The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System

Nguyen Bach, Matthias Eck, Paisarn Charoenpornsawat, Thilo Köhler, Sebastian Stüker, ThuyLinh Nguyen, Roger Hsiao, Alex Waibel, Stephan Vogel, Tanja Schultz, Alan W Black

InterACT, Language Technologies Institute,
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
awb@cs.cmu.edu

Abstract

The paper describes our portable two-way speech-to-speech translation system using a completely eyes-free/hands-free user interface. This system translates between the language pair English and Iraqi Arabic as well as between English and Farsi, and was built within the framework of the DARPA TransTac program. The Farsi language support was developed within a 90-day period, testing our ability to rapidly support new languages. The paper gives an overview of the system's components along with the individual component objective measures and a discussion of issues relevant for the overall usage of the system. We found that usability, flexibility, and robustness serve as severe constraints on system architecture and design.

1. Introduction

In our continuing efforts to construct practical two-way speech-to-speech translation systems, we have developed a generic hands- and eyes-free portable speech translation system which allows an English speaker to converse with a target language speaker.

Our systems have been evaluated on a regular basis as part of the DARPA TransTac program. These evaluations are run by NIST, and involve military users and target language users who have never used our system before. The evaluations consist of communicating through the translation device for a number of pre-designed scenarios (which were previously unknown to us). The tests take place both indoors and outdoors. Other systems in the TransTac program include those developed by BBN [5], IBM [6], SRI, Sehda/Fluential, and USC [7] [8].

2. Challenges

The two target languages in the TransTac program are Iraqi Arabic and Farsi. Iraqi Arabic is defined as the spoken form of Arabic used by the people of Iraq in everyday conversations. It is distinct from the formal Modern Standard Arabic (MSA) used in written communication. As Iraqi Arabic is normally not written, even with transcription conventions there is greater variability in the spelling conventions than in a standard written language. Farsi (Persian), mainly spoken in Iran and areas of Afghanistan, also uses the Arabic script, though is not a Semitic language.

In general, the target domain is "force protection" which includes checkpoints and house-hold searches, and extends to civil affairs, medical, and training. DARPA has collected a number of spoken dialogs within these domains that have been provided to the participants of the TransTac program.

Spoken language translation for the languages Iraqi and Farsi is challenging due to a number of reasons, such as the availability of larger amounts of suitable speech and text data, the lack of extensive language and cultural expertise, language peculiarities in large, and the lack of language convention [3].

Two-way translation between rich inflectional morphology languages, such as Iraqi and Farsi on the one side, and languages with a relatively poor inflectional morphology such as English on the other side, presents several challenges to the existing components of a speech-to-speech translation system. This inflection gap causes an abundance of surface word forms in Iraqi and Farsi compared with relatively few forms in the English language. This mismatch introduces several issues into natural language processing, such as a large number of unknown word forms in unseen data, many words occurring only once in the corpus, and many distinct words. As we have had more than one year to develop the Iraqi system but only 90 days for the Farsi system, the lack of suitable data appears more severe in the Farsi system than in the Iraqi one.

One major key engineering problem is the design of an effective, efficient, and easy-to-use real-time two-way eyes- and hands-free interface. The system needs to be robust, and portable for tactical use in the field. The primary use cases involve US military personnel and local foreign language speakers. While the military personnel will be trained to use the systems, the foreign language users are assumed to have little or no chance to become familiar with the system. Unfortunately, this important usability aspect is very difficult to assess. Our very limited cultural expertise and insights into the community of the languages in question make this task even harder. We are aware that a system which is developed by optimizing

individual components toward the respective metrics such as Word Error Rate (WER) and BLEU does not fully reveal shortcomings in the overall usability of the final system.

Finally, high environmental noise levels found to be present in the real-world application are major obstacles, especially for automatic speech recognition (ASR), and text-to-speech output (TTS).

