Rapid Development of an English/Farsi Speech-to-Speech Translation System

C.-L. Kao, S. Saleem, R. Prasad, F. Choi, P. Natarajan, D. Stallard, K. Krstovski, M. Kamali

BBN Technologies, Cambridge MA, 02138, USA
{ckao,ssaleem,rprasad,fchoi,prem,stallard,kkrstovs,mkamali}@bbn.com

Abstract

Significant advances have been achieved in Speech-to-Speech (S2S) translation systems in recent years. However, rapid configuration of S2S systems for low-resource language pairs and domains remains a challenging problem due to lack of human translated bilingual training data. In this paper, we report on an effort to port our existing English/Iraqi S2S system to the English/Farsi language pair in just 90 days, using only a small amount of training data. This effort included developing acoustic models for Farsi, domain-relevant language models for English and Farsi, and translation models for English-to-Farsi and Farsi-to-English. As part of this work, we developed two novel techniques for expanding the training data, including the reuse of data from different language pairs, and directed collection of new data. In an independent evaluation, the resulting system achieved the highest performance of all systems.

1. Introduction

As part of the DARPA TRANSTAC program, we have developed and fielded a free-form, two-way S2S translation system supporting English/Iraqi Arabic (E/I) in the military force protection domain [1]. In the latest phase of the program, a "Surprise Language" was announced to participating sites, with only a small amount of annotated data provided for training and development. The participants then had 90 days to build an end-to-end S2S translation system supporting the new language *and* new domain. Introducing a language this way was the first for the program. The program goal is to test the adaptability and scalability of the baseline system which should perform comparably regardless of the language. Farsi was chosen as the target language, and the application domain spanned a wide range of scenarios including medical interview, police interrogation, and airport security, etc.

In this paper, we describe the English/Farsi (E/F) system that was developed within the given time frame as part of this effort. The overall system architecture is briefed in Section 2. The challenges of configuring an S2S translation system in a new language pair of low resources, and the specific challenges presented by Farsi are explained in Section 3. Next in Section 4, we describe the limited data provided, and present two innovative techniques for increasing the amount of relevant data efficiently. In Section 5 and 6, we describe robust training procedures of the automatic speech recognition (ASR) and statistical machine translation (MT) component using a small amount of annotated data, and show the improvements from adding data collected effectively. Section 7 reports both the objective and subjective results from the evaluation with significant observations from error analysis using a methodology we have developed to quantify different translation errors. And finally in Section 8, we conclude with recommendations for future work.

2. Overview of the BBN S2S System

In BBN's S2S translation system architecture, for each translation direction, the system uses the BBN Byblos ASR engine to turn speech into text, then the BBN MT engine to turn the source-language text into target-language text, and finally the Cepstral Text-to-Speech (TTS) synthesizer to convert the target-language text to audio. In the English-to-foreign-language direction, the system also uses an English Canonicalizer module to check whether the English utterance is equivalent to one of the "canonical" utterances for which the system has a fluent recorded translation in the target language. The reader is referred to [1] for more details.

3. Challenges of the Farsi Language

The lack of sufficient Farsi linguistic resources and available annotated data presents a challenge for training robust ASR and MT models. The

amount of time available did not allow us to acquire the required linguistic expertise, or to prior conduct an extensive set of experiments. Little technical work on ASR and MT has been done in Farsi, so a baseline for comparison during development was not available.

Farsi shares the same issues as other languages employing the Arabic script. That is, the written text does not explicitly include vowelization information. Without the vowels in the transcription, there is usually a high degree of ambiguity in writing and translation introduced by "homographs", words with the same orthography but different pronunciations. In addition, Farsi also has a large number of "homophones", words with the same pronunciation but different orthographies, which can lead to ambiguity in speech recognition itself.

