Colloquial Iraqi ASR for Speech Translation

Shirin Saleem, Rohit Prasad, Prem Natarajan

BBN Technologies
50 Moulton Street, Cambridge, MA 02138, USA
{ssaleem, rprasad, pnataraj}@bbn.com

ABSTRACT

In this paper we describe a real-time speech recognition system developed for colloquial Iraqi Arabic. This system is currently used in our speech-to-speech translation system configured for bi-directional communication in English and Iraqi on a laptop. We present experimental results on Iraqi utterances from different speech-to-speech translation domains, and analyze the usefulness of acoustic and language modeling data from different domains. We highlight the improvements obtained by modeling techniques that are language-independent, such as lattice-based discriminative training and domain-biased language model interpolation. In addition, we report on initial experiments we have performed to address specific challenges posed by Iraqi for speech recognition such as absence of short vowels and multiple forms of glottal stop, or the hamza, in the written form.

Index Terms: Speech recognition, Iraqi Arabic, Maximum Mutual Information, Vowelization, Hamza Normalization

1. INTRODUCTION

There are far-reaching benefits for developing a two-way speech-to-speech translation system that facilitates bi-directional communication over a language barrier. Under the DARPA TransTac and Babylon programs, many sites including BBN have developed speech-to-speech (S2S) translation systems that enable information exchange between an English speaker and a foreign language speaker. An early version of BBN's prototype was developed for the medical/refugee processing domain and the target language was Levantine Arabic [1]. Since then our prototype has been significantly improved and has been configured for an entirely new language pair of English and colloquial Iraqi Arabic [2].

Success of two-way S2S translation systems is predicated on development and robust integration of multiple technologies such as English and foreign language automatic speech recognition (ASR), English to foreign language translation and vice-versa, and Text-to-Speech (TTS) synthesis in English as well as in the foreign language. This paper addresses the development of accurate, real-time foreign language ASR, specifically for colloquial Iraqi Arabic.

Developing a speech recognition system for colloquial Iraqi Arabic (IA) poses significant challenges. The absence of short vowels in the orthography and the rich morphology of the Arabic language make it inherently hard for speech recognition. Also, acquiring a large corpus of transcribed speech data is much harder for colloquial dialects than for Modern Standard Arabic (MSA). MSA is the dialect used in news broadcasts and has been researched for a few years now [3-4]. Unfortunately, pronunciations and orthographies of many words are significantly different between MSA and colloquial IA. Therefore, MSA-based speech recognition is not well suited to colloquial IA speech.

This paper is structured as follows. In Section 2, we describe the characteristics of the Iraqi Arabic speech corpus that we have used to develop the recognition system. Section 3 is an overview of the BBN Byblos speech recognition system used for this research. In Section 4, we report on experiments that highlight the importance of domain-specific data. In Section 5, we report on gains obtained from improved language modeling. In Section 6, we present improvements in acoustic modeling. Section 7 discusses experiments performed to address the issue of short vowels and different forms of glottal stop in the Arabic orthography. In Section 8, we describe the configuration used in the TransTac 2006 March evaluation and the results obtained on the offline evaluation data.

2. COLLOQUIAL IRAQI CORPUS

Under the DARPA TransTac program, BBN has collaborated with DARPA, Appen PTY LTD., and Marine Acoustics Inc. (MAI) to collect and transcribe a large corpus of colloquial IA speech data (entirely from native speakers of IA) for three speech-to-speech translation domains: Medical/Refugee processing, Force-Protection (FP), and Civic Amenities. A set of scenarios was generated, organized around the themes relevant to the domain of interest. Each scenario consisted of a series of questions that were asked of the subject. For example, one scenario dealt with collecting biographical information using questions such as "What is your name", "Which city were you born in?, etc. Using a tool BBN has developed for such data collection, responses in Iraqi Arabic to the specified questions were collected for the three domains. We refer to this collection as the "1.5-way" collection, since the set of questions for which the responses were collected was fixed.