3. System Architecture and Design

Our system is designed for eyes-free/hands-free use. This means there is no display or any other visual feedback, so that the user can focus during operation on his surrounding environment. The complete functionality - translation, control of the device, error recovering - must use human speech and audio signals alone. Additionally, the system provides two user modes: an automatic mode and a manual mode. In the automatic mode, the system will automatically detect speech segments of the two open microphones, and translate it in both directions. In the manual mode, the system opens the microphones only if the user pushes and holds a button, allowing more restrictive control over the translated utterances. Consequently, the system can be used in a hands-free operation mode, or semi-hands-free where one hand is needed while speaking, but both hands are free while listening. Furthermore, the system accepts various voice commands for system control, such as playing back instructions for the inexperienced user, repeating the last translated utterance, disabling/enabling the translation or switching between the automatic and manual mode.

3.1. Software components

As shown in Figure 1, there are 8 main system components composed of two audio segmenters, two automatic speech recognition (ASR) systems, two text-to-speech (TTS) systems, a bidirectional machine translation system (MT) and one FrameWork.

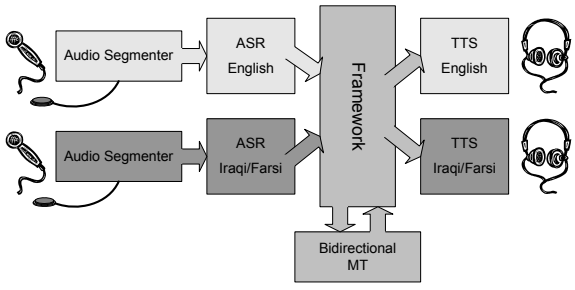

Figure 1: Software components

- The audio segmenter is a tool recording a continuous signal from a microphone. It can automatically detect, segment, and stream incoming speech to one of the ASR components for further processing. Additionally, it can be manually operated by using a push button for walkie-talkie like recording. The ASR component recognizes speech and produces a hypothesis of the transcription.
- The TTS component synthesizes speech from text. This can be the translation, but also the feedback to the speaker of the ASR result or other system information.
- The bidirectional MT component translates text from English to Iraqi/Farsi and vice versa. For the Iraqi system, only statistical machine translation (SMT) is used, while the Farsi system uses a combination of a phrasebook translation for a fast and reliable translation of high frequency sentences, and an SMT translation for more complicated sentences, due to data sparseness.
- The FrameWork is the program used to synchronize all the above components. It checks whether a result from the ASR is a voice command or an utterance for translation. In case of a voice command, the program will process that command. For example, if the user says "TransTac Instructions", the system will play the pre-recorded instructions for the foreign language speaker. In case of an utterance for translation, it sends the ASR result to the TTS while asking the MT component for the translation. Table 1 shows the list of voice commands and their functions After the MT and the TTS component are done, it sends the translation to the other TTS component.

Table 1: Voice command list

Voice Commands	Functions
transtac instructions	play pre-recorded instructions to the foreign language speaker
transtac say again	repeat the last translation
transtac say recognition	say the ASR result of the last utterance
transtac say translation	say the back-translation of the last utterance
transtac automatic mode	switch to automatic mode (hands-free mode)
transtac manual mode	switch to manual mode (push-to-talk)
transtac stand by	turn translation off
transtac listen	turn translation on
transtac status	report system ready and mode information

3.2. Hardware components

The system is composed of two microphones, two buttons, one loudspeaker, and one laptop. Due to the eyes/hands-free concept, the laptop is kept in a backpack, the remaining items are attached to a vest that the user wears as shown in Figure 2.

- Two microphones: one for the English speaker and the other for the Iraqi/Farsi speaker.

- Two buttons are used when the system is in manual mode. One button is for the English speaker and the other is for the Iraqi/Farsi speaker.
- The laptop is stored in the backpack with a simple fan-driven cooling system that prevents the laptop from overheating.
- The loudspeaker is located in the front pocket, so that the users clearly hear the speech output.

Figure 2: Hardware components

4. Automatic Speech Recognition

Compared to our last year's speech to speech translation system [1], there are some significant developments which affect the design and performance of the ASR component.