Compound Word Form I: نمیتونم (I can't)
سلام سلام آره من صداتون میشنوم نمیتونم ببینمت ولی صداتون میشنوم
Hi, hi ya. I hear you but I can't see you, but I hear you.
Compound Word Form II: نمی تونم (I can't)
به من کمک کن کمک کن نمی تونم بالا بیام که
Help me, help. I can't come up

Table 1: *Multiple forms of a Farsi compound word.*

Farsi is a highly inflected language with a complex word structure, with a different grammar from English. Unlike Modern Standard Arabic (MSA) and modern spoken Arabic dialects, which tend to prefer the same Subject Verb Object (SVO) word order as the English language, Farsi is a language with Subject Object Verb (SOV) word order; it also has a freer word order than either Arabic or English, due to the "scrambling" phenomenon. For example, almost any element, aside from adjectives, can be moved to sentence-initial position for emphasis. Verbs are marked for tense and aspect, and agree with the subject in person and number. Moreover, words can be used with their formal or informal endings, or modified to represent their colloquial pronunciations. One or more affixes can be attached to a word or to each other to form compound words, and components of compound words can be joined or separated depending on style. Table 1 shows an example of a compound word appearing in two different forms in transcriptions; Form I: components are joined, and Form II: components are separated with a half-space (the zero-width non-joiner which is used to preserve the final forms of some letters as in the middle of a word). Both forms have identical meaning: I can't, and are considered correct and present in our training transcripts. The complex word order and sentence structure, and irregularity in orthography make creating a lexicon and translation alignments challenging tasks.

4. Training Data and Expansion Techniques

The training data collected under the program includes 1.5-way (answers to a fixed set of questions) and 2-way (full dialog) recordings of transcribed speech and parallel translations. Participants also received the Babylon/CAST multilingual collection, which was translated from English to various languages, including Iraqi and Farsi, to create parallel corpora in text only. We divided the data randomly into 3 groups for training, development, and validation purposes. Table 2 and Table 3 show the amount and distribution of the speech and translation data after processing. Results reported in this paper are from the Dev set unless otherwise specified.

Set	#Hours Farsi	#Hours English
Train	82.2	3.6
Dev	3.5	0.6
Test	3.4	0.6

Table 2: *Farsi and English speech data.*

Farsi-to-English					
Set	**#Sentence Pairs**	**#Farsi Words**		**#English Words**	
		Total	**Unique**	**Total**	**Unique**
Train	75K	537K	24.6K	602K	11.3K
Dev	3.9K	28K	4.7K	32K	2.8K
Test	3.9K	27K	4.7K	30K	2.8K
English-to-Farsi					
Set	**#Sentence Pairs**	**#Farsi Words**		**#English Words**	
		Total	**Unique**	**Total**	**Unique**
Train	31.3K	178K	14.1K	200K	8.2K
Dev	1.6K	9.5K	2.6K	11.2K	1.9K
Test	1.6K	9.6K	2.7K	11.2K	1.9K

Table 3: *Description of Farsi and English parallel translation data.*

The amount of Farsi data is only approximately 1/5 of the available foreign language training data for the E/I system. In the following sections, we describe our efforts to improve system performance by augmenting the training data in a short period of time. The two novel approaches are: directed interactive data collection, and efficient reuse of parallel translation data from a different language pair. In addition, harvesting text data from the Web for language modeling was attempted, and is described later in detail in the ASR Section 5.2.

4.1. Directed Interactive Data Collection

We developed an in-house tool for viewing, editing, and collecting the translations produced by our S2S translation system. The combination of this tool and the S2S translation system was used for directed data collection to efficiently generate additional translated sentence pairs and translation variations for the in-domain application scenarios with very limited provided data.

A native English speaker and a Farsi speaker were instructed to use the S2S translation system to role-play new conversational scenarios that were not seen in the E/I data, such as medical interview, car accident, stolen identification, etc. The Farsi speaker, who also spoke English, reviewed and corrected the incorrect translations produced by the system.

In total, 1800 parallel sentence pairs were collected for different scenarios which were newly introduced in the Farsi data. These sentences were added to language modeling for both English and Farsi as additional in-domain training data, and the parallel sentence pairs were included in translation modeling. The advantage of adding this data was twofold. First, it proved useful in identifying exactly which phrases were translated incorrectly by the system and allowing the translation models to learn the correct translations for them. Secondly, it proved useful in adding new words and their translations to the system's vocabulary.

