IraqComm: A Next Generation Translation System

Kristin Precoda, Jing Zheng, Dimitra Vergyri, Horacio Franco,
Colleen Richey, Andreas Kathol, Sachin Kajarekar

SRI International, Menlo Park, CA, USA
{precoda,zj,dverg,hef,colleen,kathol,sachin}@speech.sri.com

Abstract

This paper describes the IraqComm™ translation system developed by SRI International, with components from Language Weaver, Inc., and Cepstral, LLC. IraqComm, supported by the DARPA Translation Systems for Tactical Use (TRANSTAC) program, mediates and translates spontaneous conversations between an English speaker and a speaker of colloquial Iraqi Arabic. It is trained to handle topics of tactical importance, including force protection, checkpoint operations, civil affairs, basic medical interviews, and training. Major components of the system include SRI's DynaSpeak® speech recognizer, Gemini, SRInterp, and Language Weaver translation engines, Cepstral's Swift speech synthesis engine, and the user interface. The system runs on standard Windows computers and can be used in a variety of modes, including an eyes-free, nearly hands-free mode.
Index Terms: speech translation, spoken language translation system, user interface

1. Introduction

This paper describes the IraqComm™ translation system developed under the DARPA Translation Systems for Tactical Use (TRANSTAC) program. IraqComm mediates and translates spontaneous, real-time conversations between an English speaker and a speaker of colloquial Iraqi Arabic. It is trained to handle topics of tactical importance, including force protection, checkpoint operations, civil affairs, basic medical interviews, and training. The major components of the system are described below and include the speech recognizer, translation engines, speech synthesis engine, and user interface. An excerpt of a sample dialog using the system during a NIST evaluation in February 2007 is shown and discussed.

2. Software system architecture

IraqComm includes two recognizers, one for English and one for colloquial Iraqi Arabic, which are the primary means of providing language input to the system. When the user selects the input language and speaks, the recognizer begins calculating probable word sequences given the acoustics and the statistics of previously seen training data. When recognition is complete, several top recognition hypotheses may be displayed. The most likely recognized word sequence, or another chosen by the user, is passed to two translation engines. If the primary engine produces a translation within a specified time, that translation is passed to the speech synthesis engine for playback. If not, the output of the second engine is synthesized and played.

The user may select various options to control which steps are performed automatically and which require user intervention and confirmation before proceeding. The user may also choose to enter input text via a system keyboard; this is primarily intended for making minor corrections to automatic speech recognition (ASR) errors.

3. Hardware system architecture

The IraqComm software works with standard, widely commercially available hardware components. It runs on the Windows operating system. Most standard Windows platforms do not provide sufficiently loud audio output or built-in microphones suitable for use by two speakers who may potentially not be very close together. For this reason the system usually makes use of platform-external audio input and output devices. Either one or two microphones can be used; if two, the input audio may be fed to a single audio channel or to separate audio channels per language. External speakers of some kind are usually used to provide sufficient amplitude, or the audio output can be presented through one or both earpieces in one or two headsets.

4. Speech recognizer

IraqComm uses the Dynaspeak® real-time speech recognition engine [1] to transcribe Iraqi and English speech. Dynaspeak is designed to accept both highly optimized statistical grammars, used for most of the system's recognition tasks, and flexible hierarchical dynamic grammars, used to recognize command grammars. In IraqComm, the phonetic HMMs, the pronunciation dictionary, and the n-gram language model are compiled into a state-level determinized and minimized weighted finite-state acceptor, called a state graph, on top of which a time-synchronous Viterbi search is performed to generate lattices. Language model rescoring is applied in a second pass to the resulting lattices to derive the final recognition output. This two-pass decoding structure allows the use of a small, heavily pruned language model for fast decoding and a much larger language model for fast rescoring; the rescoring can yield more than a 20% relative word error rate reduction in English. Rescoring with the language model is implemented with the SRILM toolkit [2].

