

Development of the SRI/Nightingale Arabic ASR system

D. Vergyri¹, A. Mandal¹, W. Wang¹, A. Stolcke¹, J. Zheng¹, M. Graciarena¹,
D. Rybach², C. Gollan², R. Schlüter², K. Kirchhoff³, A. Faria⁴, N. Morgan⁴

¹ SRI International

² RWTH Aachen University

³ University of Washington

⁴ International Computer Science Institute

dverg@speech.sri.com

Abstract

We describe the large vocabulary automatic speech recognition system developed for Modern Standard Arabic by the SRI/Nightingale team, and used for the 2007 GALE evaluation as part of the speech translation system. We show how system performance is affected by different development choices, ranging from text processing and lexicon to decoding system architecture design. Word error rate results are reported on broadcast news and conversational data from the GALE development and evaluation test sets.

Index Terms: speech recognition, large vocabulary, Arabic

1. Introduction

The goal of the Global Autonomous Language Exploitation (GALE) program is to develop computer software techniques to analyze, interpret, and distill information from speech and text in Arabic and Chinese. In order to achieve the high performance targets set by the program, it has been necessary to drastically improve the accuracy of the first stage of GALE systems, which involve automatic speech recognition (ASR) in the source language. In this paper we focus on the Arabic ASR system work. Since the data of interest was from broadcast news (BN) and broadcast conversations (BC) we concentrated on Modern Standard Arabic (MSA), the most common Arabic dialect used in public broadcasts.

The morphological complexity of Arabic and other language features, like the lack of short vowels in standard orthography, introduce new requirements for the design of the ASR systems. The characteristics of the Arabic language are described in detail in [1]. Previous work that dealt with the problems of MSA ASR can be found in [2], [3], [4].

This paper describes the following aspects of system design, and reports their impact on the system's performance: data preprocessing, lexicon and pronunciation probability generation, training data size, vocabulary size, sentence segmentation, front end features and discriminative training for acoustic models, and system architecture including system combination and cross adaptation. Our system includes components from two recognition engines: DecipherTM from SRI and the speech recognition system from RWTH Aachen. Results are reported for various GALE datasets that include both BN and BC speech.

2. Data

For acoustic training we used the 16 kHz Arabic data distributed by the Linguistic Data Consortium (LDC) with quick transcriptions: 45 hours of FBIS, 85 hours of TDT-4, 580 hours of GALE BC, and 740 hours of GALE BN data. We experimented with unsupervised training on an additional 3000 hours of untranscribed data, but these experiments showed ultimately no gains and are not included here. We also excluded all data sampled at 8 kHz that were distributed by LDC, since experiments showed no benefit from its use.

All acoustic data was aligned with the transcriptions using the flexible alignment procedure described in [5]. Long silences and regions with poor alignment confidence were discarded. Eventually, a total of 1120 hours was selected for training the acoustic models.

The language model (LM) data was also provided mostly by LDC, but some data came from other sources (under GALE). The total text used for LM included about 1.1 billion words: 490M of Arabic Gigaword (LDC, 2nd edition), 86M UN-data (LDC), 79M GALE data (LDC), 67 M CMU Web news data and 410M Web news data (collected by Cambridge University covering the dates between 11/2000 and 10/2006).

We evaluated the results on three development sets: the 2006 evaluation set (eval06: 3 hours, 11481 BN and 11243 BC words), the 2007 development set (dev07: 2.5 hours, 11742 BN and 6495 BC words) and the 2007 evaluation set (eval07: 4 hours, 14355 BN and 14187 BC words).

3. Text Processing and Lexicon generation

3.1. Text Data Processing

All text data was processed using the MADA¹ tool from Columbia University [6], which uses a statistical morphological tagger to fix common errors or ambiguities in the Arabic orthography, like the presence of Hamzas above or below the letters Alef, Yeh and Waw, and the marking of final Yeh and Teh Marbuta. MADA also provides full word diacritization, which was used as a means for developing the pronunciation lexicon. Other post-processing scripts were used to expand numbers and dates in the text data to words, remove nonlexical tags and punctuation marks, and normalize the orthography of some frequent names.

We did not compare our data processing method with that of other GALE sites (e.g., some sites do full normalization of Alefs and final Yeh), but we found that improved text processing had a significant effect on the accuracy of our models. Table 1 compares the word error rate (WER) results using MADA_1 (scripts released in 2005) versus MADA_2 (released in 2007 as v1.8). MADA_1 was based on the Buckwalter Arabic Morphological Analyzer² (BAMA) v.1 while MADA_2 used BAMA v.2, and the tagger was trained on more training data. The acoustic models used for this

Text Processing	Eval06	Dev07
MADA_1	36.3	25.7
MADA_2	35.7	25.2

Table 1. Effect of improved text processing on WER.