Table 2: Comparison of 2006 and 2007 ASR systems

	2006 System	2007 System
Platform	PDA	Laptop
Vocabulary	5 – 10k	(20 – 60k)
Real-time factor	2 – 5	1
Input device	Built-in Mic	Headset Mic
N-gram LM	3-gram	3-gram
# Gaussians AM	32 – 48k	200 – 400k
AM Training	ML	MMIE

First, our new 2007 system is designed to be hands-free and eyes-free, i.e. it intentionally lacks any graphical user interface (GUI). The motivation of this design is to allow effective communication in tactical environments – through speech alone. Consequently, very robust and highly reliable ASR performance is crucial for the success of the overall system. Second, we focused on the laptop platform, while our 2006 system was designed for PDAs with very limited resources. Given more computing power and memory resources, we built an ASR system with more parameters for acoustic and language model and thus better performance compared to the PDA-based system. Due to the much larger vocabulary the system can also support broader scenarios and allow more flexible conversations.

Our ASR system uses the Janus Recognition Toolkit (JRTk) featuring the IBIS decoder [16]. Table 2 lists the major differences between our last year's 2006 ASR system and this year's 2007 ASR system.

4.1. English ASR

The English ASR system utilizes 3-state sub-phonetically tied fully-continuous Hidden Markov Models composed of 4000 models with a maximum of 64 Gaussians per state and a total of 234K Gaussians. The preprocessing stacks 15 sequential, 13 dimensional vectors of Mel Frequency Cepstral Coefficients (MFCC) and reduces the resulting 195 dimensional vector to 42 dimensions with the help of discriminant analysis (LDA). The acoustic model was trained on 138h of American Broadcast News data and 124h Meeting data, using merge and split training on training samples extracted with the help of forced alignments and one global semi-tied covariance (STC) transformation. This was followed by two iterations of Viterbi training to compensate for potential errors in the forced alignments. The resulting model was then MAP adapted on 24h of data from DLI provided within TransTac. The system uses utterance based Cepstral Mean Subtraction (CMS) during training and incremental CMS and feature space constrained MLLR (cMLLR) during decoding to adapt to the current speaker.

The language model used in decoding is a trigram model with approximately 3.5M trigrams and a vocabulary size of 11k words. The language model was trained on several text corpora in force protection domain, which sum up to 1.7M words. A subsequent interpolation with a language model trained on 65M words of web data was performed, in order to gain a wider coverage of bi- and trigrams, but not to increase the vocabulary.

Using our own scoring script, the system achieves a WER of 25.2% on the TransTac January 2007 offline evaluation data, and less than 10% on the data recorded with the CMU system during the lab and field evaluation, while people were actually using our system. The difference in performance on the two sets comes from two different sources. First, the segmentation of the offline audio data was slightly corrupted and often too tight for our system. Second, the data was retrieved from a human – interpreter – human conversation, resulting in sloppier and more complicated speech than users of a machine translation device would do.

4.2. Iraqi Arabic ASR

The acoustic model (AM) of the Iraqi ASR system is a 3-state sub-phonetically tied semi-continuous HMM-based recognizer composed of 5000 context dependent triphone/quintphone models. Each model consists of a mixture of 64 Gaussians at the most, where the exact number of Gaussians is determined by a merge-and-split

training algorithm. Input speech is represented by the first 13 Mel Frequency Cepstral Coefficients (MFCC) and power, together with approximations of the first and second derivatives. Linear discriminant analysis is applied to reduce the dimensionality to 42 coefficients.

The acoustic model is trained with 320 hours of Iraqi Arabic speech data including data sets from Appen/BBN, Cepstral, IBM/DLI Pendleton, and Marine Acoustics Inc. The language model (LM) for Iraqi ASR is a trigram model using modified Kneser-Ney smoothing. The training set consists of 2.2M words including data from different domains in force protection and medical processing such as, common community interest, medical screening, traffic control check points, and other less restricted topics, such as rapport building. The 62K vocabulary is based on frequency count.

Table 3: Characteristics of Iraqi ASR System

Iraqi ASR	2006 System	2007 System
Vocabulary	7k	62k
# AM models	2000	5000
#Gaussians/ model	≤ 32	≤ 64
Acoustic Training	ML	MMIE
Language Model	3-gram	3-gram
Data for AM	93 hours	320 hours
Data for LM	1.2 M words	2.2 M words

A discriminative training algorithm, maximum mutual information estimation (MMIE), is applied on the acoustic model to further improve the system [17]. Table 3 lists the differences between our last year's 2006 and this year's 2007 Iraqi ASR system.