4.2. Efficient Reuse of Iraqi Data

In this section, we propose a novel and efficient procedure for generating parallel sentence pairs in a new language using an existing parallel corpus from a different language pair. We generated E/F sentence pairs by reusing the large E/I corpus of parallel data from the military force protection domain described in [7]. This method is efficient and beneficial because the data is (1) already cleansed and in the desired format for training, (2) likely to be in a similar domain, and (3) in the same conversational style as the new data.

First, we replaced all proper names in the E/I corpus by a single token, since names could be different across languages and cultures. We then selected complete English sentences (including English translations of Iraqi) that maximize trigram coverage on the E/I data. Next, we chose the most frequent 3K sentences that had minimal OOV rates with respect to the Farsi collection. The motivation for selecting complete sentences was that we had discovered that it was easier for human translators to translate complete sentences instead of isolated n-grams. These 3K sentences were translated to Farsi using the baseline E/F MT engine, and the output translations were reviewed and corrected by a native Farsi speaker. The resulting Farsi sentences were added to language model training, and translated sentence pairs were used to aid translation model training.

Such a selection strategy can be particularly effective when there is a significant domain overlap between the new system and the existing system. Another advantage of the above methodology is that it does not require separate data collection effort in the new language.

5. Speech Recognition

The BBN Byblos ASR system models speech as the output of context dependent phonetic Hidden Markov Models (HMMs). A detailed description of the ASR component in the E/I system is in [1]. We describe only the improvements made to the E/F system in this section.

5.1. Phonetic Transcription Scheme for Farsi

The Farsi alphabet consists of 32 letters, with 28 directly borrowed from the Arabic script but pronounced differently. Farsi has 3 short and 3 long vowels. The short vowels are almost never written, and the long vowels are represented by three letters, 'alef', 'yeh', and 'waw', which can also be consonants. Because all letters can be consonants, and some share the same phonemes, there are 23 distinct consonant sounds in Farsi.

We first created a training dictionary similar to our E/I system by using a one-to-one mapping of graphemes (in a modified Buckwalter transliteration system) to phonemes. Then we experimented with reducing the number of phonemes by mapping letters sharing the same pronunciation to the same phoneme. We tested

different phoneme sets of size from 36 down to 27. In each set, the same 4 phonemes were used to model non-speech events.

#Phonemes	36	32	31	28	27
%WER	42.7	42.2	42.3	42.5	42.5

Table 4: Farsi *WERs from varying the number of phonemes in training.*

The preliminary Farsi acoustic models were trained in the Maximum Likelihood (ML) framework, and the initial language models (LMs) were trained on 941K words of in-domain text using a lexicon of 36K words. Table 4 shows the word error rates (WERs) from training with phoneme sets of various sizes. The WER did not change as the number of phonemes was reduced. Due to this graphemic approach, vowels omitted from the written text were not present in the dictionary. With such a small amount of training data, this approach did not work as well as in the E/I system for Iraqi Arabic.

Vowelization of the training dictionary was attempted to reduce the Farsi WER. We extracted phonetic pronunciations from a 35K-word vowelized phonetic lexicon released in the program. This lexicon uses a phonetic scheme based on one developed by the University of Southern California (USC). USC's proposed class of transcription schemes, USCPers, USCPron, and USCPers+, are described in detail in [2].

Lexicon	Un-vowelized	Vowelized
%WER	42.2	38.7

Table 5: *Farsi WERs for models trained with un-vowelized and vowelized lexicons.*

To vowelize the training dictionary, words were mapped from their Buckwalter forms to the corresponding USCPers forms, and their graphemic spellings were replaced with the set of possible vocalic USCProns from the phonetic lexicon. Not all words in the original 36K graphemic training dictionary were present in the 35K phonetic lexicon; for those 967 words, their spellings remained graphemic in the final 36K training dictionary.

We trained ML models with the 36K vowelized dictionary described above. Total of 33 phonemes: 29 Farsi phonemes and 4 non-speech phonemes were used in training the phonetic models. The number of Gaussians is around 82K for the state-tied-mixture (STM) model used in the forward decoding pass, and 176K for the state-clustered-tied-mixture (SCTM) model in the backward decoding pass. As shown in Table 5, an 8% relative improvement was obtained from the vowelization of the phonetic model.