All audio from the 1.5-way collection was recorded in the MSWAV format at a sampling rate of 16 KHz and 16-bit per sample linear encoding. The Medical and FP data were collected in Australia and Jordan, while the Civic Amenities data was collected in Baghdad. A majority of speakers for all the collections were native speakers of the Baghdadi sub-dialect of IA, but there were a few speakers of native northern and southern sub-dialects.

A different kind of data collection was jointly undertaken by Defense Language Institute (DLI) and IBM Corporation. This collection was more free-form with the participants playing the role of an English speaking subject matter expert, a bilingual interpreter, and an Iraqi civilian. The collection covered the following scenarios: Traffic Control Point (TCP), Civic Amenities, Force Protection, and Common Community

Interest (CCI). The audio from this collection was recorded on 2 channels at a sampling rate of 22 KHz. We refer to this data collection as the "2-way" collection.

The audio data from all the domains was transcribed using MSA-based orthography in which the short vowels are not written. We held out two sets of 9 hours of speech data from each domain as development and test set respectively. In Table 1, we summarize the number of hours, number of utterances, number of words, and number of unique words for each domain. A total of 215 hours of transcribed speech data was available for acoustic modeling, and it included 1 million words that can be used for language modeling.

The average number of words per utterance for the 2-way datasets is about 2 times that of the 1.5-way set. Since the 2-way responses were not directed answers to pre-determined English questions, they are more spontaneous and diverse in content than the 1.5-way data.

Domain	Set	Hrs	#Utts.	#Words	Unique Words
1.5-way	Train	165	209K	730K	28.5K
	Dev.	6.3	7.5K	27.7K	5K
	Test	6.5	7.4K	27.9K	5K
2-way	Train	31	27K	240K	20K
	Dev.	2.2	1.9K	17.1K	4.3K
	Test	2.5	2.1K	18.3K	4.5K

Table 1: Description of the colloquial Iraqi corpus.

3. OVERVIEW OF THE ASR SYSTEM

We used the BBN Byblos [5-7] speech recognition system for the experiments described in this paper. The BBN Byblos recognition system uses phonetic hidden Markov models (HMM) with one or more forms of the following parameter tying: Phonetic-Tied Mixture (PTM), State-Tied Mixture (STM), and State-Clustered-Tied Mixture (SCTM) models. The states of each phonetic model are clustered based on the triphone or quinphone context into different "codebooks" (groups of Gaussian components). The mixture weights are clustered using linguistically-guided decision trees.

Decoding in the basic BBN Byblos system is performed in two passes [6]. The forward pass uses PTM or STM acoustic models and a composite set bigram language model (LM). The output of the forward decoding pass consists of the most likely word ends per frame, along with their partial forward likelihood scores. The backward decoding pass then operates on the output of the forward pass using SCTM within-word acoustic models and an approximate trigram LM to either generate an N-Best list or a word lattice. The word lattice or the N-best is typically rescored with a more detailed between-word SCTM models.

For the experiments described in this paper, we used STM models in the forward pass and the SCTM models in the backward pass. Given our focus on real-time performance on COTS laptops we did not use crossword rescoring of word lattices or N-best lists. We also used shortlists [7] to reduce the time taken to compute the Gaussian distances for each feature frame. Grammar spreading [7] is used to reduce the search errors during beam pruning.

4. TRAINING WITH DOMAIN DATA

For our baseline experiments, we used domain-specific STM and SCTM acoustic models estimated using Maximum Likelihood criterion via the EM algorithm. The features used for model estimation were the normalized energy and Mel-Frequency Cepstral Coefficients (MFCC) together with their first and second derivatives. RASTA normalization was performed to improve robustness to channel variations.

Since we did not have a dictionary of phonetic spellings for all the Iraqi words and also given that the orthographic representation of the words in the transcripts did not contain the short vowels, we used the "grapheme-to-phoneme" approach introduced in [3] to automatically generate the phonetic spellings. The grapheme to phonemes mapping approach used a phonetic set consisting of 39 phones.

We estimated domain-specific LMs using the acoustic training transcripts from each domain. Witten-Bell (WB) smoothing was used to mitigate the effects of data sparseness. The decoding lexicons for each domain consisted of all the words observed in the training data for that domain, which as shown in Table 1 are 28.5K and 20K words.