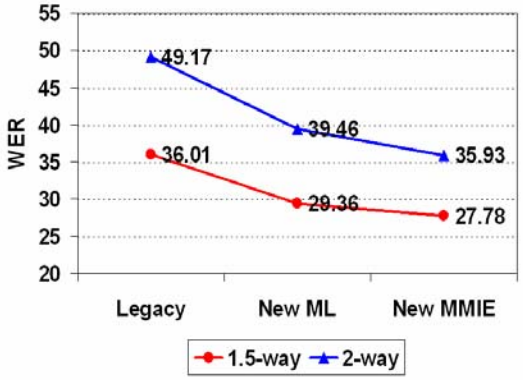

Figure 3: Performance of the Iraqi ASR system

To evaluate the Iraqi ASR, we kept held out data from the data sets collected in 2006 by DLI and Appen. These test data are separated into two sets: 1.5-way data and 2-way data. 1.5-way data consist of mainly basic questions and answers while the 2-way data set is conversational speech. Each data set has roughly an hour of speech data. To allow for a fair comparison between the 2007 Iraqi ASR and our 2006 legacy system, we retrained the latter system with the new data, and increased the vocabulary such that it is consistent with the new system. In addition to increasing the amount of speech and text training data and vocabulary, we also measured the impact of applying MMIE discriminative training. Figure 3 shows that our new Iraqi 2007 ASR system outperforms the 2006 legacy system, and that the MMIE discriminative training further reduces the WER by 5–10% relative. The Out-Of-Vocabulary rates on the 1.5-way and 2-way data sets are 1.6% and 2.0% respectively.

4.3. Farsi ASR

The Farsi acoustic model has the very same topology as the Iraqi ASR system. It is trained with about 110 hours of Farsi speech data collected by Appen, DLI, and University of Southern California. The acoustic model consists of 3000 models, each has at most 64 Gaussians which is determined by merge-and-split training. The acoustic model was bootstrapped from the Iraqi acoustic model. The two phones of Farsi not covered by the Iraqi phone set were initialized by phones of the same phone category. After this phone mapping a first Farsi context independent acoustic model was bootstrapped from the Iraqi acoustic model. This first Farsi context independent system was used to force-align all data. Based on these new forced alignments we initialized a second context independent system before using our regular context dependent training routine. Finally, MMIE discriminative training was applied to further reduce the Word Error Rate.

The language model is a trigram model using modified Kneser-Ney smoothing, and it is trained with 900K words. The vocabulary size is around 33K words, which consists of all available words in the provided corpora.

Similar to the Iraqi ASR, we kept a 1.5-way data set and a 2-way data set for testing. Those two data sets are roughly one hour each. Table 4 shows the performance of the Farsi ASR system. Similar to the Iraqi ASR, MMIE gave around 5–10% relative WER reduction.

Table 4: Performance of Farsi ASR [WER in %]

Farsi ASR	ML built	MMIE built
1.5-way	28.73	25.95
2-way	51.62	46.43

5. Machine Translation

5.1. Dialog structure

The typical dialog structure in TransTac is based on the English speaker trying to gather information from or give information to the Iraqi/Farsi conversation partner. This means the English speaker mainly uses questions or gives instructions and commands. The Iraqi/Farsi speaker generally gives relatively short answers to these questions. Dialogs are from the military, relief, and medical domain as described above. Some example for typical commands

and questions with corresponding answers are given in Table 5.

Table 5: Examples of dialog structure

English speaker	Farsi/Iraqi speaker
Do you have electricity?	
	No, it went out five days ago
How many people live in this house?	
	Five persons.
Are you a student at this university?	
	Yes, I study business.
Open the trunk of your car.	
You have to ask him for his license and ID.	

5.2. Training & Testing Data

Table 6 shows an overview over the Training data that was available for Iraqi ↔ English. The Training data was provided in separate corpora depending on the language of the original utterance during the data collection (which was later translated to the other language). This means the English→Iraqi part mainly contains the typical sentences of the English speaker while the Iraqi→English part typically contains the answers. In both cases we see that the English sentence contains about 1.4 times the number of words of the Iraqi sentence (after preprocessing).