5.2. Improvements in Language Modeling

Training and improving LMs typically requires a large corpus of text that is matched to the target task domain both in terms of style (conversations, broadcast news, questions and answers, etc.) and topic (business, politics, etc.), and an extensive lexicon for reducing the out-of-vocabulary (OOV) rate for the recognition task. To improve n-gram coverage, we applied techniques described in [3] to harvest additional training data from the Web, and discarded documents with OOV rate higher than 0.2 measured on a 29K English or a 36K Farsi lexicon.

Farsi LM Data	Grammar Perplexity	Farsi %WER
E/F Farsi	200	32.6
E/F Farsi +Web	191	32.2

Table 6: *Effect of adding Web data on Farsi WER and grammar perplexity.*

Table 6 shows a slight positive impact on the grammar perplexity on the test set and a small 0.4% absolute reduction in Farsi WER from adding the Web data (44M words) to the in-domain E/F corpus (1.7M words) in LM training.

English LM Data *	Lexicon Size	English %WER
E/F	22K	31.1
E/F	29K	30.9
E/F + E/I	29K	26.0
E/F + E/I + BC	29K	25.8
E/F + E/I + BC + W	29K	25.6

Table 7: *Effect of adding data and lexicon size on English WER (* BC: Babylon/CAST corpus, W: Web data).*

For English, we experimented with five different LMs. First, without creating a new lexicon, we trained a language model using the provided in-domain E/F corpus (900K words), and the 22K lexicon directly taken from the E/I system.

The second model was trained on the same 900K-word E/F corpus, but with a 28% larger lexicon (29K) to cover new words from the target application domain that is out-of-domain for the E/I system. Then we created three more LMs by interpolating the model trained on the E/F corpus with one or more models trained on (a) the 5M-word E/I corpus, (b) the 700K-word Babylon/CAST corpus, and (c) 12M words of Web data. As shown in Table 7, we obtained a total of 17.7% relative gain from reusing existing data, increasing the lexicon size to reduce the OOV rate, and adding harvested Web data.

5.3. Improvements in Acoustic Modeling

For English acoustic training, the provided 3.6 hours of speech data was combined with the 130 hours from the E/I system, and the resulting model was used in both the E/I and E/F systems. The Farsi acoustic model was trained solely on the 82 hours of provided data (Table 2).

Training Criterion	Farsi %WER	English %WER
ML	36.4	34.7
MPE	33.1	29.3
HLDA-MPE	32.2	25.6

Table 8: *Improvements from MPE and HLDA.*

The acoustic models described in Section 5.1 were further improved with lattice-based Minimum Phone Error (MPE) discriminative training criterion [4] and the use of Heteroscedastic Linear Discriminant Analysis (HLDA) [5] to estimate feature transformations. We found that MPE models perform well even with ~80 hours of data. Estimated MPE models with HLDA transforms outperformed the earlier MPE models. The overall improvement with HLDA and MPE is 26.2% relative for English and 11.5% for Farsi (Table 8) using the LMs described in Section 5.2.

We also ported the incremental speaker adaptation [6] technique designed for "batch" operation to a low-latency, "streaming" mode for on-the-fly online adaptation in the HLDA feature space during decoding. On our offline 2-way test set, we obtained 13% relative WER improvement for English and 5% for Farsi using this adaptation technique.

6. Machine Translation

The BBN MT system uses a phrase-based engine described in [7]. The basic principle is the same as in [8] and [9]. Automated metrics, BLEU [10], METEOR [11], and Translation Error Rate (TER) [12], were used in assessing the translation quality.

In the following sections, we briefly describe the baseline configuration of the MT engine, and report the translation performance improvements obtained from adding the additional translation data collected using the novel techniques described in Section 44.