We decoded the test set from each of the domains using the domain-specific acoustic and LM pairs. We used the standard two-pass decoding described in Section 3. The word error rate (WER) obtained on the test set for each domain is summarized in Table 2. Given that there is significantly less training data from the 2-way collection than the 1.5-way collection it is not surprising that the WER is worse on the 2-way test set than the 1.5-way test set. Also, from Table 3 we can see that the out-of-vocabulary (OOV) rate and test set perplexity for the 2-way test set is much worse than the 1.5-way domain.

The WER in the off-diagonal elements of Table 2 shows that the performance is significantly worse for the "mismatched" training and test condition than the matched condition.

Training Data	%WER		
	1.5-way	2-way	Combined
1.5-way	**33.0**	69.5	47.5
2-way	64.8	**41.1**	55.4
Overall WER with matched train and test			**36.2**

Table 2: Decoding results for matched and mismatched test and train condition for domain-specific models.

5. LM IMPROVEMENTS

We explored the use of Kneser-Ney [8] (KN) smoothing and domain-biased interpolation of the domain-specific LMs. First, we re-estimated LMs for each domain using KN smoothing. Then, we decoded the test set for each domain using domain-specific acoustic models and KN-smoothed LMs trained for that domain. In Table 3, we compare the perplexity and WER obtained using KN-smoothed LMs with WB-smoothed LMs. As seen from the table, there is a modest improvement in perplexity for KN smoothing but no significant improvement in the WER. We have found that KN smoothing does result in an improvement in WER, if we build separate LMs for each sub-domain for the 1.5-way and 2-way data. Next, we generated LMs biased to each domain by interpolating the domain-specific KN-smoothed LMs, as well as additional 500K words from conversational Iraqi collection

from a different effort. The interpolation weights were estimated on the held-out development set using perplexity as a minimization criterion. While generating the interpolated LMs, we also expanded our lexicon to include all the 48.5K unique words observed in the training data and an additional 3.5K words from the out-of-domain conversational Iraqi data. We decoded the test set for each domain using the domain-specific acoustic models and the domain-biased interpolated LMs. As shown in Table 3, the domain-biased interpolated LMs (denoted as "DomainName-Inter" in Table 3) outperform the domain-specific LMs.

Test Set	LM	%OOV	Perp.	%WER
1.5-way	1.5-way-WB	1.8	83	33.0
	1.5-way-KN	1.8	73	33.0
	1.5-way-Inter	1.3	71	32.6
2-way	2-way-WB	6.0	302	41.1
	2-way-KN	6.0	265	41.0
	2-way-Inter	3.6	241	40.3

Table 3: Comparing different LM estimation techniques for the matched train and test condition. Domain-specific acoustic models were used for decoding the test set.

6. ACOUSTIC MODELING IMPROVEMENTS

In this section, we report on improvements made to the acoustic models by incorporating data from both 1.5-way and 2-way domains, and lattice-based discriminative training. For experiments reported in this section, we merged the 1.5-way and 2-way test sets to generate a "combined" test set. We also generated an interpolated "global" LM optimized on the combined development set from 1.5-way and 2-way development sets.

We estimated STM and SCTM "global" acoustic models in the ML framework with 196 hours of data from both 1.5-way and 2-way data collections. Next, we decoded the combined test set with different acoustic models: global model, 1.5-way model, and 2-way model. All decoding experiments used the interpolated LM optimized on the combined development set. As shown in Table 4, the global ML acoustic model significantly outperforms the domain-specific acoustic models. The global SCTM model consisted of 175K Gaussians, whereas the 1.5-way and 2-way data had 147K and 35K Gaussians respectively.

Acoustic Model	%WER
1.5-way ML	43.6
2-way ML	41.6
Global ML	38.1
Global MMI	34.2
Global MMI w/ lattice regeneration	33.7

Table 4: Comparison of different acoustic models on combined test set.

