THE IBM 2006 GALE ARABIC ASR SYSTEM

Hagen Soltau, George Saon,
Brian Kingsbury, Jeff Kuo, Lidia Mangu, Daniel Povey and Geoffrey Zweig

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
e-mail: {hsoltau,gsaon}@us.ibm.com

ABSTRACT

This paper describes the advances made in IBM's Arabic broadcast news transcription system which was fielded in the 2006 GALE ASR and machine translation evaluation. These advances were instrumental in lowering the word error rate by 42% relative over the course of one year and include: training on additional LDC data, large-scale discriminative training on 1800 hours of unsupervised data, automatic vowelization using a flat-start approach, use of a large vocabulary with 617K words and 2 million pronunciations and lastly, a system architecture based on cross-adaptation between unvowelized and vowelized acoustic models.

***Index Terms*—** Speech recognition

1. INTRODUCTION

Under the auspices of DARPA's Global Autonomous Language Exploitation (GALE) project, a tremendous amount of work was done by the speech research community toward improving speech recognition performance. The goal of the GALE program is to make foreign language (Arabic and Chinese) speech and text accessible to English monolingual people, particularly in military settings. A core component of GALE is ASR and research in this area spans multiple fields ranging from traditional speech recognition to speaker segmentation and clustering, sentence boundary detection, etc.

In this work, we focus primarily on Arabic broadcast news transcription although many of the techniques that are described here have been successfully applied to our Mandarin and English BN systems as well. Research on Arabic ASR has been fertile over the past few years as attested by the numerous papers published on this subject [1, 2, 3]. In the following, we describe the key design characteristics of our system:

- A cross-adaptation architecture between unvowelized and vowelized speaker-adaptive trained (SAT) acoustic models. The distinction between the two comes from the explicit modeling of short vowels which are pronounced in Arabic but almost never transcribed. Both models use a pentaphone acoustic context and comprise 5K context dependent states and 400K 40-dimensional Gaussians. The gains from cross-adaptation are estimated to be about 1% absolute.
- Large-scale discriminative training on 1800 hours of unsupervised data. The aforementioned models were trained with a combination of fMPE and MPE [4] on 135 hours of supervised data and 1800 hours of TDT4 BN-03 data. The gains from the unsupervised data are 1.3% absolute after discriminative training.
- Automatic vowelization using a flat-start approach. This results in a 2.0% absolute gain over unvowelized models on the broadcast news part while having a similar performance on broadcast conversations.
- Use of a vocabulary of 617K words which for the vowelized system translates into roughly 2 million pronunciations. The gains over a 129K word vocabulary are 1.5% absolute.

The following sections describe the training and test data (section 2), the system overview (section 3) and the vowelization experiments (section 4). Section 5 summarizes our findings.

2. TRAINING AND TEST DATA

For acoustic model training, we used the following corpora:

- 85 hours of FBIS + TDT4 with transcripts provided by BBN
- 51 hours of GALE data (first and second quarter releases) provided by the Linguistic Data Consortium (LDC)
- 1800 hours of unsupervised data (i.e. without transcripts) from the TDT4 BN-03 corpus

The following resources were used for language modeling:

- Transcripts of the audio data
- Arabic Gigaword corpus
- Web transcripts for broadcast conversations collected by CMU/ISL (28M words from Al-Jazeera)

The resulting language model was a 4-gram LM with 56M n-grams trained with modified Kneser-Ney smoothing. Throughout this paper, we report results on the following test sets:

- RT-04: 3 shows of 25 minutes each (totaling 75 minutes of BN) from 2 sources (ALJ, DUB)
- BNAT-05: 12 shows of 30 minutes each (totaling 5.5 hours of BN) from 5 different sources (VOA, NTV, ALJ, DUB, LBC) provided by BBN
- BCAD-05: 6 shows of 30 minutes each (totaling 3 hours of BC) from ALJ provided by BBN
- EVAL-06: 37 shows of 5 minutes each (totaling 3 hours of BN and BC) from 9 different sources

The table 1 shows the OOV rates and the word error rates on RT-04 for an ML-trained vowelized system (without pronunciation probabilities) as a function of the vocabulary size. As can be seen, the benefit from increasing the vocabulary from 129K to 589K words is 1.5% absolute. The final evaluation system used an even larger vocabulary of 617K words.

This work was partially supported by the Defense Advanced Research Projects Agency under contract No. HR0011-06-2-0001.

Nb. words	Nb. pronunciations	OOV	WER
129K	538K	2.9%	19.8%
343K	1226K	1.2%	18.6%
589K	1967K	0.8%	18.3%

Table 1: OOV and word error rates on RT-04 as a function of vocabulary size.