6.1. Baseline System

The baseline MT models were trained on all parallel training data provided (Table 3) and 64K sentence pairs from the Babylon/CAST corpus. Table 9 shows the baseline performance (denoted by 'B') of the Farsi-to-English (F2E) and English-to-Farsi (E2F) translation models measured on held-out Dev sets. The large difference in performance between F2E and E2F is due to the morphological complexity of the Farsi language, which contributes to the larger Farsi lexicon size and the higher perplexity of the Farsi set (191 vs. 46 on the Dev set).

Farsi-to-English			
F2E Data*	**BLEU**	**METEOR**	**100-TER**
B	31.2	61.0	38.9
B+D	32.1	61.6	38.7
B+D+I	32.5	62.3	39.5
English-to-Farsi			
E2F Data*	**BLEU**	**METEOR**	**100-TER**
B	18.0	47.2	40.4
B+D	18.7	47.8	41.5
B+D+I	18.9	47.9	41.5

Table 9: *F2E and E2F performances (* B: Baseline E/F data, D: Directed interactively collected data, I: E/I data)*

6.2. Improvements in MT Modeling

In this section we report the improvements from using novel approaches described in Section 4 to increase the quantity and enhance the translation quality of training data.

We presented a directed data collection method using our S2S translation system and an interactive

translation editing tool. While the system was being exercised and tested, domain-matching dialog-like data was collected at the same time. Adding this data to training can help fill in the gaps between the ASR and the MT components of the system. For instance, when a phrase or word is recognized correctly by the ASR engine, but translated incorrectly or imperfectly by the MT engine, adding this data in training helps the MT system to learn the correct and reliable translation; or if it is due to an unknown word to the MT system, this new word and its translation would be added to the system's vocabulary.

The gains from adding the 1800 directed interactively collected sentence pairs described in Section 4.1 (denoted by 'D') are summarized in Table 9. A significant gain in all metrics was noticed for E2F (3.9% relative gain in BLEU, 1.3% in METEOR, and 2.7% in TER), and a modest improvement in BLEU and METEOR is seen for F2E.

Next, the 3K sentence pairs selected and re-translated using the technique described in Section 4.2 was added to training. This new data provides a broader coverage for the translation phrase table. As shown in Table 9, in spite of the domain mismatch, applying this novel data reuse technique to include a small amount of re-translated data (denoted by 'I') in MT training resulted in a modest improvement in the E2F direction and achieved significant gains in all metrics in the F2E direction.

In total, 4.8K sentence pairs were added to training, and the significant improvements suggest that the quality of the alignments and hence the phrase table used for translation had improved from adding only a small amount of data collected using the techniques proposed in this paper.

7. Evaluation

The system described was evaluated as part of the DARPA TRANSTAC Evaluation conducted by NIST and held in July 2007. The evaluation was carried out in two modes:

(I) ***Lab Evaluation***: system usability was evaluated through live interactions between pairs of a US Marine and a Farsi speaker in a lab using structured scenarios. The objective was to provide an overall score to the capabilities of the whole system.

(II) ***Offline Evaluation***: ASR and MT component testing was conducted through the use of offline recorded speech audio files for speech-to-text (S2T) evaluation, and the proper transcription of the audio utterances for text-to-text (T2T) evaluation. The offline evaluation was performed so that component testing would be conducted on identical inputs for all systems. The purpose was to test individual system components to see how well they perform in isolation.

7.1. Evaluation Metrics

For the Lab Evaluation, three quantities were measured by bilingual human judges:

1. *Complete Exchange*: the number of high-level concepts that the US Marine was able to successfully retrieve in a 10-minute period.
2. *Proper Question*: the number of English utterances correctly translated.
3. *Proper Answer*: the number of Farsi utterances correctly translated.

In addition, the questionnaire completed by US Marines and Farsi speakers after each scenario they participated in was analyzed.

For the Offline Evaluation, automated MT metrics (BLEU, METEOR, TER) for both S2T and T2T, and ASR WER were computed. Subjective Likert [13] judgment for semantic adequacy by bilingual human judges was also performed on a subset of the T2T output. On a 1 to 5 Likert scale, a score of 5 denoted perfect translation, 4 adequate translation, 3 semi-adequate, and so on.