To further improve our acoustic models, we estimated acoustic models using the Maximum Mutual Information (MMI) criterion [9]. First, we generated lattices for the entire 196 hours of acoustic training data by decoding with the global ML models and a bigram LM trained with Witten-Bell smoothing on the same training data. Next, we used the global ML model as an initial estimate and performed 6 iterations of the Extended Baum-Welch (EBW) [9] algorithm to update the means and variances of Gaussians in the model. We trained using unigram LM probabilities in the lattices and also used I-smoothing [9] to avoid over-training.

We decoded the combined Iraqi test set with MMI models from different iterations using the interpolated LM optimized on the combined development set. We found that the MMI model from the 5th iteration resulted in the best WER. As shown in Table 4, the 5th iteration MMI model resulted in a WER of 34.2%, i.e. a 3.9% absolute reduction in the WER over the ML model. We further experimented with a variant of the conventional MMI training by regenerating lattices with the 5th iteration MMI model and performing additional iterations of MMI training using the regenerated lattices. Decoding with the MMI model after three iterations of training with regenerated lattices resulted in a WER of 33.7% on the combined Iraqi test set, which is 4.4% absolute better than decoding with the global ML model.

We also decoded test sets from each domain with the MMI model trained with lattice regeneration and domain-biased interpolated LM. In Table 5, we have summarized the overall improvements over the baseline domain-specific results presented in Section 4. The best results (Overall WER of 32.7%) were obtained by decoding with MMI models trained on the entire Iraqi training corpus and domain-biased interpolated LM trained for each domain. The overall decoding time with the above configuration was 0.6xRT (times real-time) on a 2.8 GHz Xeon CPU. It is possible that a better result can be obtained by training domain-specific MMI models given the diversity in the data collection and we plan to investigate that approach in the near future.

Acoustic Model	Language Model	%WER (Overall)
Domain-specific ML	Domain-specific	36.2
Global MMI	Domain-biased-interpolated	32.7

Table 5: Summary of improvement in WER over baseline decoding with models trained on domain data only.

7. IRAQI-SPECIFIC MODELING

In this section, we report on experiments we have performed to address the specific challenges offered by the Iraqi language. We report on results obtained by normalizing different forms of glottal stops. In addition, we report on experiments that compare the effect of using multiple "vowelized" pronunciations for Iraqi words as opposed to using a straightforward grapheme-to-phoneme mapping.

Hamza Normalization: Hamza (ء), which represents the glottal stop, can be written above or below the alif (ا) or on waaw (و), yaa (ي), or can appear by itself. The use of madda over alif (آ) and the interchanging of alif maksuura (ى) for yaa (ي) is also common in written Arabic. The writing conventions are not stringent, and hence different forms of the same word may appear with or without the hamza in the transcriptions. Therefore, while measuring WER we usually do not penalize the system, when the recognized word contains a different hamza than the one in the reference orthography.

We trained acoustic and language models with three common forms of the hamza – the hamza over alif, the hamza under alif, and madda over alif normalized to one common character alif in the training transcriptions. This change

reduced the dictionary size from 52K words to 49K words after the normalization. In Table 6, we summarize the results obtained by normalizing hamzas. Although, significant improvement was obtained by normalizing hamza for ML estimated models, no improvement was obtained following MMI.

ASR Models	Normalization in WER	%WER	
		ML	MMI
Un-normalized	No	39.7	35.5
Un-normalized	Yes	38.1	33.7
Hamza Norm.	Yes	37.4	33.9

Table 6: Results on overall test set with un-normalized and hamza normalized models.

Vowelization: The absence of short vowels in the orthography is a severe problem for speech recognition systems and has led to grapheme-to-phoneme as a popular approach to mitigating the problem. Recently, the Linguistic Data Consortium (LDC) has provided a pronunciation lexicon for approximately 13K Iraqi Arabic words from the 2-way TransTac dataset. There are on the average 1.5 pronunciations per word in the lexicon provided by LDC. Though this lexicon does not cover our complete dictionary, we attempted to use the LDC lexicon to build acoustic models for our system.