¹http://www1.cs.columbia.edu/~rambow/software-downloads/MADA_Distribution.html

² LDC distributions LDC2002L49 and LDC2004L02

	Eval06	Dev07
No pron. probs	36.8	26.0
With pron. probs	35.7	25.2

Table 2. *Effect of pronunciation probabilities on WER.*

comparison were maximum likelihood (ML), speaker independent (SI) MFCC models trained on 450h. The bigram LM used a 400K word vocabulary and was trained on ~500M words of the data released before 2007.

3.2. Lexicon

Since standard (undiacritized) Arabic orthography lacks the symbols that indicate the position of short vowels and the lengthening (doubling) of consonants [1], all Arabic ASR systems have the problem of representing these in the pronunciation lexicon ([2], [3], [4]). In our system we used the diacritized MADA output as the basis of our pronunciation lexicon. The mapping between the diacritized orthography and the phonemic representation is fairly straightforward. It is mostly a 1-1 mapping with a few rules that describe the pronunciation of Arabic (e.g., the Teh Marbuta is pronounced “a t”).

For acoustic model training the MADA-diacritized word form was used with a single pronunciation per word. We found that this improved the acoustic models by about 0.5%, compared to using undiacritized orthography and multiple word pronunciations. For the test lexicon we used the undiacritized orthography, and all diacritizations found for each word in the training data were used as pronunciation variants. This procedure resulted in an average of 3.3 pronunciations per word, for the words that MADA could diacritize. MADA failed to produce output for about 3% of training word tokens, which account for about 20% of the total training data word types (mostly infrequent words). For these words we used an automatic pronunciation generation model that was trained on the rest of the pronunciation lexicon to obtain their pronunciation, as is described in [7].

The pronunciation probabilities for the words appearing in the acoustic training data were obtained using acoustic forced alignment of the data with the phonetic pronunciations, and calculating smoothed empirical frequencies of the pronunciations of each word. For rare words or words not found in the acoustic training data, we obtained the probability for each pronunciation using a phone n-gram model trained on the forced alignment phone sequences. The n-gram probabilities were scaled and normalized. The effect of using pronunciation probabilities in the test lexicon on the first-pass WER is shown in Table 2. The same acoustic models were used as for the experiments in Table 1.

4. Language Modeling

4.1. Language Model Training

The LM was trained using the SRILM toolkit [8] with the undiacritized MADA-processed data. We trained source-specific LMs, each with modified Kneser-Ney smoothing, which were then interpolated for the final LM. The training sources were FBIS, TDT-4, GALE BN and BC transcripts (manual transcriptions and web data), Arabic Gigaword, CMU web data, CU web data, and UN data, for a total of about 1.1B words. Interpolation weights were tuned on a balanced held-out mixture of GALE BN and BC manual transcripts, of about 300K words. The final LM used a 600K

Vocab. Size	OOV(%)		WER(%) 2gram/4gram	
	Eval06	Dev07	Eval06	Dev07
64K	5.36	3.88	36.4/33.6	25.3/22.5
128K	3.09	2.05	35.8/32.7	24.6/21.7
256K	2.10	1.34	35.6/32.5	24.5/21.6
500K	1.20	0.69	35.4/32.4	24.3/21.5
600K	1.07	0.58	35.5/32.5	24.3/21.6

Table 3. *Effect of vocabulary size on OOV and WER*

vocabulary and was entropy-pruned to 40M bigrams for lattice generation decoding. For rescored purposes a much bigger 4-gram was used with about 100M bigrams, 73M trigrams, and 73M four-grams. These LMs were used in both the SRI and the RWTH systems.

We also experimented with morphological-class LMs and factored LMs, but as more data became available the improvement from these models diminished, and they were not used in the final system.

4.2. Vocabulary Selection

To deal with the high morphological variability of Arabic, we had to enlarge the vocabulary far beyond what is typically used for English ASR. We used the approach described in [9] to choose the vocabulary, using the same held-out set as for LM tuning. Table 3 shows the effect of vocabulary size on OOV (out-of-vocabulary) rate and WER, using all the available data for LM.

