Proc. of ASRU 2003, Dec 2003, U.S Virgin Islands

TRANSONICS: A SPEECH TO SPEECH SYSTEM FOR ENGLISH-PERSIAN INTERACTIONS

S. Narayanan,[*] S. Ananthakrishnan,[*] R. Belvin,[†] E. Ettaile,[*] S. Ganjavi,[*]
P. G. Georgiou,[*] C. M. Hein,[†] S. Kadambe,[†] K. Knight,[*] D. Marcu,[*]
H. E. Neely,[†] N. Srinivasamurthy,[*] D. Traum,[*] D. Wang[*]

[*]University of Southern California, Los Angeles
[†]HRL Laboratories, Malibu, California
shri@sipi.usc.edu

ABSTRACT

In this paper we describe the first phase of development of our speech-to-speech system between English and Modern Persian under the DARPA Babylon program. We give an overview of the various system components: the front end ASR, the machine translation system and the speech generation system. Challenges such as the sparseness of available spoken language data and solutions that have been employed to maximize the obtained benefits from using these limited resources are examined. Efforts in the creation of the user interface and the underlying dialog management system for mediated communication are described.

1. INTRODUCTION

Creating a *speech-to-speech* (S2S) translation system presents significant challenges. These arise not only due to the complex nature of the individual technologies involved in a S2S translation, but also due to the intricate interaction that these technologies have to achieve. Additionally, a great challenge for the specific S2S translation system we are presenting stems from the great discrepancy in the structure of the English and Persian (Modern Persian is called Farsi by native speakers) languages, as well as the extremely limited amount of data for the Persian language. Furthermore, for the most part the Persian writing system (employing the Arabic script) lacks the explicit inclusion of vowel sounds, thus resulting in a very large amount of one-to-many mappings from transcription to acoustic and semantic representations.

There have been major efforts in creating end-to-end S2S translation systems including the Spoken Language Translator [1], the VerbMobil project [2], NESPOLE! [3], and several ongoing undertakings by the participants of the DARPA Babylon project [4]. It should be noted that some of the language pairs addressed by the Babylon project, including Persian, have not been well investigated in the speech technology community. One major challenge this creates is lack of readily available spoken language resources. Some of the steps taken by this team to tackle this challenge are be reported in this paper.

2. SYSTEM OVERVIEW

Our system comprises several spoken language components that act in a collaborative manner, and an optional visual and control graphical user interface (GUI). A functional block diagram is shown in Fig. 1. In our architecture, all messages are broadcast with a tag that includes among other information the message's originating and proposed terminating point, and are visible to all subsystems. This allows for ease of collaboration and monitoring of the internal communication channels by the dialog manager, which can interrupt and request corrective action by the user. The most probable corrective actions are requests for repeat, rephrase, confirmation, and disambiguation where the user is asked to choose the utterance from a list of options (using the speech and/or the GUI).

The individual subsystems are the *Automatic Speech Recognition* (ASR) subsystem, which works both using *Fixed State Grammars* (FSG) and *Language Models* (LM) and produces n-best lists/lattices along with the decoding confidence scores. The output of the ASR is subsequently "re-scored" by the *Dialog Manager* (DM) according to the history of the conversation, before being passed along to the *Machine Translation* (MT) unit. The MT unit also works in two modes: Classifier based MT and a fully Stochastic MT. Finally, a unit selection based *Text To Speech* synthesizer (TTS) provides the spoken output.

3. DATA COLLECTION & TRANSCRIPTION

The development of S2S systems in the Babylon project rely on a "limited" domain approach. For example, one of the major domains considered is interaction between an English-speaking medical professional and a Persian language speaking patient. The available data in a mediated doctor-patient interaction even in English are indeed sparse. This highlights data needs, not