The first experiment we performed was to estimate ML acoustic models using the "vowelized" dictionary consisting of multiple pronunciations for each word. For fair comparison, we trained acoustic models with vowelized and un-vowelized dictionary only on the acoustic data that contained the 13K words from the LDC lexicon (75 hours of speech). In case of the training with vowelized dictionary, we introduced 5 more phonemes to model the short vowels. Pronunciation probabilities were also estimated for each pronunciation.

In Table 7, we show the effect of the two models by comparing the WER on a subset of the test set which did not contain any OOV words with respect to the 13K dictionary. From the table, we see that using the vowelized form of pronunciation dictionary did not result in an improvement over the grapheme-to-phoneme pronunciation dictionary. We believe this is due to lack of training data for different pronunciations.

Acoustic Model Configuration	%WER
ML STM and SCTM w/ unvowelized dict.	42.3
ML STM and SCTM w/ vowelized dict.	43.2

Table 7: Comparison of acoustic models trained using grapheme-to-phoneme dictionary with vowelized dictionary using 75 hrs of training data.

In our next experiments, we will explore combining the vowelized dictionary with the grapheme-to-phoneme dictionary to estimate acoustic models for the complete system. We will also explore discriminative training using Minimum Phone Error (MPE) criterion which is better suited to this problem than the MMI criterion.

8. EVALUATION SYSTEM

The Iraqi ASR system used in BBN's Speech-to-Speech translation system for the recently concluded March 2006 TransTac live evaluations was configured with the "global" MMI STM and SCTM models and "global" LM described in Section 6. The processing speed was faster than real-time on a 1.4 GHz Pentium M laptop. The same ASR configuration was used for decoding the offline evaluation data. The offline data consisted of a combination of speech recorded in clean and noisy conditions from various scenarios such as "Person of Interest", "Intelligence Screening", "Medical", etc.

In Table 8, we summarize the WER obtained on 2-way and 1.5-way test conditions with and without MLLR based speaker adaptation. Speaker adaptation results in a 20% relative improvement in the WER.

Decoding	%WER	
	1.5-way	2-way
Un-adapted	28.7	28.9
Adapted	23.0	22.6

Table 8: WER obtained on the TransTac March 2006 offline Iraqi evaluation data.

9. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a detailed investigation of the impact of various modeling techniques in the context of Iraqi Arabic speech recognition. We were able to derive large gains by using standard techniques such as use of more data, discriminative acoustic modeling, and speaker adaptation. Preliminary experiments with Iraqi Arabic specific modeling techniques such as hamza normalization and use of a vowelized lexicon did not result in significant improvements, but are worth pursuing further.

10. REFERENCES

[1] D. Stallard et al., "Design and Evaluation of a Limited Two-way Speech Translator," *Proc. EUROSPEECH, ISCA,* Geneva, Switzerland, pp. 2221-2223, Sept. 2003.

[2] D. Stallard et al., "A Hybrid Phrase-based/Statistical Speech Translation System" *Submitted to ICSLP 2006,* ISCA.

[3] J. Billa et al., "Audio Indexing for Arabic Broadcast News," *Proc. ICASSP,* IEEE, Orlando, FL, May 2002.

[4] M. Afify et al., "Recent progress in Arabic broadcast news transcription at BBN," *Proc. EUROSPEECH*, ISCA, Lisbon, Portugal, 2005.

[5] R. Prasad et al., "The 2004 BBN/LIMSI English Conversational Telephone Speech Recognition System," *Proc. EUROSPEECH,* ISCA, Lisbon, Portugal, Sept. 2005.

[6] L. Nguyen and R. Schwartz, "Efficient 2-pass N-best Decoder," *Proc. EUROSPEECH,* ISCA, Rhodes, Greece, Sept. 1997.

[7] J. Davenport et al., "Towards a Robust Real-Time Decoder," *Proc. ICASSP, IEEE*, Phoenix, AZ, March 1999.

[8] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," *Proc. ICASSP, IEEE*, pp. 181-184, 1995.

[9] P. C. Woodland and D. Povey, "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition," *Computer Speech and Language*, Vol. 16, pp. 25-47, 2002.