7.2. Evaluation Results

Our system demonstrated the best overall system performance in the Lab Evaluation as well as robust component capability and accuracy in the Offline Evaluation.

In the Lab Evaluation, the system produced the highest counts on all three areas measured, namely, having the most number of correctly transferred concepts with the largest number of properly translated questions and answers in both languages. Moreover, observations from the post-scenario survey results ranked this system as the most preferable one for its ability to help English and Farsi speakers communicate effectively.

Condition		BLEU	MET.	100-TER	100-WER
S2T	E2F	19.3	45.5	41.0	84.7
	F2E	29.7	56.7	37.7	72.3
T2T	E2F	23.3	50.3	44.3	N/A
	F2E	35.7	63.2	45.7	N/A

Table 10: *BBN July 2007 offline evaluation results on automated MT and ASR metrics.*

In the Offline Evaluation, the results of the automated metrics are shown in Table 10. The BBN system received the highest E2F BLEU score for the T2T condition; the highest F2E scores for all three MT metrics (BLEU, METEOR, TER) under both the S2T and T2T conditions; and the best TER scores in both E2F and F2E directions for both the S2T and T2T conditions. The English ASR had the highest accuracy, and the Farsi ASR accuracy was the second best. In addition, Likert-scale analysis at utterance level was performed by bilingual human judges. Our system had the largest "Completely Adequate" Likert percentages, and the smallest "Inadequate" Likert percentages for both E2F and F2E directions. This favorable subjective result was supported by the automated metrics described. In summary, our system was the top performer in both automated and subjective tests.

7.3. Subjective Analysis of MT Output

The MT component of our system was evaluated subjectively on a subset of the T2T test set consisting of 216 English and 206 Farsi utterances. In this section, we describe our effort to determine the relative importance of different translation errors using a method we have recently developed.

In [14], we define the "Likert Error" (LER) for a translation as 5 minus its Likert score, and "Total Likert Error" (TLE) of a set of translations as the sum of the LERs of the translations:

$$TLE(C) = Count(C)*LER(C),$$

where *LER(C)* is the average damage done by instances of *C*, and quantifies the "seriousness" of the error.

Table 11 gives the LER weights and the estimated TLEs for each language direction. We can see the percentage of errors caused by each error category, and that the four major error categories for both F2E and E2F are: (a) using the wrong word sense of a word because the two languages do not share the same conflation on the same word; (b) wrong word order; (c) missing a concept word whose omission really matters to the meaning of the utterance; and (d) wrong concept, meaning that a word or phrase translation that is wrong in all contexts, regardless of word sense. Interestingly, the "Wrong Word Order" error was not ranked among the high-frequency error categories in our E/I system [14] because Iraqi Arabic is often spoken with the same word order as English.

Farsi-to-English			
Error Category	**%Count**	**LER**	**%TLE**
Missing Concept Word	39.1	1.13	41.5
Wrong Word Sense	21.2	1.08	21.5
Wrong Word Order	20.5	0.89	17.2
Wrong Concept	10.4	1.29	12.6
Pronoun Error	2.9	0.86	2.4
MT OOV Word	2.0	1.0	1.8
Other	2.1	1.33	2.9
English-to-Farsi			
Error Category	**%Count**	**LER**	**%TLE**
Wrong Word Sense	29.8	1.22	30.6
Wrong Word Order	23.4	1.13	22.3
Wrong Concept	11.1	1.58	14.7
Missing Concept Word	15.7	1.09	14.4
Extra Concept Verbiage	6.0	0.9	4.5
MT OOV Word	2.6	2.06	4.4
Pronoun Error	4.7	1.02	4.0
Other	6.8	0.87	5.0

Table 11: *Estimated Likert Error values.*

The system missed lexical syntax when it translated words separately which were really part of a syntactic structure. Among other errors, we also observed that often the system did not correctly translate plural nouns due to Farsi's complex compound word structure. When a plural suffix comes after the noun with a full space, the system might consider it an independent entry and translate it to the noun's singular form in English. Lastly, the system performed poorly on names possibly due to variances of transliteration of the same names and the lack of names in the lexicon. This analysis helps quantify the distributions of the errors and direct our efforts towards improving the system.