Table 6: Iraqi training data statistics

	Source	Target
Iraqi→English		
Sentence pairs	502,380	
Unique pairs	341,149	
Average length	5.1	7.4
Words	2,578,920	3,707,592
English→Iraqi		
Sentence pairs	168,812	
Unique pairs	145,319	
Average length	9.4	6.7
Words	1,581,281	1,133,230

Table 7: Farsi training data statistics

	Source	Target
Farsi→English		
Sentence pairs	56,522	
Unique pairs	50,159	
Average length	6.5	8.1
Words	367,775	455,306
English→Farsi		
Sentence pairs	75,339	
Unique pairs	47,287	
Average length	6.7	6.0
Words	504,109	454,599

We used two test sets from the TransTac January 2007 evaluation. There are 415 sentences in the English→Iraqi test set and 433 sentences in the Iraqi→English test set. For both test sets 4 reference translations were available.

Table 7 shows the statistics of the parallel data in Farsi→English and English→Farsi. The development sets of Farsi→English and English→Farsi are extracted from the training data, and have only one reference. To generate unseen test sets for the Farsi system, the English side of the January evaluation set was manually translated.

5.3. Data normalization

The MT component interfaces between ASR and TTS, hence data normalization for Iraqi, Farsi, and English plays an important role in the S2S system.

The major goal of data normalization is to minimize the mismatch in vocabulary between ASR, MT, and TTS components. The mismatch in vocabulary can happen due to a number of reasons such as different preprocessing steps in individual components, lack of data, or different encoding of the same orthography form. To fulfill the requirement data normalization steps need to be agreed upon. However, it is not easy to reach a consensus since Iraqi Arabic lacks a standard writing system.

Furthermore, there are issues with speaking style. Words can be used with their formal or informal/colloquial endings for example *raftin* vs. *raftid* "you went". The word forms (inside of the word) may be modified to represent their colloquial pronunciation for instance *khune* vs. *khAne* 'house', *midam* vs. *midaham* 'i give').

5.4. Language models

The language model is a standard 6-gram language model with Good-Turing smoothing implemented as a suffix array (SA LM) [10]. Another option of language model is the 4-gram modified Kneser-Ney smoothing trained using the SRI language modeling toolkit (SRI LM) [11].

5.5. Translation models

5.5.1. PESA phrase extraction

In Iraqi-English we applied the PESA phrase extraction method [9]. For a given source phrase PESA tries to find the optimal sentence splits of the training sentences containing this source phrase based on inner and outer IBM1 word alignment probabilities. We applied PESA as an online phrase extraction which means that phrase pairs are dynamically extracted from the training data as needed during the translation of the test set. We compared the performance here with a standard Pharaoh phrase table but we saw considerable improvements using the PESA approach.

For Iraqi-English a considerable amount of training data is available and parts of the test dialogs are repetitive which leads to the fact that we actually find an unusual number of longer ngrams. The phrase pair extraction in the Pharaoh

toolkit uses the same length restriction for source and target side. However, for Iraqi to English translation we typically need to generate more English words then we have on the source side. This makes many of the phrase pairs with 6 or 7 source words un-usable, as they do not give the correct translations of 8, 9, or more words. PESA is able to match arbitrarily long phrases and this happens quite often in typical TransTac data.

5.5.2. Combined online and offline phrase extraction

One problem with the online phrase extraction method is that it is relatively slow, taking up to 20 seconds for long sentences. To shorten this time we use a combined approach. A pre-extracted phrase table for shorter phrases and the most common source phrases was combined with the online phrase extraction.

For these pre-extracted phrases the online phrase extraction does not have to extract the phrases pairs dynamically. Instead, the online phrase extraction is only used for long or rarely seen phrases. This did not give any significant change in performance but resulted in a considerable speedup. The system uses the same corpora to extract online PESA phrases for both translation directions so we combined the Iraqi-English and English-Iraqi corpora for this. However, the pre-extracted phrases were extracted separately for each direction from the respective corpus.

