The BBN 2007 Displayless English/Iraqi Speech-to-Speech Translation System

David Stallard, Fred Choi, Chia-lin Kao, Kriste Krstovski, Prem Natarajan, Rohit Prasad, Shirin Saleem, Krishna Subramanian

BBN Technologies
10 Moulton Street Cambridge, MA
{stallard,fchoi,ckao,krstovs,prem,rprasad,ssaleem,ksubrama}@bbn.com

10.21437/Interspeech.2007-648

Abstract

Spoken communication across a language barrier is of increasing importance in both civilian and military applications. In this paper, we present an English/Iraqi Arabic speech-to-speech translation system for the military force protection domain (checkpoints, municipal services surveys, basic descriptions of people, houses, vehicles, etc). The system combines statistical N-gram speech recognition, statistical machine translation, hand-crafted rules, and speech synthesis in order to translate between the two languages. The system is designed for "eyes-free", or "displayless" use. That is, it does not make use of a screen, mouse, or keyboard, but is instead operated by a handheld microphone with just two push buttons: one for English, and the other for Arabic.

Index Terms: speech-to-speech translation, Iraqi Arabic

1. Introduction

Military and humanitarian personnel often need to communicate with residents of a host country who do not speak English. Human interpreters are inevitably in short supply, and training personnel to speak a new language is difficult. Portable devices for speech-to-speech language translation would therefore be very useful in such environments.

This paper describes new work on a speech-to-speech (S2S) translation system earlier reported on in [1]. The system is designed to allow an English-speaking soldier, termed the Subject Matter Expert (SME), to engage in translingual dialog with a native Iraqi speaker, termed the Foreign Language Expert (FLE). The domain of the system is military force protection, including checkpoints, house searches, civil affairs, etc. The system combines Automatic Speech Recognition (ASR), Machine Translation (MT) and Text to Speech (TTS) technologies.

The research reported in this paper was performed under DARPA's TRANSTAC program, which conducts evaluations of systems and provides common training data to develop them. Other systems in the program include systems developed by IBM [2], SRI [3], and CMU [4]. In Phase I of TRANSTAC, systems were allowed to use a Graphical User Interface (GUI). In Phase II of the program, however, systems were required to move to an "eyes-free", or "displayless" interface, in which the user was not allowed to use a keyboard or view a screen, and was only allowed to control the system through a handheld control device of their own choosing or devising.

In what follows, we first discuss the overall architecture of the system, which was substantially redesigned to meet the challenges of displayless use. We next present the speech recognition and translation components. Finally, we give results of the common TRANSTAC evaluation, which was carried out in January of 2007.

2. User Interface and System Architecture

The system architecture is shown in Figure 1. For each translation direction, the system uses ASR to turn speech into text, MT to turn source language text into target language text, and finally TTS to turn this text into speech of the target language. In the English-to-Iraqi direction, the system also uses an English Canonicalizer module to check whether the English utterance is equivalent to one of the ~700 "canonical" utterances for which the system has a fluent recorded Iraqi translation. A central dialog manager orchestrates the operation of the system, and receives control inputs from the user interface. The English and Iraqi ASR components are BBN's Byblos system [6]. The English-to-Iraqi (E2I) and Iraqi-to-English (I2E) MT are BBN's phrase-based statistical MT engine. The complete system runs on a Windows XP laptop with 2 GB RAM.

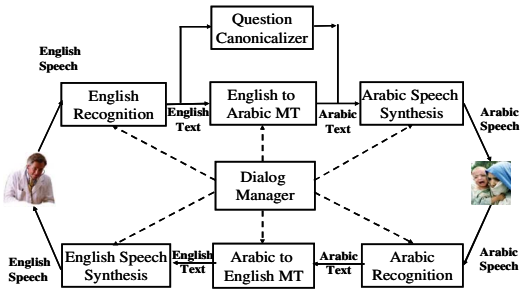

Figure 1: Block Diagram

In designing the displayless version of the system, we sought to make it as effective, efficient, and easy to use as possible. Transferring crucial functions of the GUI-based system to the displayless environment posed a number of challenges. One challenge was to squeeze the manual controls of the GUI onto a device that could be operated with just one hand. The GUI included multiple on-screen buttons, including "Listen English", "Listen Arabic", "Abort", "Reprompt", etc. For the sake of simplicity and ease of use, however, we wanted the hand control have just two buttons