8. Conclusions and Future Work

In this paper, we presented novel techniques for developing an English/Farsi S2S translation system with limited resources within 90 days. The techniques spanned the following: making best use of limited data for ASR and MT training; reusing existing data; and augmenting training data with human-in-the-loop directed interactive data collection. The described system performed considerably better than others in the evaluation.

On analyzing the system output, we located top translation error categories. Future research will focus on algorithms for addressing these errors. We plan to explore techniques on better syntax translation and name recognition, such as name-aware translation. For dealing with Farsi's complex word order, we plan to investigate techniques for re-ordering English or Farsi source utterances to bridge the gap in linguistic constructs between the two languages before translation. We found that vowelization was critical for improving the ASR performance in our system. In future work, we will explore techniques for automatic vowelization for Farsi. We also plan on tighter integration of our ASR and MT engines.

Because there are neither standardized guidelines for Farsi text normalization nor well-defined orthographic transcription rules, transcripts done by professionals still contain a large number of inconsistencies in compound word forms. We believe that having standardized Farsi normalization guidelines is essential to reducing data inconsistency and improving system performance.

9. References

[1] David Stallard, Fred. Choi, Chia-lin. Kao, Kriste Krstovski, Prem. Natarajan, Rohit Prasad, Shirin Saleem, and Krishna Subramanian, "The BBN 2007 Displayless English/Iraqi Speech-to-Speech Translation System," *Proc. Interspeech 2007, pp. 2817-2820,* Antwerp, Belgium, August 2007.

[2] Shadi Ganjavi, Panaviotis G. Georgiou, and Shrikanth Narayanan, "ASCII Based Transcription Systems for Languages with the Arabic Script: The Case of Persian," *Proc. ASRU, pp. 595-600*, St. Thomas 2003.

[3] Ivan Bulyko, Andreas Stolcke, Mari Ostendorf, "Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling Using Class-dependent Mixtures," *Proc. HLT/NAACL, pp-7-9, 2003*

[4] Dan Povey and Phil C. Woodland, Minimum Phoneme Error and I-smoothing for Improved Discriminative Training," *Proc. ICASSP '02.*

[5] Spyros Matsoukas and Richard Schwartz, "Improved Speaker Adaptation Using Speaker Dependent Feature Projections," *Proc. IEEE ASRU, pp. 273-278*, St. Thomas, Nov 30 - Dec 3, 2003.

[6] Daben Liu, Daniel Kiecza, Amit Srivastava, and Francis Kubala, "Online Speaker Adaptation and Tracking for Real-Time Speech Recognition," *Interspeech 2005*.

[7] Shirin Saleem, Krishna Subramanian, Rohit Prasad, David Stallard, Chia-lin Kao, Prem Natarajan, and Raid Suleiman, "Improvements in Machine Translation for English/Iraqi Speech Translation," *Proc. Interspeech 2007, pp. 2445-2448*, Antwerp, Belgium, August 2007.

[8] P. Koehn, F. Oec, and D. Marc, "Statistical Phase-Based Translation," *Proc. HLT/NAACL 2003*.

[9] P. Brown, S. Della Pieta, V. Della Pietra, and Robert Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics, 19(2), 263 - 311, 1991*.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: A method for automatic evaluation of machine translation," *Proc. ACL, pp. 331-318*, 2002.

[11] S. Banerjee and A. Lavis, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," *Proc. ACL 2005*.

[12] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," *Proc. Association for Machine Translation in the Americas, 2006*.

[13] V. Barnett, "Sample Survey Principles and Methods," *Hodder Publisher, 1991*.

[14] David Stallard, Chia-lin Kao, Kriste Krstovski, Daben Liu, Prem Natarajan, Rohit Prasad, Shirin Saleem, Krishna Subramanian, "Recent Improvements and Performance Analysis of ASR and MT in a Speech-to-Speech Translation System," *Proc. ICASSP 2008. pp. 4973-4976*, Mar. 30 – Apr. 4 2008, Las Vegas, Nevada USA.