4.1. Recognition front end

IraqComm uses a recognition front end with a 16KHz sampling rate, 10 ms frame advance rate, and mel frequency cepstral coefficients (13 coefficients and their first and second derivatives). The Iraqi recognizer also uses third derivatives, and heteroscedastic linear discriminant analysis to reduce the feature dimensionality to 39.

A combination of two methods is used for noise robustness: (1) acoustic model training in noise and (2) feature compensation. During model training an augmented training set is used with additional noisy speech generated by adding a set of noise samples to the original training data. The noisy data has SNRs ranging between 15 dB and very clean

signals, to minimize degradation in recognition of clean speech. The additive noise is selected to match expected kinds of noise in the target environment. The relative error reduction achieved by this approach has ranged from 13% for a test set with SNRs in the range of 15 to 25 dB, to 27% for a test set with SNRs in the range of 5 to 15 dB.

The feature compensation approach uses the Probabilistic Optimum Filtering (POF) algorithm ([3], [4]), which applies a piecewise linear transformation to map a noisy feature space to a clean one. During recognition, different POF mappings are dynamically selected based on estimates of the SNR, and applied when the measured SNR is worse than 15 dB. For SNRs between 5 and 15 dB relative error reductions of 30% to 75% have been observed depending on the type of noise.

4.2. Acoustic models

Decision-tree state-clustered triphone models were trained for both languages, and discriminatively trained using Minimum Phone Frame Error (MPFE) training ([5],[6]). Gaussian shortlists [7] are used to speed up the Gaussian computation during recognition. The English models are trained on the Wall Street Journal corpus (80 hours), the ATIS corpus (30 hours), and the Broadcast News HUB4 release for 1996–1997 (100 hours), and tuned on in-domain data. Part of the data was corrupted by noise and added to the training data to improve the noise robustness of the models. The models have an inventory of 45 phones, and 600 triphone state clusters were trained, with 48 Gaussians per cluster.

The Iraqi Arabic models were trained with about 300 hours of data collected by the TRANSTAC program. Lacking a full pronunciation lexicon, the graphemic (orthographic) representation is used with some linguistic rules applied to derive the pronunciation lexicon used to train the acoustic models. The pronunciation lexicon is based on an inventory of 33 "graphones". 1200 triphone state clusters were trained with 64 Gaussian mixtures per cluster.

4.3. Language models

The English recognizer uses a vocabulary of 60K words, including a few common n-grams defined to be "multi-words". 21 domain-based word classes were also defined and used; these include, for example, body parts, weekdays, place names, and so on. The final language model (LM) was obtained by interpolating an LM trained only on about 5M words of in-domain data with a larger, general English LM. The final 4-gram LM uses 10M n-grams after pruning.

The Iraqi Arabic recognizer currently uses a vocabulary of 52K words. A 4-gram LM with about 1.2M n-grams was trained from the approximately 3M words in the TRANSTAC data collection. Techniques for automatic word clustering, word-class LMs and morphology-based LMs gave small improvements in performance.

4.4. ASR accuracy

The ASR components were evaluated on several different test sets. In various in-domain tasks the English recognizer showed a word error rate (WER) of between 6-15%, while the Iraqi Arabic recognizer has a WER between 18-30%. However, it should be noted that inconsistencies in the Iraqi orthography make it difficult to evaluate the accuracy consistently. On data from human-human conversations, the system's performance degrades significantly (about 25% WER for English, 40% for Iraqi Arabic).

5. Translation components

IraqComm uses two independently developed translation components in each translation direction, minimizing the risk of various kinds of failures and tending to increase accuracy.

In translating from English to Iraqi Arabic, IraqComm utilizes a rule-based engine developed by SRI and a statistically-based translation engine supplied by Language Weaver, Inc. The former is implemented in SRI's Gemini framework [8] and consists of two manually developed context-free unification-based grammars, one for parsing English and one for generating Iraqi Arabic. Connecting the two is a common interlingua based on (quasi-)logical forms (see [9] for a similar approach for Pashto).

