The LIMSI RT-04 BN Arabic System

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

ABSTRACT

This paper describes the LIMSI Arabic Broadcast News system used in the RT-04F evaluation. The 10x system uses a 3 pass decoding strategy with MAP adapted gender- and bandwidth-specific acoustic models, vowelized 65k pronunciation lexicon, and a word class 4-gram language model where a word class regroups all vowelized forms for each non-vowelized entry.

The primary system was trained on about 150 hours of audio data and almost 600 million words of Arabic texts. A contrast system, trained only on resources distributed by the LDC, was also submitted. The word error rates of the primary system were 16.0% and 18.5% on the dev04 and eval04 data, and the respective word error rates were 17.6% and 20.2% for the contrast system.

1. INTRODUCTION

This paper describes some recent work improving our broadcast news transcription system for Modern Standard Arabic as described in [10]. 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, and 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 [7].

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

[*†] Visiting scientist from the Vecsys Company.

2. ARABIC LANGUAGE RESOURCES

The audio corpus contains about 150 hours of radio and television broadcast news data from a variety of sources including VOA, NTV from the TDT4 corpus, Cairo Radio from FBIS (recorded in 2000 and 2001 and distributed by the LDC), and Radio Elsharq (Syria), Radio Kuwait, Radio Orient (Paris), Radio Qatar, Radio Syria, BBC, Medi1, Aljazeera (Qatar), TV Syria, TV7, and ESC [10].

For the 70 hours of TDT4 and FBIS data, we used time-aligned segmented transcripts, shared with us by BBN, which had been derived from the associated closed-captions and commercial transcripts. These transcripts are not vowelized as is typically the case for Arabic texts, and have about 520k words (45k distinct forms).

The remaining audio data were collected during the period from September 1999 through October 2000, and from April 2001 through the end of 2002 [10]. These 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 realizations depending upon the speaker's knowledge of the language of origin and how well-known the particular word is to the target audience. These vowelized transcripts contain 580k words, with 50k distinct non-vowelized forms (85k different vowelized forms).

Combining the two sources of audio transcripts results in a total of 1.1M words, of which 70k (non-vowelized) are distinct.

The written resources consist of almost 600 million words of texts from the Arabic Gigaword corpus (LDC2003T12) and some additional Arabic texts 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[1] 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 [5, 7]. 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 600 million words, of which 2.2 M are distinct.

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 a fully expressed vowelized 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. This latter format is used for the 65k word lexicon, where a pronunciation graph is associated with each word so as to allow for alternate pronunciations. 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 words in the vowelized training data 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 65k vocabulary contains 65539 words and 528,955 phone transcriptions. The OOV rate with the 65k vocabulary ranges from about 3% to 6%, depending upon the test data and reference transcript normalization (see Table 2).

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.

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).

4. 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 LIMSI segmentation and clustering is based on an audio stream mixture model [4, 5]. First, the non-speech segments are detected and rejected using GMMs representing speech, speech over music, noisy speech, pure-music and other background conditions. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments. The result of the procedure is a se-

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

quence of non-overlapping segments with their associated segment cluster labels. Each segment cluster is assumed to represent one speaker in a particular acoustic environment and is modeled by a GMM. The objective function is the GMM log-likelihood penalized by the number of segments and the number of clusters, appropriately weighted. Four sets of GMMs are then used to identify telephone segments and the speaker gender. Segments longer than 30s are chopped into smaller pieces by locating the most probable pause within 15s to 30s from the previous cut.

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 LIMSI BN speech recognizer [5] uses 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.5kHz for telephone data) every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalization. Thus each cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

Word recognition is performed in three passes, where each decoding pass generates a word lattice which is expanded with a 4-gram LM. Then the posterior probabilities of the lattice edges are estimated using the forward-backward algorithm and the 4-gram lattice is converted to a confusion network with posterior probabilities by iteratively merging lattice vertices and splitting lattice edges until a linear graph is obtained. This last step gives comparable results to the edge clustering algorithm proposed in [9]. The words with the highest posterior in each confusion set are hypothesized.

Pass 1: Initial Hypothesis Generation - This step generates initial hypotheses which are then used for cluster-based acoustic model adaptation. This is done via one pass (less than 1xRT) cross-word trigram decoding with gender-specific sets of position-dependent triphones (5700 tied states) and a trigram language model (38M trigrams and 15M bigrams). Band-limited acoustic models are used for the telephone speech segments. The trigram lattices are rescored with a 4-gram language models.