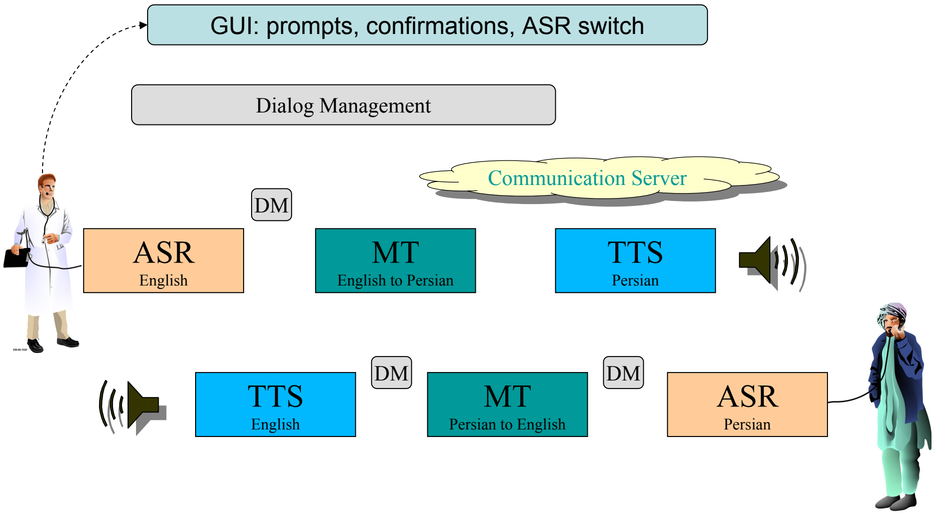

Fig. 1. Block diagram of system. Note that the communication server allows interaction between all subsystems, and the broadcast of messages. Our vision is that only the doctor will have access to the GUI, and the patient will only be given a phone handset.

only in terms of speech data for Persian, but also multilingual spoken language data providing adequate linguistic coverage in the target domains. As mentioned earlier, readily available resources for our target language and domain are severely limited. Hence, a significant portion of the initial efforts are focused on data gathering. Our approach to this problem was multipronged: leverage and adapt existing resources, and develop new resources.

Table 1 provides a summary of the currently identified data sources for the Transonics system development. It includes domain material gathered from existing resources (and translated) or material being specifically collected as a part of the project. The first 4 rows are target-domain material completely transcribed/translated in both English and Persian. For the medical domain, the initial bootstrapping was based on the availability of a large amount of common medical expressions, obtained from Marine Acoustics Inc (http://www.sarich.com) and a number of medical phrase books. These were not only useful for the creation of fixed state grammars and the Classifier based MT, but are also valuable in enriching our medical domain vocabulary (especially the medical phrase book data). In addition, for supporting development of language models for generic larger vocabulary recognition in Persian, we have been gathering Persian text corpora from mining publicly available newspapers. Due to the tremendous amount of data needs and processing involved, we are continuing the collection and transcription process.

The other major focus of our data needs is spoken language interaction data in the target domains. Although we identified some limited sources of existing spoken dialog interaction data (in English for *e.g.,* doctor-patient dialogs), these are significantly different from the mediated dialogs of the Transonics S2S system. Hence, a significant portion of current efforts is focused on generating actual interaction data (both monolingual and bilingual modes) in the target domain. In addition to feeding the ASR and MT modules, these data are valuable for designing the dialog interface. These data collection processes are well under way and the results will be progressively incorporated into our S2S system.

3.1. USC Standardized Patient Data Collection

The most important data collection effort that has been undertaken by USC/HRL, and in collaboration with the USC Keck School of Medicine, is the *Standardized Patient* data collection. The practice of using *Standardized Patients* began in Los Angeles in the 1960s as a way of allowing medical students to gain experience interacting with and diagnosing patients, and with a greater degree of consistency in terms of symptoms displayed; moreover, the patients are trained to rate the students on their bedside manner, handling of the physical examination, and methods of diagnosis.

Standardized Patient cases are created by MD's and RNs, ideally ones who have had first-hand experience with such medical instances. The cases consist of a detailed description of the symptoms the standardized patient is to report (some brief samples are shown below), as well as a one-page synopsis of some of the patient's vital signs, which will differ from their actual vital signs, but which will serve as important indicators to the students in forming a proper diagnosis (see Fig. 2). The SP goes

(A) Brief patient instructions:
The cough started about 3 months ago. It is constant and produces sputum that is usually thick and yellow and occasionally has some flecks of blood in it. The sputum does not have any bad smell. The cough is deep and you have occasional coughing "fits." The cough is fairly constant, happening often during both the day and the night. You have also lost weight during this time without dieting. You have noticed your skirt/pants have become very loose.

(B) Doctors chart of vital signs:

Temperature:	99 degrees F
Pulse:	100
Respiration:	18
Blood Pressure:	112/80 mm Hg

Fig. 2. (A) Only seen by patient. (B) Seen by both medical student and patient

through extensive training and two practice run examinations by qualified MD's.

The cases, which include among others, tuberculosis (TB), malaria, flu, heart attack, severe diarrhea, *etc.*, were chosen not only to get better balance of illnesses to injuries – the vast majority of the IBM data collection are injuries – but also after research into published material by military institutions (such as the Naval Medical Research Institute, Army sources, Army Research Inst. of Environmental Medicine, *etc.*).