Language Weaver's engine uses a statistical phrase-based translation approach. It uses a dynamic programming algorithm to find the best translation among a large number of possibilities, guided by an automatically built probabilistic bilingual phrase table and an n-gram language model for the output language. Since there is little parallel English/Iraqi Arabic data, this engine uses out-of-domain Modern Standard Arabic (MSA)/English bitexts to increase coverage.

In the Iraqi-to-English direction, IraqComm uses two statistical engines, from Language Weaver and SRI ("SRInterp"). Both are phrase-based statistical systems, though with independent design and implementation. Since Language Weaver's system is augmented with MSA/English training data, it has better coverage and is better able to handle code switching between Iraqi and MSA. Conversely, SRI's engine has been trained exclusively on in-domain Iraqi/English corpora and appears to have higher precision on purely in-domain input.

This redundancy offers enhanced accuracy, robust translation quality and software reliability. While brittle in cases of unknown words and ill-formed input, including some kinds of recognition errors, the rule-based translator often produces output that is of higher quality and more easily understood by native speakers than that of the statistical translation engine. At run time, the two operate in parallel and if the rule-based translator returns a result within a certain amount of time, its output is sent to the speech synthesizer. Otherwise, the output of the statistical translator is displayed and synthesized. In the Iraqi-to-English direction, Language Weaver's output is generally the primary translation, with the SRInterp output also provided visually. If the user notices the two translations differ significantly in meaning, s/he may wish to confirm the information conveyed via the system. In the future, more sophisticated methods of system combination relying on translation confidence estimation may be developed and incorporated.

6. Speech synthesis

The speech synthesis for both languages is provided by Cepstral LLC's Swift engine [10], which performs concatenative synthesis. Both voices are male. The English voice is a fairly standard one, customized for this application only in the pronunciations of certain domain-specific lexical items. The Iraqi Arabic voice was developed expressly for the DARPA TRANSTAC program and is based on a speaker with a Baghdadi accent. Intelligibility is generally fairly good in Iraqi Arabic; in cases where intelligibility is less than desired, it is difficult to distinguish the contribution of potentially poorly formed machine translation output utterances from that of imperfect synthesis.

7. Interface and configurability

A screenshot of the graphical user interface for IraqComm is shown in Figure 1. The string of words output by the recognizer is shown in the "Text to Translate" box in the upper left of the screen; their (primary) translation is displayed in the "Translation" box below. The recognizer displays the n-best hypotheses in the upper right of the screen, and the outputs of the two translation engines are shown in the lower right. The lower left of the screen contains a list of frequent phrases, which may be played by simply clicking on the desired phrase.

Figure 1: *IraqComm screenshot.*

7.1. Software configurations

IraqComm offers various configurations so the application can be customized to the needs of any specific situation. Clearly, different situations have different requirements, in terms of the physical constraints of the situation, the communicative goals of the users, the cost associated with errors, and so on. For example, the Multi-National Security Transition Command-Iraq (MNSTC-I) considered a relatively large laptop screen desirable. That screen facilitates both the English and Iraqi speaker being able to see written feedback on the screen, become equally familiar with the system, and have joint planning discussions between approximately equal partners. Other situations may require eyes-free usage, and hence screen size or even the presence of a screen is unimportant during an actual interaction.

If the system is used in eyes-free mode, getting feedback to the user(s) via another modality becomes important. IraqComm makes audio cues available to indicate when either user should speak, when the system is busy, and for confirmation of recognition results before translation. The user may choose whether to hear the recognition output before it is translated. If the recognition output is played, the user may further choose to configure the application so that a recognition result is automatically translated, or so that it requires an explicit confirmation from the user before translation is performed. The latter case allows substantial user control when the cost of an error is higher than the cost of the additional time needed. For greatest control, backtranslations of an Iraqi Arabic translation into English are also available, for comparison with the original English input.