5.5.3. Interpolate Pharaoh and PESA

We observed that the Pharaoh [19] phrase table does not contain entries for all words in the source vocabulary. This comes from the heuristics applied to avoid unlikely translations. Therefore, some words will not be translated, even though they appear in the training corpus, because they occur in the phrase table only embedded in longer phrases. This leads to an unnecessary high number of untranslated words. On the other side, the PESA phrase alignment will generate translations for all n-grams including all individual words, which can be found in the training corpus.

To guarantee that the phrase table can cover all source vocabulary and to leverage the PESA's strength in arbitrary long matching, we trained two phrase tables and interpolated them. The interpolation parameters are optimized through a minimum-error-rate training framework [12].

5.5.4. Speed constraint

To limit delays, the translation has to be performed during the replay of the ASR output. This has to be the case for even very long sentences. For all practical considerations we assume to have about 200 ms on average to do the translation.

Some of speeding strategies we applied is phrase table pruning and restrict the search space during the decoding process. Those techniques help to decrease the system running time significantly.

5.5.5. Decoder

For this evaluation the system is running on a standard laptop with 2 GB of memory so we could use our regular decoder [2]. The previous system described in [1] was running on a PDA. Due to lack of memory and computing power an earlier version of the decoder described in [18] had to be used that did not support word reordering and required heavily pruned models.

5.5.6. Translation results

We report the performance of translation component in terms of BLEU score [20]. On the test sets the system achieved a score of 42.12 for English to Iraqi and 63.49 for Iraqi to English.

The Farsi systems use similar technologies as the Iraqi systems. Table 8 shows the translation performance of the provided training data on various setups.

Table 8: Farsi translation performance (in BLEU)

Farsi→English		
	Dev.	Unseen
Pharaoh + 4-gram SRI LM	24.64	23.3
PESA + 6-gram SA LM	23.06	19.9
English→Farsi		
Pharaoh + SRI LM	10.07	14.87
PESA + SA LM	9.45	14.67
Pharaoh + SA LM	10.41	15.42
Pharaoh + PESA + SA LM	10.23	16.44

6. Text-to-Speech

Text-to-speech was provided by Cepstral, LLC's SWIFT speech synthesis engine. The Iraqi voice was created two years ago when Iraqi became the chosen language within the TransTac program.

Farsi was introduced to the program this year and offered an opportunity to test our skills at building understandable synthesis in a new language at short notice. In order to build a synthetic voice it is necessary: to design a speech database that contains adequate phonetic and prosodic variation found within the language; record the data from a native speaker; construct a pronunciation lexicon; and build the synthetic voice itself.

SPICE (Speech Processing - Interactive Creation and Evaluation Toolkit for new Languages) is an NSF sponsored project that aims to significantly reduce the amount of time and effort involved in building speech and language processing systems for new languages [13]. SPICE provides web-based tools that enable non-expert users to develop speech and language processing models, collect appropriate speech and text data, as well as evaluate the models' performance allowing for iterative

improvements. Using transcribed prompts from recorded examples of Farsi speakers as provided by DARPA as part of the program, SPICE searches for utterances of moderate length containing high frequency words with the largest trigram letter variation possible. Our goal is to get easy-to-read sentences (so the voice talent does not make an error), that most likely cover the phonetic space. At time of prompt selection we did not yet have a pronunciation lexicon, so selected our prompts for letter rather than phoneme coverage. Using an initial text database of 6MB characters we selected 14,000 sentences that were between 4 and 15 high frequency words (approximately one sixth of the total database). We then selected 2021 prompts that had optimal letter context coverage.

We then passed this list to a native Farsi speaker who checked the list and noted sentences that would be difficult to say. The "error" sentences were mistyping, unclear grammar, or just strange. As these sentences were automatically selected from transcriptions of natural speech some may be strange to say even though they have interestingly varied letter contexts. This further reduced the prompt list from 2021 to 1905. These 1905 sentences were then recorded by a male native Farsi speaker in a professional studio. We continue to note that finding suitable target language speakers can take a significant time in the whole voice building process. Finding and setting up the recording process took over a month, thus around half the time required for the voice construction process was taken up by finding a suitable speaker and recording environment.