Pass 2: Word Graph Generation - Unsupervised acoustic model adaptation is performed for each segment cluster using the MLLR technique [8] with only one regression class. The lattice is generated for each segment using a bigram LM and position-dependent triphones with 11500 tied states (32 Gaussians per state).

partitioning	546s (.12xRT)
1st decoding pass	4647s (1.0xRT)
2nd decoding pass	21369s (4.7xRT)
3rd decoding pass	7244s (1.6xRT)
Total	33806s (7.4xRT)

Table 1: Times for the different decoding steps.

Pass 3: Word Graph rescoring - The word graph generated in pass 2 is rescored after carrying out unsupervised MLLR acoustic model adaptation using two regression classes.

The primary system's Total Processing Time (TPT) was 33806s, as measured on a Dell Workstation 360 Pentium 4 Extreme 3.2GHz loaded with 4Gb of memory. The Source Signal Duration (SSD) for this test was 4556.42s. The system's Speed Factor (SF) was therefore 7.4xRT. For the contrast system the TPT was 33396s, and the SF 7.3xRT. The time spent for each decoding step is given in Table 1.

Acoustic models

The acoustic models are context-dependent, 3-state left-to-right hidden Markov models with Gaussian mixture. Two sets of gender-dependent, position-dependent triphones are estimated using MAP adaptation of SI seed models for wide-band and telephone band speech [6]. The triphone-based context-dependent phone models are word-independent but word position-dependent. The first decoding pass uses a small set of acoustic models with about 5700 contexts and tied states. A larger set of acoustic models, used in the second and third passes, cover about 15800 phone contexts represented with a total of 11500 states, and 32 Gaussians per state. State-tying is carried out via divisive decision tree clustering, constructing one tree for each state position 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 [5]. A set of 152 questions concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones.

The acoustic models for the contrast system were trained only on the audio data from LDC. This is about 72 hours of data from VOA, NTV, and Cairo Radio. The small set of acoustic models used in the first decoding pass have 5500 contexts and tied-states, and the larger set has 12000 contexts and 11500 tied states with 32 Gaussians per state.

The training data were also used to build the Gaussian mixture models with 2048 components, used for acoustic model adaptation in the first decoding pass.

Language models

The word class n-gram language models were obtained by interpolation [11] backoff n-gram language models trained on subsets of the Arabic Gigaword corpus (LDC2003T12) and some additional Arabic texts obtained from the Internet. Component LMs were trained on the following data sets:

1. Transcriptions of the audio data, 1.1M words
2. Agence France Presse (May94-Dec02), 94M words
3. Al Hayat News Agency (Jan94-Dec01), 139M words
4. Al Nahar News Agency (Jan95-Dec02), 140M words
5. Xinhua News Agency (Jun01-May03), 17M words
6. Addustour (1999-Apr01,) 22M words
7. Ahram (1998-Apr01), 39M words
8. Albayan (1998-Apr01), 61M words
9. Alhayat (1998), 18M words
10. Alwatan (1998-2000), 29M words
11. Raya (1998-Apr01), 35M words

The language model interpolation weights were tuned to minimize the perplexity on a set of development shows from November 2003 shared by BBN.

For the contrast system, the transcriptions of the non-LDC audio data were removed from the language model training corpus, reducing the amount of transcripts to about 520k words.

Table 2 gives the OOV rates and perplexities with and without normalization of the reference transcripts for the language models used in the Primary and Contrast systems. Normalization of the reference transcripts is seen to have a large effect on the OOV rate.

Unnormalized	*dev03*	*eval03*	*dev04*	*eval04*
% OOV	4.3	7.3	7.8	7.1
Px Primary	272.4	305.4	416.1	458.1
Px Contrast	271.7	306.2	422.8	462.9
Normalized	*dev03*	*eval03*	*dev04*	*eval04*
% OOV	3.3	4.0	4.8	6.4
Px Primary	267.8	307.3	423.8	459.3
Px Contrast	269.2	308.9	430.9	464.6

Table 2: OOV rates and perplexity of the 4 test sets (dev03, eval03, dev04 and eval04) with the Primary and Contrast system language models without (top) and with (bottom) normalization of the reference transcripts.