System	RT-04	BNAT-05	BCAD-05	EVAL-06
SI	24.4%	25.1%	30.3%	40.8%
U-SA	14.4%	15.3%	22.1%	29.7%
V-SA	12.6%	13.7%	21.2%	27.4%

Table 2: Word error rates of the different decoding steps on various test sets.

3. SYSTEM OVERVIEW

The operation of our system comprises the following steps depicted in Figure 1: (1) segmentation of the audio into speech and non-speech segments, (2) clustering of the speech segments into speaker clusters, (3) estimation of several speaker compensation transforms (VTLN, FMLLR and MLLR), (4) decoding with an unvowelized speaker adapted model (U-SA), (5) estimation of FMLLR and MLLR based on the U-SA output, (6) decoding with a vowelized speaker adapted model (V-SA). The performances of the various decoding steps on all the different test tests are summarized in table 2.

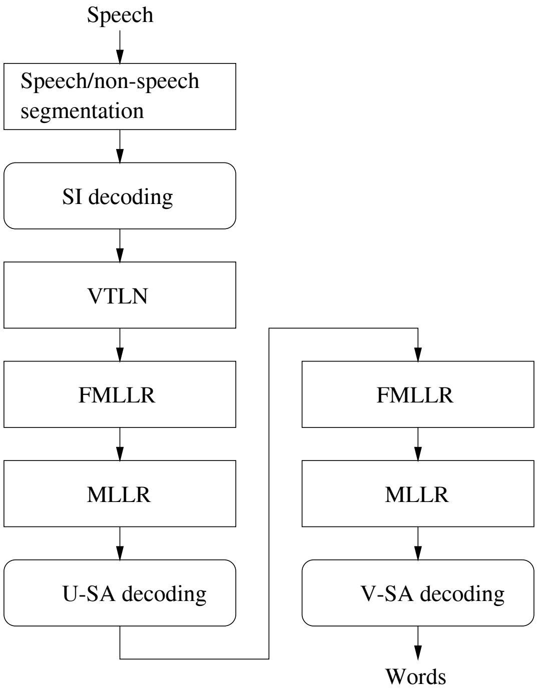

Figure 1: System diagram.

System	Nb. leaves	Nb. Gaussians
SI	3.5K	150K
U-SA	5.0K	400K
V-SA	4.0K	400K

Table 3: System statistics.

3.1. Speaker segmentation and clustering

We use an HMM-based segmentation procedure very similar to the one described in [5]. Speech and non-speech segments are each modeled by five-state, left-to-right HMMs with no skip states. The output distributions in each HMM are tied across all states in the HMM, and are modeled with a mixture of diagonal-covariance Gaussian densities obtained by clustering the Gaussians of the speaker independent system (to 240 Gaussians for speech and 16 for non-speech). Hypothesized speech segments are extended by an additional 30 frames to capture any low-energy segments at the boundaries of the speech segments and to provide sufficient acoustic context for the speech recognizer.

Next, the speech segments are clustered in the following way. Each segment is modeled by a single diagonal-covariance Gaussian in the speaker independent feature space. The Gaussians are clustered via K-means to a predefined number of clusters using the metric:

$$d(i,j) = (\mu_i - \mu_j)^T (\Sigma_i + \Sigma_j)^{-1} (\mu_i - \mu_j)$$

Additionally, we perform a bottom-up clustering of the resulting K Gaussians up to a prespecified maximum merge distance.

3.2. Acoustic modeling

Speech is coded into 25 ms frames, with a frame-shift of 10 ms. Each frame is represented by a feature vector of 13 perceptual linear prediction cepstral coefficients which are VTL-warped for the speaker adapted decoding. Every 9 consecutive cepstral frames are spliced together and projected down to 40 dimensions using LDA. The range of this transformation is further diagonalized by means of a maximum likelihood linear transform. Prior to splicing and projection, the cepstra are mean normalized for the speaker independent system and mean and variance normalized for the speaker adapted systems on a per speaker basis.

We use FMLLR [6] to map the VTL-warped data to a canonical feature space which in turn is transformed via fMPE [4]. The speaker adapted models are discriminatively trained in the resulting space with MPE. All models have pentaphone acoustic context and the number of context dependent units and 40-dimensional Gaussians are summarized in Table 3.

3.3. Unsupervised training

The starting point for unsupervised training are 1800 hours of audio from the TDT4 BN-03 dataset. First, we decode the entire data with an unvowelized speaker adapted system following steps (1)-(4) from section 3 (left-hand side of Figure 1). Then, we select a subset of the data based on per utterance average word posteriors. The word posteriors are obtained using consensus processing on the word lattices [7]. We then build the acoustic models by varying the amount of unsupervised data in addition to the 135 hours of supervised data (cf. section 2). In table 4, we report the posterior threshold, the amount of unsupervised data retained and the

Posterior thresh.	Nb. of hours	WER
1.0 (baseline)	0	17.2%
0.95	751	15.9%
0.9	1321	15.7%

Table 4: Amount of data and word error rates as a function of posterior threshold.

word error rates of a vowelized system trained on that data. As can be seen, a satisfactory performance can be obtained with only 751 hours of unsupervised training.