4.3. Effect of LM Training Data Size

Using about half the available LM data (about 500M words from Gigaword, UN and the first 2 years of GALE) we found that the performance of the LM (after 4-gram rescoring) degrades about 1-1.5% absolute on various test sets. The experiment was performed using a 500K vocabulary.

5. Acoustic Modeling

5.1. SRI Acoustic Models

As described in Section 6.3, the SRI system uses multiple models at different steps: non-cross-word (non-cw) models are used for first-pass lattice generation, and for lattice rescoring. Cross-word (cw) models are used only for lattice rescoring and N-best list generation.

5.1.1. Front Ends

We used an MFCC front end with 16 kHz sampling rate and a PLP one with 8 kHz sampling rate, to help decorrelate errors for system combination. (The 8-kHz front end is motivated by the fact that some of the BC data is narrowband and/or noisy.) The MFCCs were augmented with multilayer perceptron (MLP) features as described in [10]. The non-cw MFCC and the PLP front ends have 13 cepstral coefficients while the cw MFCC has 19. They all use 1st, 2nd and 3rd order derivatives, with HLDA for reduction to 39-dimensional feature vectors. Table 4 shows the performance of the various models on different test sets. The PLP model without MLP features does not perform as well as the MFCC models, but because of the 8-kHz front end it performs well on the BC part of the data.

	<i>Eval06</i>			<i>Dev07</i>		
	<i>BC</i>	<i>BN</i>	<i>all</i>	<i>BC</i>	<i>BN</i>	<i>all</i>
ML models						
noncw MFCC+MLP	33.6	21.8	27.6	19.7	11.3	14.3
cw PLP	32.1	21.4	26.7	18.8	12.0	14.4
cw MFCC19+MLP	31.8	21.1	26.4	18.2	10.9	13.5
MPFE models						
noncw MFCC+MLP	30.9	20.8	25.8	18.4	10.5	13.3
cw PLP	30.5	21.1	25.7	18.6	11.2	13.8
cw MFCC19+MLP	29.8	19.7	24.7	16.5	9.8	12.2
cw PLP+fMPE	30.1	20.4	25.2	17.8	10.8	13.3

Table 4. Results with different front-ends used at SRI acoustic models, contrasting ML and discriminative training conditions. All results are obtained with SI models after rescoring 4gram lattices created with the same non-cw MFCC model.

5.1.2. Discriminative Training

All models were trained with discriminative minimum phone frame error training (MPFE) [11]. We can see the gains over ML training in Table 4. We also experimented with using feature-level minimum phone error (fMPE) training [12],[13]. We found that the MLP features, which are already discriminatively trained, do not combine well with fMPE (there was practically no improvement from fMPE). That was the reason we decided not to include MLP features for the PLP front end and to use fMPE instead. We see that the PLP+fMPE models are comparable to the MFCC+MLP non-cw models after discriminative training. The cross-word MFCC19+MLP models performed the best.

5.1.3. Gender Dependent Models

Since we did not have data annotated for speaker gender, we used a gender prediction GMM trained on English data, to predict speaker gender in both training and test data. Then we performed MPE-MAP adaptation on the final discriminatively trained models, using the single-gender data only. This gave about 0.3% absolute improvement for individual models.

5.1.4. Duration Models

We used the approach described in [14] to train duration models for each acoustic model in our system. Single multivariate Gaussians are used to model phone frame durations in words and triphones, normalized for a speaker’s average speaking rate. We employ these duration models for rescoring all N-best lists used in system combination (see Section 6.3). On average, duration modeling contributed a relative gain of ~2.5% for each component system. The improvement in the final word error was about 0.2% absolute.

5.2. RWTH Acoustic Models

The RWTH system consists of two subsystems using two different acoustic models that share the same acoustic front end. MFCCs (16 cepstral coefficients) normalized by cepstral mean and variance normalization are augmented with a voicing feature. An LDA matrix projects the concatenation of 9 consecutive feature vectors in a sliding window to 45 components. As in the SRI MFCC models, this reduced feature vector is augmented with phoneme posterior features estimated by a neural network [15]. The two acoustic models

	non-cw system	cw system
Pass 1	18.6	-
Pass 2	14.4	14.4
Pass 3	13.9	14.2
Weight Optimization and Combination	12.7	

Table 5. Intermediate results of the RWTH subsystems on the dev07 corpus.

used differ in their treatment of cross-word context; one subsystem uses within-word models only. In training, speaker variations are compensated by applying VTLN to the MFCC filterbank and speaker adaptive training (SAT) based on CMLLR. The models were enhanced by performing discriminative training with the MPE criterion.

