointly. At the same time several improvements were made to the recognizer which explains the improved performance on the Radio Qatar data from 2000 which is the same as that used in the first set of experiments. A 4-gram LM was estimated using a slightly larger vocabulary of 65k words. In addition, the texts were further normalized to correct errors which had been missed and to ensure better consistency with word spellings, particularly for proper names which may have many different written forms.

Some observations can be made about the results. Data from sources which were seen in the audio training corpus and/or are closer in time to the training epoch tend to have lower word error rates, as can be expected. In general, the radio sources have lower word error rates, with the exception of Medi1, which is new and the only source from Morocco.

Language Model	Vocab. Size	OOV %	Px	Word Error Vowel	- final Vowel	no Vowel
LM_v vowelized, transcripts	57k	15.6	1952	39.2	28.1	24.5
LM_{nv} non-vowelized, transcripts	33k	7.5	751	45.8	31.7	26.3
LM_t non-vowelized texts	60k	4.1	550	42.9	28.3	22.2
LM_{nv+t} non-vowelized, trans+text	60k	4.1	151	40.1	26.6	19.9

Table 2: OOV and word error rates on 7 hours of data from Radio Qatar with four contrastive language models, LM_v trained on vowelized transcripts, LM_{nv} trained on non-vowelized transcripts, LM_t trained only on texts, and LM_{nv+t} LM_{nv} interpolated with LM_t. In the interpolated LM the coefficient for the texts is 0.75.

Audio Source	No Normalization %OOV	GI	GD	W/T	GD,W/T	Normalization %OOV	%WER
Radio Qatar 2000	6.2	16.4	16.0	16.0	15.7	3.2	12.7
Radio Qatar 2001	6.2	14.4	14.6	14.2	14.2	2.6	10.7
Radio Kuwait 2001	6.1	14.4	14.2	14.4	14.0	2.3	10.0
Radio BBCA2 2001	6.2	26.4	26.6	26.0	26.3	4.5	24.4
Radio Medi1 2002	7.6	28.8	28.7	28.8	28.5	5.9	27.1
TV Syria 2001	6.7	16.3	15.4	16.3	15.6	3.9	12.9
TV Aljazeera 2002	8.1	36.8	36.1	36.8	35.7	5.4	33.2
TV ESC 2002	13.9	33.4	33.4	33.4	33.5	6.5	27.0
TV7 2002	8.9	33.2	32.4	33.2	32.2	6.0	29.4
Average	6.9	21.5	21.1	21.2	20.9	4.0	17.8

Table 3: Word error rates (ignoring errors on short vowels and geminates) on test data from eight audio sources. The recognizer uses a 65k 4-gram language model. The left part of the table gives results with unormalized reference transcripts using gender-independent acoustic models (GI), gender-dependent AMs (GD), bandwidth-dependent AMs (W/T), and gender and bandwidth dependent AMs (GD,W/T)). The right part gives results after normalization of the reference transcripts using gender and bandwidth dependent AMs.

In classifying the system's errors we noticed that a substantial number were due to different orthographic forms for the same word. Some words occurred with or without the Hamza mark above or below the Alif. For example, the word "*economy*" is written both as 'IqtSAd' and 'AqtSAd'. This is quite common for words of foreign origin, such as "*democrat*" for which four written forms were found: 'dymwkrAT', 'dmwkrAT', 'dymkrAT' and 'dmkrAT'. Based on these observations, the test reference transcriptions were normalized. The recognition results after normalization are given in the right part of Table 3. It can be seen that the average OOV and word error rates are reduced by 2.9% and 3.1% respectively.

Of the remaining errors, 22% of the confusions are due to incorrect gluing of the word and an affix, where the word stem is correctly recognized. Such errors are more common for prefixes than suffixes. Of these errors 5.6% involve the article 'Al' (*the*), 6.6% the conjunction 'wa' (*and*, and 2.4% and 2.2% respectively for the propositions 'li' (*for*) and 'bi' (*with*). Other frequent errors concern gender and number agreement, where the different forms can have quite similar pronunciations, differing only in the final vowel (long or short).

5. DISCUSSION AND CONCLUSIONS

This paper has reported on some recent work we have done on transcribing Modern Standard Arabic broadcast news data. As described above, the Arabic language presents a number of challenges for speech recognition, due in part to the large lexical variety arising from inflections, and also to the difference between the spoken language and the written language, with the convention of writing texts without vowels. Recognition experiments carried out with 12 hours of test data from 8 sources resulted in a word error rate of about 18%, without counting errors on characters that are not written (mostly short vowels and geminates). In looking at the errors it is apparent that some of the mismatches between the recognizer hypotheses and manual reference transcripts are due to the use of different conventions and multiple spellings for the same word. If these differences are ignored there is an absolute word error reduction of about 2-3%. The explicit internal representation of vowelized word forms in the lexicon may be useful to provide an automatic (or semi-automatic) method to vowelize transcripts.

REFERENCES

[1] C. Barras et al. (2002), "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, **33**(1-2):5-22.

[2] J. Billa et al. (2002), "Audio Indexing of Arabic Broadcast News," *ICASSP'02*, **1**:5-8.

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

[4] J.L. Gauvain, L. Lamel (2000) "Fast Decoding for Indexation of Broadcast Data," *ICSLP'2000*, **3**:794-798.

[5] L. Lamel, J.L. Gauvain (2002), "Automatic Processing of Broadcast Audio in Multiple Languages," *Eusipco'02*.

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