A further database reduction occurred due to errors in delivery, giving us a final set of 1815 prompts, about 107 minutes of speech. This was then used to build a unit selection synthesis voice using Cepstral's VoiceForge (TM) build process [14].

In addition to the recorded database a pronunciation lexicon was required. This was based on USC's Farsi Pronunciation Lexicon. It was converted from its romanization to Arabic script. Letter-to-sound rules were trained from this data to provide pronunciations for unknown words. We used our standard CART-based techniques [15], though used a new fully automatic grapheme/phoneme alignment technique to find initial alignment. Hence the letter to sound rules could be built without any knowledge of the target language. For a held out set of words these models produce 77% word correct, which we note is better than our Iraqi prediction (68%). Farsi is probably easier than Iraqi Arabic as it has a standardized spelling. This lexicon and letter-to-sound rules was used to generate the pronunciations for both the TTS engine and the ASR pronunciation lexicon.

7. Practical Issues and Conclusions

As noted, the user interface is very important in a usable translation system. In order to inform the user what was happening and to indicate the likelihood of a successful translation, we echo back the recognized text using the text-to-speech system. Thus the user hears what the system is going to translate. When users detect errors they learn to repeat or rephrase their sentences. In an earlier version of the system we echo back the “back-translation” to the user (that is we recognized, translated then translated back into the speaker’s language). Although as developers we found this information very useful, it clearly confused the users. They did not understand for example why “car” was echoed as “vehicle”. They also would mimic the non-grammatical output (resulting from translating twice) lowering the overall performance of the system.

We offered many options to the users but without experience they could not decided what may be best for them. Although we offered hands-free communication, most users disliked that everything spoken was translated, and therefore picked the manual push-to-talk mode over the automatic open microphone mode.

As described above we use the time during play back of the recognized speech for the translation component. The system seems fast, with no obvious delays. Although the time taken to echo back the translation does slow down the overall system, we feel this feedback is useful.

One observation that came up with the text to speech output, was that some users said it was too fast to understand. This we believe is related to the fact that speech synthesizers are designed to speak fluent speech, while the output of an MT system may not actually be fully grammatical. [4] showed how understandability of text to speech falls with machine translation output when compared to human MT output. This issue, we believe may be attenuated if the text string was marked up into better phrases, even down to isolated words. The “too fast” comment is not just that speed of each phoneme but with more phrase breaks in the speech there would be more time for the listener to understand it. However we have not pursued this direction yet.

Another observation when working on rapid development of speech translation components is how to use language expertise efficiently and effectively. We followed the 10 suggestions in [3] to assign tasks for native speakers. We found that having a native speaker analyze the output of the prototypical component systems and identify the possible error sources related to the language is crucial. This step helps technology experts improve their components significantly.

The user interface is another important factor for a successful device. It needs to be simple and clear so that the time for training a new user is minimal. Our system adopts a push-to-talk mechanism and a simple protocol which allow users start using the device after a few minutes of training.

Putting the whole system in a backpack with a vest to carry the accessories makes the system neatly self-

contained. However, there were other engineering issue we had to address. Although there is no over-heating issue when used in a air-conditioned office we added a USB fan to the system to increase airflow through the pack. This was sufficient to keep the system cool for the 3 hour use in the 85°F (~30°C) outdoor evaluations.

8. Acknowledgements

This work is in part supported by the US DARPA under the TransTac (Spoken Language Communication and Translation System for Tactical Use) program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

We would like to thank Akram Kamrani for helping us in the Farsi systems. We would also like to thank Cepstral LLC and Mobile Technologies LLC, for support of some of the lower level software components.

9. References

[1] Hsiao, R., Venugopal, A., Kohler, T., Zhang, Y., Charoenpornsawat, P., Zollmann, A., Vogel, S., Black, A., Schultz, T., and Waibel, A. ``Optimizing Components for Handheld Two-way Speech Translation for an English-Iraqi Arabic System", *In the Proceedings of INTERSPEECH, Pittsburgh, USA, 2006*.