A second challenge was to provide confirmation for the SME. The GUI-based system displayed the text of the English ASR output as a confirmation, so that the SME could tell when a mistake had been made, and take corrective action. Without this capability, the SME would have no way to tell when the system had misunderstood him, and confusion between him and the FLE could quickly mount. The only way for a displayless system to provide the confirmation would be to speak it. But adding an extra system utterance to each SME

turn would increase the time taken by the overall interaction, conflicting with the goal of efficiency.

A third challenge was managing the interaction between the system and the SME. There are two aspects to this. First, the system must not allow the SME's actions to place it in an inconsistent state, such as trying to listen and speak at the same time. Second, the SME must know what he can do at all times, so that he does not become confused or frustrated. The GUI system managed the interaction with a rigid state-transition network. This network controlled what the SME could do at any given time by disabling buttons, for example disabling "Listen" when the system was doing TTS output. To let the SME know what actions were currently possible, the system displayed its current state ("Speaking", "Listening", etc) on the screen, and grayed out the buttons it had disabled.

The displayless system, by contrast, cannot prevent the user from pressing the mechanical buttons of the hand control, nor can it convey to him graphically what it is doing, and why he has to wait before he can press the button that he wants to press now. We chose instead an entirely different model of interaction, in which the system is completely interruptible in all aspects of its operation, and the SME is free to take any action at any time. As we shall see, this redesigned interaction also allows us to reduce the number of buttons required to operate the system, and helps make audio confirmation more efficient.

The system's hand control device is a handheld microphone on which two buttons were mounted. This approach integrates manual control and voice input in a single device, and also provides a built-in visual cue for turn-taking. One button, labeled "YOU", is for the English speech of the SME, while the other, labeled "HIM", is for the Iraqi speech of the FLE. Pushing the YOU button causes the system to begin performing English ASR on speech from the microphone. Releasing the button finishes ASR, and passes ASR output to English-to-Iraqi MT. The Iraqi MT output is then passed to the Iraqi TTS and spoken out to the FLE. The "HIM" button operates analogously for Iraqi speech.

In the displayless system's interaction model, the SME can press the YOU button at any time and speak, no matter what the system is doing, even if it is still doing ASR on the SME's previous utterance. Any currently active ASR, MT, or TTS components are told to abort and prepare for new input. If a component's abort procedure has latency that makes it unable to begin processing the new input immediately, the system simply queues the input until the component is ready. This process is invisible from the outside; at most all that will be observed is a slight delay in the speech translation output.

For the YOU button (SME voice input), the system first speaks a confirmation utterance before speaking. To determine the confirmation utterance, the system first tries to match the ASR output to one of the canonical English utterances for which it has a recorded translation. The confirmation utterance is the canonical utterance if one is found, and otherwise just the ASR output. The system then begins speaking the confirmation via English TTS. In the case that a canonical utterance was not found, it also begins to E2I MT on the ASR output, so that translation can be translated in parallel while the confirmation utterance is being spoken..

If the confirmation utterance satisfies the SME, he simply lets the system proceed to produce the spoken Iraqi translation. If the confirmation was not correct, or if the SME simply changes his mind about what he wants to say, he presses the YOU button and speaks a new utterance. This causes the system to abort out of whatever stage of the processing it was in, and start a new ASR episode. If the SME wants to abort without speaking, he simply presses the YOU button and releases it immediately, generating an empty ASR output that the system discards. In this way, the YOU button also performs the function of the "Abort" button of the GUI system.