6. System Architecture

6.1. Acoustic Segmentation

Acoustic segmentation was applied in tasks where no manual transcription was available. Our segmentation method operates on the output of a speech recognizer. Several segment features are used to optimize the complete segmentation of a recording. A detailed description of this method can be found in [16]. In experiments where we used the non-cw system without LM rescoring and cross adaptation, we achieved a WER of 16.1% on Dev07 using the described segmentation. This result is quite close to the 15.9% achieved when using manual segmentations. However, there is still potential for improvement, which becomes apparent when splitting the manual segmentation before and after overlapping speech, yielding an error rate of 15.4%.

6.2. RWTH ASR Architecture

Speaker-independent within-word models are used in the initial recognition pass. The output of this system is used to estimate the text-dependent CMLLR transforms for both subsystems. Segment clusters obtained by generalized likelihood ratio clustering with a BIC-based stopping criterion act as speaker labels for the speaker adaptation methods applied. The second recognition pass is carried out using the CMLLR transformed features and acoustic models trained with SAT. The lattices produced by the two systems are rescored using the full 4-gram language model, while a pruned bigram LM is used for generating the lattices. The acoustic models of both systems are adapted to this rescored output using MLLR on the means. We used a cross-adaptation scheme to benefit from having two subsystems. The results of the final recognition pass are again rescored using the larger LM. Eventually, the lattices produced are combined by a method based on min-fWER decoding [15]. Intermediate results for both systems are given in Table 5.

6.3. SRI ASR Architecture

Figure 1 illustrates the design of the SRI ASR system. We generated lattices with the non-cw model twice (before and after adaptation) and cross-adapted all models used for final lattice rescoring to outputs of the other systems. The best system (MFCC19+MLP) was cross-adapted to the rover-combination of the other two systems. The results for different stages of the system are given in Table 6.

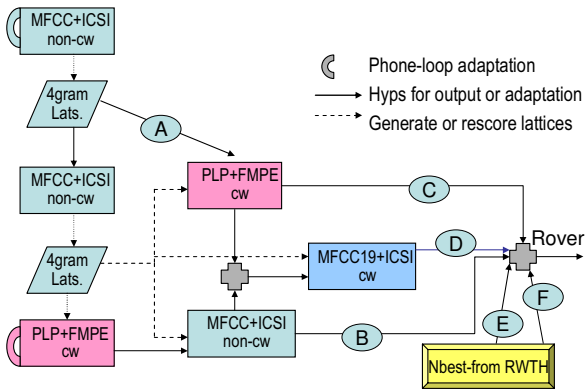


Figure 1. SRI System Architecture

Output	Eval06 (manual seg.)			Dev07 (manual seg.)		
	BN	BC	all	BN	BC	all
A	20.7	33.1	26.7	10.7	19.1	15.3
B	19.5	31.2	25.2	9.7	16.5	12.2
C	19.3	31.1	25.1	9.8	15.9	12.0
D	19.1	29.9	24.4	9.3	15.2	11.4
E-noncw	19.8	31.1	25.4	10.6	16.7	12.7
F-cw	20.4	31.8	26.0	10.6	16.7	12.8
Rover3	18.8	29.0	23.8	9.2	14.7	11.1
Rover5	18.2	27.9	23.0	8.9	14.3	10.8

Table 6. WER results at different system stages. Rover3 includes SRI-only systems B, C and D, while Rover5 includes also the two RWTH systems E and F.

6.4. System Combination

We used N-best ROVER on the three N-best lists obtained from the SRI system, and the two N-best lists from RWTH, to obtain the final output. The results are shown in the last line of Table 6 for the Eval06 and Dev07 test sets, while the final results for the latest Eval07 test are shown in Table 7.

Cross-adaptation schemes were also explored where the output from an intermediate pass from RWTH was used to adapt the SRI models, but since the SRI system was already using cross-adaptation among fairly different systems, cross-site adaptation did not give any further improvements.

7. Discussion and Conclusions

We have described the development process of a large vocabulary ASR system for MSA. The system was based on our English ASR experience, but we paid special attention to lexicon development and increased vocabulary size to deal with the needs of the new language. The different models had different enough features to achieve good cross-adaptation and system combination results. Our system obtained a WER of about 10% on BN and 15-25% on BC on different testsets. Since the goal of the GALE project was speech-translation performance, we compared the machine translation (MT) results obtained using the output of our system to those obtained using the reference hypotheses and found a difference of less than 1 point of BLEU or TER scores in all the test sets (when manual segmentation was used).