[2] Eck, M., Lane, I., Bach, N., Hewavitharana, S., Kolss, M., Zhao, B., Hildebrand, S., Vogel, S., and Waibel, S. ``The UKA/CMU Statistical Machine Translation System for IWSLT 2006", *In the Proceedings of IWSLT-2006, Kyoto, Japan, 2006.*

[3] Schultz, T. and Black, A. ``Challenges with Rapid Adaptation of Speech Translation Systems to New Language Pairs", *In the Proceedings of ICASSP, Toulouse, France, 2006*.

[4] Tomokiyo, L., Peterson, K., Black, A., and Lenzo, K. ``Intelligibility of Machine Translation Output in Speech Synthesis", *In the Proceedings of INTERSPEECH, Pittsburgh, USA, 2006*.

[5] Stallard, D., Choi, F., Kao, C., Krstovski, K., Natarajan, P., Prasad, R., Saleem, S., and Subramanian, S. ``The BBN 2007 Displayless English/Iraqi Speech-to-Speech Translation System", *In the Proceedings of INTERSPEECH, Antwerp, Belgium, 2007*.

[6] Gao, Y., Zhou, B., Sarikaya, R., Afify, M., Kuo, H., Zhu, W., Deng, Y., Prosser, C., Zhang, W., and Besacier, L. ``IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-Speech Translator", *In the Proceedings of the First International Workshop on Medical Speech Translation, ACL , New York, USA, 2006*.

[7] Georgiou, P., Sethy, P., Shin, J., and Narayanan, S. ``An English-Persian Automatic Speech Translator: Recent Developments in Domain Portability and User Modeling", *In the Proceedings of ISYC2006, Ayia Napa, Cyprus, July 2006.*

[8] Riesa, J., Mohit, B., Knight, K., Marcu, D. ``Building an English-Iraqi Arabic Machine Translation System for Spoken Utterances with Limited Resources", *In the Proceedings of INTERSPEECH, Pittsburgh, USA, 2006*.

[9] Vogel, S. ``PESA: Phrase Pair Extraction as Sentence Splitting", In the Proceedings of MT Summit X, Phuket, Thailand, September 2005.

[10] Zhang, Y., and Vogel, S. ``An Efficient Phrase-to-Phrase Alignment Model for Arbitrarily Long Phrase and Large Corpora", *In the Proceedings of the 10th EAMT conference "Practical applications of machine translation", 30-31 May 2005, Budapest; pp. 294-301*.

[11] Stolcke, A. ``SRILM - An extensible language modeling toolkit", *In the Proceedings of ICSLP 2002, vol. 2, (Denver, CO), pp. 901--904, September 2002.*

[12] Venugopal, A., and Vogel, S. ``Considerations in Maximum Mutual Information and Minimum Classification Error training for Statistical Machine Translation", *In the Proceedings of EAMT-2005, Budapest, Hungary May 30-31, 2005.*

[13] Schultz, T., Black, A., Badasker, S., Hornyak, M., and Kominek, J. ``SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems", *In the Proceedings of INTERSPEECH, Antwerp, Belgium, 2007*.

[14] Cepstral, LLC VoiceForge voice building system http://voiceforge.com

[15] Black, A., Lenzo, K. and Pagel, V. ``Issues in Building General Letter to Sound Rules", *In the Proceedings of The 3rd ESCA Workshop on Speech Synthesis, pp. 77-80, Jenolan Caves, Australia, 1998.*

[16] Soltau, H., Metze, F., Fugen, C., and Waibel, A. ``A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment", *In the Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2001.*

[17] Valtchev, V., Odell, J. J., Woodland, P. C., and Young, S. J. ``Lattice-based discriminative training for large vocabulary speech recognition", *In the Proceedings of ICASSP, 1996.*

[18] Zhang, Y., and Vogel, S. ``PanDoRA: A Large-scale Two-way Statistical Machine Translation System for Hand-held Devices", *In the Proceedings of MT Summit XI, Copenhagen, Denmark, Sep. 10-14 2007*.

[19] P. Koehn. 2004. ``Pharaoh: A beam search decoder for phrase-based statistical machine translation models", *In the Proceedings of AMTA 2004*.

[20] Papineni, K., Roukos, S., Ward, T., and Zhu, W. ``BLEU: A method for automatic evaluation of machine translation", *In the Proceedings of ACL-2002, ACL, Philadelphia, PA, July 2002.*