Accepting the confirmation thus requires no interaction at all, and is essentially instant. Rejecting the confirmation requires just a button press, but no specific dialog, so it is also very fast. All that is spoken is the new utterance. This strategy of assuming the user approves if he does not intervene is termed "implicit confirmation" [5], as opposed to "explicit confirmation", in which the system explicitly asks the user if he approves or not.

Once the confirmation utterance completes, the system can speak the Iraqi translation as soon as the E2I MT result is available. But because the E2I was running in parallel while the confirmation was being spoken, it will almost always be done by this time, thus allowing the Iraqi TTS to begin at once. Because the E2I would have to be done anyway, whether or not confirmation was used, running E2I in parallel with confirmation subtracts the E2I processing time from the time the confirmation actually costs the dialog. Effectively, the system is speaking the confirmation utterance faster.

Of course, the fact that the English ASR output is correct does not guarantee that it will be correctly translated to Arabic. An alternative would be to translate the E2I output back into English, and use that for the confirmation. We have found these back-translations to be very noisy, however, and to often appear wrong even for adequate E2I translations. By contrast, if the English ASR has serious (concept word) errors, the resulting translation to Arabic is sure to be wrong, so the ASR error criterion is at least more precise. It is also a much easier criterion for non-technical users to understand and apply.

3. Speech Recognition

The BBN Byblos speech recognition system [6], models speech as the output of context-dependent phonetic Hidden Markov Models (HMMs). The outputs of the HMM states are mixtures of multi-dimensional diagonal Gaussians. Different forms of parameter tying are used in Byblos, including State Tied Mixture (STM) triphone and State Clustered Tied Mixture (SCTM) quinphone models. The mixture weights in both these cases are shared based on the decision tree clustering.

Recognition is performed using a two pass search strategy [9]. The forward pass is a fast match beam search using an STM acoustic model and an approximate bigram language model. The output of the forward pass consists of the most likely word-ends per frame along with their partial forward likelihood scores. The backward pass operates on the set of choices from the forward pass to restrict the search space, and uses the more detailed SCTM quinphone model and a trigram language model to produce the best hypothesis.

The Iraqi Arabic acoustic models were trained on 405 hours of speech collected under the TRANSTAC effort. This consists of 1.5-way (simple Iraqi answers to questions) and 2-way (full dialog) data of colloquial Iraqi speech from the force protection domain. The baseline acoustic models were estimated in the Maximum Likelihood (ML) framework. The models were further improved with lattice-based discriminative training. Both the Maximum Mutual Information (MMI) [7] and Minimum Phoneme Error (MPE) [8] criteria were explored for improving the acoustic models.

The language models were trained using approximately 2.8 million words of data and a lexicon of 80k words. Phonetic word pronunciations were written using a set of 39 phonemes, and we used the graphemic approach introduced in [11]. The language model employs Kneser-Ney smoothing [12].

Table 1 shows the results on a held-out test set of approximately 12 hours of speech. MPE gives 9.4% relative improvement in WER over the baseline ML models. All forms of the glottal stop, or "hamza", were normalized for WER computation.

Configuration	**%WER**
Maximum Likelihood	35.2
Maximum Mutual Information	33.7
Minimum Phoneme Error	31.9

Table 1: Iraqi WER for different acoustic model estimation criteria

Table 2 shows the results on the 1.5-way and 2-way test sets of the TRANSTAC March 2006 Offline Evaluation. A smaller dictionary (~60K) was used than the one used to generate the results in Table 1. The table also reports results with online speaker adaptation, where adaptation statistics are continuously updated [13]. A 23% relative gain in WER was obtained via adaptation.

Decoding	%WER	
	1-5way	2-way
Unadapted	25.7	25.7
Adapted	19.8	19.6

Table 2: WER for March 06 Offline Iraqi test set

The English recognizer has the same configuration as the Iraqi recognizer. The English acoustic models are trained on 110 hours of speech. The language model is an interpolation of 4.4 million words of in-domain data, and 41 million words of out-of-domain data. The lexicon contains 19K words and a set of 52 phonemes. The results on the held-out test set of 1 hour of speech are shown in Table 3. MPE models give a gain of 22.5% relative over the ML models.