4. VOWELIZATION

While most ASR techniques are language independent, there are a few issues that are unique to Arabic. An excellent introduction to the Arabic language in the context of ASR can be found in [8]. One of the problems in Arabic speech recognition is the handling of short vowels and diacritics.

1. Fatha /a/
2. Kasra /i/
3. Damma /u/
4. Shadda (Consonant doubling)
5. Sukun (no vowel)

These five symbols are normally not written in Arabic texts. Only important religious texts such as the Koran are fully vowelized. Most of the training transcripts for the model building are not vowelized. This leads to a considerable mismatch between the acoustic data and the training transcripts.

When building vowelized models, it is important to keep in mind that the word error rate calculation is still based on the unvowelized forms. In other words, we map the vowelized words back to their unvowelized counterparts prior to scoring (NIST style scoring). Another reason for using the unvowelized forms for calculating the error rate is that the translation component used at IBM expects unvowelized Arabic text as input in the context of the DARPA GALE speech-to-text translation program.

One approach to bootstrap vowelized models is by using training transcripts manually vowelized by Arabic experts. This approach was chosen by LIMSI [3]. The obvious disadvantage is that writing vowelized transcripts is quite labor intensive. BBN reported in [2] on an automatic procedure based on Buckwalter's Morphological Analyzer [9]. Following the recipe in [2], we discuss our bootstrap procedure and some issues related to scaling up to large vocabularies.

4.1. Vowelized pronunciation dictionary

We use the Buckwalter Morphological Analyzer (Version 2.0), and the Arabic Treebank to generate vowelized pronunciations. The vowelized pronunciations are modeled as variants of unvowelized word forms, both in training and decoding. An example of vowelized and unvowelized pronunciations of the word *Abwh* is given below.

Abwh(deny/refuse/+they+it/him)	A a b a w o h u <
Abwh(desire/aspire/+they+it/him)	A a b b u w h u <
Abwh(father+its/it)	A a b u w h u <
Abwh(reluctant/unwilling+his/its)	A b u w h u <
Abwh(01)	A b w h

Training Method	WER (RT-04)
Flat-Start	23.0%
Bootstrap	22.8%

Table 5: Comparison of different Initialization methods for vowelized models.

Topology	RT-04	BCAD-05
3state	19.0%	28.9%
2state	18.5%	27.4%

Table 6: Comparison of different HMM topologies for short vowels.

The vowelized training dictionary has about 243368 vowelized pronunciations, covering a word list of 64496 words. The vowelization rate is about 95%. In other words, we couldn't find vowelized forms for 5% of our training word list. For these words, we back off to the unvowelized forms.

4.2. Flat-Start training vs. manual transcripts

Our flat-start training procedures initializes context independent HMMs by distributing the data equally across the HMM state sequence. We start with one Gaussian per Mixture, and increase the number of parameters using mixture splitting interleaved within 30 Forward/Backward iterations.

Now, the problem is that we have 3.8 vowelized pronunciations per word on average, but distributing the data requires a linear state graph for the initialization step. To overcome this problem, we simply select one single pronunciation variant randomly in this step. It should be noted, that all subsequent training iterations operate on the full state graph representing all possible pronunciations.

We compare this approach to manually vowelized transcripts where the correct pronunciation variant is given. BBN distributed 10 hours of manually vowelized development data (BNAD-05, BNAT-05) that we used to bootstrap vowelized models. These models are then used to compute alignments for the standard training set (FBIS + TDT4). A new system is then built using fixed alignment training, followed by a few Forward/Backward iterations to refine the models.

The error rates in table 5 suggest that manually vowelized transcripts are not necessary. The fully automatic procedure is only 0.2% worse. We opted for the fully automatic procedure in all our experiments, including the evaluation system.

4.3. Short models for short vowels

We noticed that the vowelized system performed poorly on Broadcast Conversational speech. It appeared that the speaking rate is much faster, and that the vowelized state graph is too large to be traversed with the available speech frames. One solution is to model the three short vowels with a shorter, 2-state HMM topology. The results are indicated in table 6. The improvements on RT-04 (Broadcast News) are relatively small, however there is an 1.5% absolute improvement on BCAD-05 (Broadcast Conversations).

Vocabulary	OOV Rate	WER (RT-04)
129K	2.9%	20.3%
589K	0.8%	19.0%

Table 7: OOV/WER ratio for an unvowelized system.

Vocab.	Variants	Vowel. Rate	OOV Rate	WER (RT-04)
129K	538K	-	2.9%	19.8%
343K	1226K	79.5%	1.2%	18.6%
589K	1967K	72.6%	0.8%	18.3%

Table 8: OOV/WER ratio for a vowelized system.