Ongoing work that has not yet been integrated in the system includes improved pronunciation generation and modeling, unvoiced acoustic models as an alternate system, factored LMs, partial diacritized vocabulary for LM and system optimization for MT purposes.

	Eval07			
	Manual seg.	Automatic seg.		
Output	BN	BN	BC	all
Rover3	10.3	10.1	16.0	13.0
Rover5	9.8	9.9	15.5	12.7

Table 7. The final system WER on Eval07. Manual segments were available only for the BN part.

8. Acknowledgments

We thank the CADIM group at Columbia University for providing and supporting the MADA tools. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 (approved for public release, distribution is unlimited). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA.

9. References

- [1] Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Gang J., Feng H., Henderson, J., Daben L., Noamany, M., Schone, P., Schwartz, R. and Vergyri, D., "Novel approaches to Arabic speech recognition: report from the 2002 JHU Summer Workshop", in *Proc. ICASSP*, vol.1, pp. 344–347, April 2003.
- [2] Afify, M., Nguyen, L., Xiang, B., Abdou, S. and Makhoul, J., "Recent progress in Arabic Broadcast News transcription at BBN", in *Proc. Interspeech*, pp. 1637–1640, 2005.
- [3] Messaoudi, A., Gauvain J.-L., and Lamel, L., "Arabic broadcast news transcription using a one million word vocalized vocabulary", in *Proc. ICASSP*, vol. 1, pp. 1093–1096, 2006.
- [4] Soltau, H., Saon, G., Povey, D., Mangu, L., Kingsbury, B., Kuo, J., Oma, M., and Zweig, G., "The IBM 2006 GALE Arabic ASR system", in *Proc. ICASSP*, vol. 4, pp. 349–352, 2007.
- [5] Venkataraman, A., Stolcke, A., Wang, W., Vergyri, D., Gadde, V.R.R., and Zheng, J., "An efficient repair procedure for quick transcriptions", in *Proc. ICSLP*, pp. 1961–1964, Oct. 2000.
- [6] Habash, N. and Owen, R., "Arabic tokenization, morphological analysis, and part-of-speech tagging in one fell swoop", in *Proc. 43rd Conf. of American Assoc. for Comp. Linguistics*, pp. 573–580, 2005.
- [7] Bisani, M. and Ney, H., "Multigram-based grapheme-to-phoneme conversion for LVCSR", in *Proc. EUROSPEECH*, vol. 2, pp. 933-936, 2003.
- [8] Stolcke, S., "SRILM - An extensible language modeling toolkit", in *Proc. ICSLP*, vol. 2, pp. 901–904, Sept. 2002.
- [9] Venkataraman, A. and Wang, W., "Techniques for effective vocabulary selection in domain specific speech recognition", in *Proc. EUROSPEECH*, pp. 245-248, 2003.
- [10] Zhu, Q., Stolcke, A., Chen, B. Y., and Morgan, N., "Using MLP features in SRI's conversational speech recognition system", in *Proc. Interspeech*, pp. 2141–2144, Sep. 2005.
- [11] Zheng, J. and Stolcke, A., "Improved discriminative training using phone lattices", in *Proc. Eurospeech*, pp. 2125–2128, Sep. 2005.
- [12] Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H., and Zweig, G., "fMPE: Discriminatively trained features for speech recognition", in *Proc. ICASSP*, vol. 1, pp. 961–964, 2005.
- [13] Zheng, J., Cetin, O., Hwang, M.-Y., Lei, X., Stolcke, A., and Morgan, N., "Combining discriminative feature, transform, and model training for large vocabulary speech recognition", in *Proc. ICASSP*, pp. 633–636, Apr. 2007.
- [14] Gadde, V.R.R., "Modeling word durations", in *Proc. ICSLP*, vol. 1, pp. 601–604, October 2000.
- [15] Hoffmeister, B., Plahl, C., Fritz, P., Heigold, G., Loof, J., Schluter, R. and Ney, H., "Development of the 2007 RWTH Mandarin GALE LVCSR system", in *Proc. ASRU*, pp. 455–460, 2007.
- [16] Rybach, D., Hahn, S., Gollan, C., Schluter, R. and Ney, H., "Advances in Arabic broadcast news transcription at RWTH", in *Proc. ASRU*, pp.449–454, 2007.