IraqComm may also be used in nearly hands-free mode and may be operated for example by putting the system in a backpack and only using one finger to control two buttons on a small external device. The two buttons control recording of input speech and allow a translation in process to be aborted. In this mode the interface is augmented by a few spoken commands, including, for instance, commands to play pre-recorded usage instructions to an Iraqi and to replay the last item played by the system.

7.2. Audio hardware configurations

The choice of an appropriate microphone configuration is also dependent on the task at hand. There are several considerations which may be important. Among these may be the desire to maintain eye contact and observe body language, the distance between the English and Iraqi Arabic speakers, possible interference, security issues, and risk of loss of wireless microphones, the feasibility of the two speakers being connected by wires for wired microphones, and possible mistrust by an Iraqi of using any visible microphone.

Audio output presents a similar array of points to consider. Most currently available laptops and handheld devices do not provide sufficiently loud output for any but fairly quiet environments. External speakers are readily available and can be powered either through a USB connection, batteries, or AC power. AC-powered speakers can provide considerable amplitude but require a nearby power output. Battery-powered speakers can also be very loud but have the disadvantage of limited battery life and extra weight. Finally, USB-powered speakers are typically not as loud as other types but do have the advantage of not requiring either a power outlet or batteries.

In some circumstances, the users may not want translations played out of speakers, and instead may prefer to have the output played over headphones to one or both speakers. IraqComm also supports this usage.

8. Dialog sample

While IraqComm translations convey the speaker's intended meaning the majority of the time, recognition and translation errors do occur. Table 1 presents an excerpt of a dialog produced during a recent NIST-run evaluation of IraqComm. The evaluation was conducted with the system configured for hands-free, eyes-free usage. The users were thus not able to see the recognition or translation or to use the lists of alternate recognition hypotheses and translations.

This excerpt was selected to show utterances with no errors, utterances with errors that do not affect intelligibility, and utterances with errors that either render the utterance unintelligible or cause a miscommunication. Utterance (a) shows a typical English recognition error, of a word that is unstressed and exhibits reduced articulation. Over a third of the English recognition errors – the most common error type – during the unscripted evaluation were misrecognitions of unstressed, reduced words. Despite these errors, the translation into Iraqi Arabic still conveys the English speaker's meaning. Utterance (b) is not as successful, and, in fact, could cause miscommunication. The problem arises from misrecognition of the object pronominal suffix on the Iraqi Arabic verb *detain* and mistranslation of its subject. كان المفروض ([it] was supposed to) is an expression for generic obligation; the following verb indicates the subject. The machine translation component uses the 3rd person singular form of "was supposed to" instead of the 1st person plural form of "we detain" and translates the utterance using the English pronoun "he." The misrecognition of نعتقله (detain-

Table 1: *Excerpt from dialog. Errors are highlighted in bold; literal translations of Iraqi Arabic are added in parentheses.*

	Utterance	Recognition	Translation
(a)	you should have detained him for questioning	**he** should **be** detained him for questioning	لازم يكون له موقوف للإستجواب (should be to him detained for questioning)
(b)	كان المفروض نعتقله (we were supposed to detain him)	كان المفروض **نعتقلك** (we were supposed detain **you**)	**he** was supposed to detain you
(c)	no you should have held him	no you should **of help** him	لا إنت لازم من يساعدوه (no you should of help him)
(d)	ما دأفهم (I don't understand)	ما دأفتهم (I don't understand)	I don't understand
(e)	why did you stop the teenagers	why did you stop the teenagers from	ليش إنت **وقف** المراهقين من (why you stopped the teenagers from)
(f)	علمود يفتشون جوة الكراسي (so that we search under the seats)	علمود يفتشون جوة الكراسي (so that we search under the seats)	in order to search under the seats

him) as نعتقلك (detain-you) also causes misunderstanding. Iraqi Arabic is highly inflected, with many prefixes and suffixes indicating such things as agreement, object and possessive pronouns, and prepositions. A small error in recognition, such as one grapheme in (b), can have significant effects on the translation. About two fifths of the Iraqi Arabic recognition errors during the unscripted evaluation were misrecognitions of affixes. Utterance (e) has a translation error: the 3rd person singular form of the verb is used instead of the 2nd person singular, resulting in the ungrammatical "you stopped(3sg)." This is a common translation error from English into Iraqi Arabic that usually leads to an ungrammatical but understandable translation. Utterances (d) and (f) have no errors.