4.4. Vowelization coverage for the test vocabulary

As mentioned before, we back off to unvowelized forms for those words not covered by Buckwalter and Arabic Treebank. The coverage for the training dictionary is pretty high at 95%. On the other hand, for a test vocabulary of 589K words, the vowelization rate is only about 72%.

The question is whether it is necessary to manually vowelize the missing words, or whether we can get around that by backing off to the unvowelized pronunciations. One way to test this – without actually providing vowelized forms for the missing words – is to look at the OOV/WER ratio. The assumption is that the ratio is the same for a vowelized and an unvowelized system if the dictionary of the vowelized system doesn't pose any problems. More precisely, if we increase the vocabulary and we get the same error reduction for the vowelized system, then, most likely, there is no fundamental problem with the vowelized pronunciation dictionary.

For the unvowelized system, when increasing the vocabulary from 129K to 589K, we reduce the OOV rate from 2.9% to 0.8%, and we reduce the error rate by 1.3% (table 7).

For the vowelized system, we see a similar error reduction of 1.5% for the same vocabulary increase (table 8). The system has almost 2 million vowelized pronunciations for a vocabulary of 589K words. The vowelization rate is about 72.6%. In other words, 17.4% of our list of 589K words are unvowelized in our dictionary. Under the assumption that we can expect the same OOV/WER ratio for both the vowelized and unvowelized system, the results in table 7 and table 8 suggest that the back-off strategy to the unvowelized forms is valid for our vowelized system.

4.5. Pronunciation probabilities

Our decoding dictionary has about 3.3 pronunciations per word on average. Therefore, estimating pronunciation probabilities is essential to increase the discrimination among the vowelized forms. We estimated the pronunciations probabilities by counting the variants in the training data (incl. unsupervised BN-03 data).

The setup consists of ML models, and includes all the adaptation steps (VTLN, FMLLR, MLLR). The vocabulary has about 617K words, and about 2 million pronunciations. The test sets are RT-04 and BNAT-05 (both Broadcast News), and BCAD-05 (Broadcast Conversations).

Adding pronunciation probabilities gives consistent improvements between 0.9% and 1.1% on all test sets (table 9). Also, pronunciation probabilities are crucial for vowelized models; they almost double the error reduction from vowelization. Furthermore, we investigated several smoothing techniques and longer word context, but didn't see any further improvements compared to unigram pronunciation probabilities.

System	RT-04	BNAT-05	BCAD-05
Unvowelized	17.0%	18.7%	25.4%
Vowelized	16.0%	17.3%	26.0%
+ Pron. Prob	14.9%	16.4%	25.1%

Table 9: Effect of pronunciation probabilities on WER.

5. CONCLUSION

In this paper, we presented a set of techniques for Arabic broadcast news transcription. While some of them are also relevant for other languages (like speech/non-speech segmentation, front-end processing, speaker adaptation, unsupervised training, etc.), the emphasis has been put on those methods which are specific to Arabic ASR. Among these, vowelization is a key component in our system. We opted for a flat-start approach with pronunciations generated automatically using the Buckwalter morphological analyzer and studied some aspects related to coverage, HMM topologies and pronunciation probabilities. Another technique that was found to be beneficial is to combine a vowelized and an unvowelized system through cross-adaptation.

6. REFERENCES

[1] D. Vergyri, K. Kirchhoff, R. Gadde, A. Stolcke, and J. Zheng, "Development of a conversational telephone speech recognizer for Levantine Arabic," in *Interspeech-2005*, 2005.

[2] M. Afify, L. Nguyen, B. Xiang, S. Abdou, and J. Makhoul, "Recent progress in Arabic broadcast news transcription at BBN," in *Interspeech-05*, 2005.

[3] A. Messaoudi, Gauvain J.-L., and L. Lamel, "Arabic broadcast news transcription using a one million word vocalized vocabulary," in *ICASSP-2006*, 2006.

[4] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *ICASSP-2005*, 2005.

[5] G. Saon, G. Zweig, B. Kingsbury, L. Mangu, and U. Chaudhari, "An architecture for rapid decoding of large vocabulary conversational speech," in *Eurospeech-2003*, 2003.

[6] M.J.F. Gales, "Maximum likelihood linear transformation for HMM-based speech recognition," in *Tech. Report CUED/F-INFENG/TR291*. 1997, Cambridge University.

[7] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: lattice-based word error minimization," in *Eurospeech-1999*, 1999.

[8] Kirchhoff et al., "Novel approaches to Arabic speech recognition: Report from the 2002 Johns-Hopkins workshop," in *ICASSP-03*, 2003.

[9] T. Buckwalter, "Buckwalter Arabic morphological analyzer version 2.0," in *LDC2004L02*. 2004, Linguistic Data Consortium.