5. EXPERIMENTAL RESULTS

Table 3 gives the performance of the RT-04 Primary and Contrast systems on the RT-03 and RT-04 development and test data sets. The RT-03 development data was shared by BBN, and consists of the four 30-minute broadcasts from January 2001 (two from VOA and two from NTV). The RT-03 evaluation data was comprised of one broadcast from VOA and one from NTV, dating from February 2001. The RT-04 development data consist three broadcasts from the end of November 2003: 20031122_133544_ALJ_ARB, 20031128_113212_DUB_ARB, 20031130_180000_ALJ_ARB, and the RT-04 evaluation data come from the same sources, but from the month of December: 20031208_060215_ALJ_ARB, 20031211_113227_DUB_ARB, 20031217_083227_ALJ_ARB. The results given in the table use the eval04 glm files distributed by NIST.

Condition	*dev03*	*eval03*	*dev04*	*eval04*
Baseline	19.3	24.7	24.4	23.8
LDC AM	17.7	23.6	24.8	-
Base+LDC	17.4	23.0	21.9	23.3
+new word list	17.7	22.0	21.5	23.4
+mllt, cmllr	16.4	21.6	20.3	21.7
+gigaword LM	14.7	20.0	18.4	20.6
+pron	13.2	16.6	16.0	18.5
Contrast system	13.5	16.4	17.6	20.2

Table 3: Word error rates on the RT-03 and RT-04 development and evaluation data sets for different system configurations.

The baseline system had acoustic models trained on only the non-LDC audio data, and the language model training made use of about 200 M words of newspaper texts with most of the data coming from the years 1998-2000, and early 2001. With this system, the word error is about 20% for dev03, and 24% for the other data sets. The second entry (LDC AM) gives the word error rates with the acoustic models trained only on the LDC TDT4 and FBIS data. The word error is lower for the dev03 data, which can be attributed to the training and development data being from the same sources. The error rates are somewhat higher on the other test sets. Pooling the audio training data, as done for the primary system acoustic models, gives lower word error rates, and also exhibits less variation across the test sets. The remaining entries show the effects of other changes to the system. A new word list was selected using an automatic method, that did not necessarily include all words in the audio transcripts. Incorporating MLLT feature normalization and CMLLR resulted in a gain of over 1% absolute on most of the data sets. Finally, the language model and word list were updated using the Gigaword corpus which also included more recent training texts, and pronunciation probabilities were used during the consensus network decoding stage, resulting in a word error rate of 16.0% on the

dev04 data and 18.5% on eval04. This entry corresponds to our primary system submission. The results of the contrast system are shown in the last entry of the table.

6. CONCLUSIONS

This paper has reported on our recent development work on transcribing Modern Standard Arabic broadcast news data in preparation for the RT-04 evaluation. Previous work on broadcast news transcription at LIMSI in Arabic is reported in [10]. This same system had a word error rate of about 24% on the RT-04 dev and eval data. By improving the acoustic and language models, updating the recognizer word list and pronunciation lexicon, and the decoding strategy, a relative word error rate reduction of over 30% was obtained on the dev03, eval03 and dev04 data sets. On a set of 14 recent BN shows from July 2004 (about 6 hours of data from 12 sources), we obtain a word error of about 16.5% with the primary system.

Our acoustic models and lexicon explicitly model short vowels, even though these are removed prior to scoring. 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, E. Geoffrois, Z. Wu and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, **33**(1-2):5-22 January, 2001.

[2] 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, Orlando, 2002.

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

[4] J.L. Gauvain, L. Lamel and G. Adda, "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*, **5**:1335-1338, Sydney, December 1998.

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

[6] J.L. Gauvain and C.H. Lee, "Maximum A Posteriori for Multivariate Gaussain Mixture Observation of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, **2**(2):291-298, April 1994.

[7] L. Lamel and J.L. Gauvain, "Automatic Processing of Broadcast Audio in Multiple Languages," *Eusipco'02*, Toulouse, September 2002.

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

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

[10] A. Messaoudi, L. Lamel and J.L. Gauvain, "Transcription of Arabic Broadcast News," *ICSLP'04*, Jeju, October 2004.

[11] P.C. Woodland, T. Neieler and E. Whittaker, "Language Modeling in the HTK Hub5 LVCSR," presented at the 1998 Hub5E Workshop, September 1998.