In this evaluation, in the judgement of a translator and linguist working together, about 75% of the utterances successfully conveyed the speaker's meaning, despite minor recognition and translation errors in some cases. About one quarter, like (b) and (c), communicated the incorrect meaning or were not intelligible. The vast majority of these cases included recognition errors. As the development of IraqComm progresses, additional data and improvements in the ASR and MT components should lead to fewer errors, better translations and fewer miscommunications.

9. Conclusions

IraqComm is a good example of how far spoken language translation technology has come in recent years. The system can be used in a variety of situations and offers a reliable alternative when human interpreters are scarce or unavailable. While the system is not perfect, cooperative users can successfully communicate and exchange information, occasionally paraphrasing or repeating an utterance. The system's highly configurable user interface can be adapted to many different use cases and offers a choice of operating points balancing speed and thorough confirmation.

10. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and the Department of Interior-National Business Center (DOI-NBC) under contract number NBCHD040058.

We would also like to acknowledge the important contributions of other SRI IraqComm team members, including Murat Akbacak, Michael Frandsen, Venkata Ramana Rao Gadde, Martin Graciarena, Huda Jameel, Donald Kintzing, Shane Mason, Susanne Riehemann, Talia Shaham, and Wen Wang.

11. References

[1] Franco, H., Zheng, J., Butzberger, J., Cesari, F., Frandsen, M., Arnold, J., Gadde, V.R.R., Stolcke, A., and Abrash, V., "DynaSpeak: SRI's scalable speech recognizer for embedded and mobile systems," in *Proceedings Human Language Technology Conference*, (San Diego, CA), March 2002.

[2] Stolcke, A., "SRILM — An extensible language modeling toolkit," in *Proc. International Conference on Spoken Language Processing*, vol. 2, (Denver, CO), pp. 901-904, September 2002.

[3] Neumeyer, L., and Weintraub, M., "Probabilistic optimum filtering for robust speech recognition," in *Proc. ICASSP*, 1994.

[4] Graciarena, M., Franco, H., Myers, G., and Abrash, V., "Robust feature compensation in nonstationary and multiple noise environments," in *Proc. of EUROSPEECH*, 2005.

[5] Zheng, J. and Stolcke, A., "Improved discriminative training using phone lattices," in *Proc. Eurospeech*, (Lisbon), pp. 2125-2128, September 2005.

[6] Povey, D. and Woodland, P.C., "Minimum Phone Error and I-Smoothing for Improved Discrimative Training," *Proc. ICASSP,* 2001.

[7] Digalakis, V., Monaco, P., and Murveit, H., "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers," *IEEE Transactions on Speech and Audio Processing*, 4(4): 281-289, 1996.

[8] Dowding, J., Gawron, J. M., Appelt, D. E., Bear, J., Cherny, L., Moore, R., and Moran, D. B. "Gemini: A natural language system for spoken language understanding," *Proc. ACL 1993*, pp. 54–61, 1993.

[9] Kathol, A., Precoda, K., Vergyri, D., Wang, W. and Riehemann, S., "Speech translation for low-resource languages: The case of Pashto," in *Proc. Eurospeech*, Lisbon, pp. 2273-2276, September 2005.

[10] Cepstral, LLC, "Swift™: Small footprint text-to-speech synthesizer", http://www.cepstral.com, Pittsburgh, PA, 2005.