Configuration	%WER
Maximum Likelihood	29.8
Maximum Mutual Information	25.3
Minimum Phoneme Error	23.1

Table 3: Improvements in English WER using different acoustic model estimation criteria

The results on the March 2006 Offline Evaluation test sets are shown in Table 4. Unsupervised adaptation gave a gain of 30% relative on the 2-way offline set. The greater improvements for English than for Iraqi may result from the lower perplexity (55 vs. 145) and better LM for English.

Decoding	%WER	
	1-5way	2-way
Un-adapted	10.5	10.9
Adapted	7.2	7.5

Table 4: Results on March 06 Offline English test set.

4. Translation

BBN's Statistical Machine Translation (SMT) engine is a phrase-based translation system based on the noisy channel model [1]. Given, an input foreign language sentence 'f', we estimate the most likely translation into the target sentence 'e' as:

$$\hat{e} = \arg\max_{e} P(e \mid f) \quad (1)$$

Word alignments between source-target sentence pairs are generated using GIZA++ based on IBM's Models 1 to 4 [15]. In order to improve the quality of the alignments, word alignments in the forward and backward direction are merged as in [16]. Phrase pairs are automatically extracted from the word alignments by merging neighboring alignment groups using a set of rules. The decoder uses a log-linear model of different features to choose between competing translation hypotheses. The parameters of the model are estimated using statistics of the phrase pairs extracted from the word alignments. The interpolation weights are optimized by minimizing the translation errors on a held out development set.

Table 5 shows the corpus used for training, development and test purposes.

Iraqi → English					
#English Words		**#Arabic Words**		**#Utt Pairs**	**Set**
Unique	**Total**	**Unique**	**Total**		
36K	2.7M	75K	1.9M	440K	Train
8K	116K	13K	80K	15K	Dev
7.5K	105K	11K	72K	14K	Test
English → Iraqi					
#English Words		**#Arabic Words**		**#Utt Pairs**	**Set**
Unique	**Total**	**Unique**	**Total**		
6.2K	329K	24K	224K	29K	Train
1.4K	11.5K	2.7K	7.5K	1K	Dev
1.8K	18.3K	3.8K	11.7K	1.6K	Test

Table 5: MT Training and Test Data

Detailed results and new research on the MT component are reported on in [17].

5. Evaluation Results

The system described here was evaluated as part of the program-wide TRANSTAC evaluation, held in January 2007 and conducted by the National Institute of Standards and Technologies (NIST). The evaluation had three components: offline measurements of ASR and MT component performance, live interactions in a quiet lab environment, and live interactions in a "field" indoor environment with intermittent noise. USMC personnel with experience in Iraq played the part of the SMEs, and Iraqi expatriates played the part of the FLEs. The SME and FLE used the system to work through 5 scripted and 10 "structured" scenarios, including sewage surveys, job recruiting, and checkpoint dialogs. For scripted scenarios, the parties were told what to say; while for structured scenarios, they were simply given a description of the information to be transferred. Each scenario had a time limit of 10 minutes to transfer 30 specific pieces of information. Table 6 gives the average number of concepts successfully transferred per scenario. For the Lab and overall conditions, the BBN system achieved the best results of all systems.

Lab	Field	Overall
21.7	19.7	20.5

Table 6: High-Level Concept Transfer

The automated measures WER, BLEU, Translation Error Rate (TER) [18], and METEOR [19] were computed on an offline data set for MT on ASR output (S2T), and MT on ASR transcriptions (T2T).

		WER	BLEU	1-TER	MET
I2E	S2T	27.2	48.0	54.9	67.4
	T2T	---	61.9	67.3	77.5
E2I	S2T	20.3	31.0	46.3	48.2
	T2T	---	42.3	56.4	58.8

Table 6: Automated Metrics on Offline Set

NIST also performed a detailed analysis of the transfer of lower-level lexical and phrasal concepts for the scripted live interactions and the offline data set. A concept transfer odds measure was computed, defined as the ratio of the number of successfully transferred concepts to the sum of inserted, deleted, and substituted concepts. These results are given in Table 7. Results on the offline set are lower, because the data there is less focused and transactional than the live interactions. All are the highest numbers for all systems.

Scripted	Offline	Overall
3.9	2.9	3.1

Table 7: Low-Level Concept Transfer Odds

One of the goals of our work was to implement audio confirmation efficiently. For the online lab evaluation, the average duration of confirmation utterances was 2963 ms while the average time to perform E2I MT was 1247 ms. Performing E2I in parallel with confirmation thus reduced the effective cost of confirmation to 1716 ms per utterance, just 58% of the literal duration. Notably, the average time to successfully transfer a high-level concept in structured scenarios was 23 seconds, which was tied for fastest overall with a system that did not perform confirmation.

SMEs also evaluated systems subjectively, along several dimensions of usability. Since we have tried to address usability issues in this work, this is an important aspect of the evaluation. Survey questions were expressed in the form of a statement, such as "It was easy to get the information I needed", "The system worked the way I expected it to", and "I found the system easy to understand". The SME would respond to with a 1 – 5 Liker rating, with 1 meaning "strongly disagree", 4 "agree", and 5 "strongly agree". The BBN system achieved scores ranging from approximately 4.2 to 4.7, which were the highest scores of all evaluated systems on these measures.

6. Acknowledgements

The work reported in this paper was performed under the DARPA TRANSTAC Program.

7. References

[1] D. Stallard, F. Choi, P. Natarajan, R. Prasad, and S. Saleem, "A Hybrid Phrase-Based/Statistical Speech Translation System," Proc ICSLP 2006, Pittsburgh, PA, September 2006.

[2] B. Zhou, S. Chen, and Y. Gao, "Constrained Phrase-Based Translation Using Weighted Finite-State Transducers," Proc ICASSP 2005, Philadelphia, PA. March 2005 pp. 1017-1020

[3] A. Kathol et al., "Speech translation for low-resource languages: the case of Pashto," Proc. INTERSPEECH-2005, 2273-2276.

[4] Schulz, T., and Black, A., "Challenges with Rapid Adaptation of Speech Translation Systems to New Language Pairs," Proc. ICASSP 2006, Toulouse, France, May 2006

[5] P. Heisterkamp, "Ambiguity and uncertainty in spoken dialogue," in Proceedings of EUROSPEECH Conference, pp. 1657-1660, Berlin, Germany, 1993.

[6] S. Matsoukas, R. Prasad, S. Laxminarayan, B. Xiang, L. Nguyen, R. Schwartz, "The 2004 BBN 1xRT Recognition Systems for English Broadcast News and Conversational Telephone Speech," Proc. EUROSPEECH 2005, Lisbon, Portugal, Sep. 2005

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

[8] D. Povey and P. C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," ICASSP, Orlando, FL, May 2002.

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

[10] D. B. Paul, "An Investigation of Gaussian Shortlists," Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Dec. 1999

[11] J. Billa et al., "Audio Indexing of Arabic Broadcast News", Proc. ICASSP, Orlando, FL, May 2002.

[12] R. Kneser and H. Ney, "Improved Backing-off for m-gram Language Modeling," Proc. ICASSP, IEEE, pp. 181-184, 1995.

[13] D. Liu, D. Kiecza, A. Srivastava, F. Kubala, "Online Speaker Adaptation and Tracking for Real-Time Speech Recognition," Proc. Interspeech 2005, Lisbon, Portugal, pp. 281-284, 2005.

[14] P. Koehn, F. Och and D. Marc, "Statistical Phrase-Based Translation," Proc. of the HLT and NAACL Conference, 2003

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

[16] F. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," Computational Linguistics, 29(1), 19 - 51, 2003

[17] S. Saleem, R. Prasad, K. Subramanian, D. Stallard, C. Kao, R. Suleiman, P. Natarajan, "Improvements in Machine Translation for English/Iraqi Speech Translation", Interspeech, 2007

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

[19] S. Banerjee, A. Lavie,. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005