3.2. Transcription Methods

Data transcription in Persian as mentioned previously is a harder task than in most languages. To better understand this point, consider the example of Fig. 3. The first choice one has to make in the transcription process is the use of a character set to represent the Arabic script that allows easy access to non-Persian speakers, avoids the multiple character forms that exist in Persian according to the relative position of the character in the word, and reduces the transcription overhead. This mapping leads to the USCPers transcription scheme, which employs the ASCII code and is shown on the third column of Fig. 3. Additionally, we need to provide pronunciations of each word (for ASR & TTS), and we have created the USCPron transcription system to address this issue. Furthermore, Persian written system does not include the vowel sounds in its written form, and thus multiple transcriptions on USCPers or the original Arabic script can result in both different pronunciations and different meanings. This prompts the need for a different transcription scheme that enables a one-to-one mapping between the acoustic representation (excluding user variability) and the transcription method, which we introduce as USCPers+. An in-depth analysis of our transcription schemes is given in [5].

4. DM

The dialog manager of the Transonics system performs the following functions:

- keep track of the discourse history
- provide hypotheses of most likely next utterances
- manage the grounding interaction between speakers

Gloss	Arabic Script	USCPers	USCPers+	USCPron
One Hundred	صد	$d	$ad	sad
Dam	سد	sd	sad	sad
Six	شش	SS	SeS	SeS
Lung	شش	SS	SoS	SoS

Fig. 3. Examples of the limitations of both the Arabic script and USCPers (a direct romanized version of Arabic script), which share a one-to-one mapping, in representing transcribed speech. Due to the lack of vowels in the orthographic transcription, it is extremely common that two completely different acoustic and semantic representations lead to the same orthographic transcription. A newly proposed lexicalization convention called USCPers+ includes vocalic information thereby retaining the unique mapping between the acoustic input and lexical output of ASR. USCPron provides the pronunciations for the lexical items represented in USCPers+.

Our current representation of the dialog history includes a linear sequence of utterances, each tagged with results of speech recognition, concept recognition and translation. This is combined with sets of dialog plans for sub domains yielding a frame-like structure of concepts that have been discussed during the course of a dialog and concepts that are expected to be discussed. The dialog plans are partly constructed by interviews with subject matter experts, and partly drawn from corpus examples of dialogs. Using this information, the dialog manager can make predictions on likely next utterances.

An issue is how this kind of information, to the extent that it is useful, should be integrated within the spoken translation pipeline. One option is to integrate within the speech recognizer, e.g., by changing the grammar (or probability distribution within a grammar) based on the dialog state. Another option is to fit within a concept-recognizer/translation module, giving probability distributions of possible concepts. Yet a third possibility is to have the dialog function as a separate module, adapting hypotheses provided by a recognizer to serve as input to translation. For the moment, for ease of prototyping and runtime efficiency, we have adopted the last approach. The dialog manager will prefer contextually appropriate responses (such as answers to a pending question) and dis-prefer already expressed or contradictory expressions (with explicit indications of contradiction).

The final task of the dialog manager is to track the grounding of the participants, that is how they reach sufficient confidence that they have been understood correctly. There are several "grounding modes" that the user can choose:

1. flow through (no feedback)
2. confidence-based clarification
3. show interpretation (allow blocking)
4. always clarify

Data Description	Size and original form	Used in
English questions and answers (Marine Acoustics)	600 sentences in English and Persian - text	FSG, CBMT
Paraphrasing above	2000 sentences - text and audio	CBMT
WoZ experiments	100+ utterances in audio format	All modules
Medical phrasebooks	600+ Q&A utterances. Translated and transcribed in all necessary forms	All modules
Persian newspaper mining	Virtually unlimited text. Continuously converted to USCPers+	LM
DLI data - spontaneous speech	About 5h of mediated doctor-patient interaction – audio	All modules
CECOM data - scripted	500 English with Persian translations - No transcription	
CECOM data - semi-spontaneous	75 Q&A pairs in both English and Persian - No transcription	
IBM data - semi-spontaneous	Force protection data. About 350 short interactions.	LM, DM
USC Medical school	Videotaped medical interactions – standardized patient examinations	LM, DM
USC/HRL medical data collection	Doctor patient interaction, using trained patients and medical students. 200 Dialogs in audio format.	LM, DM

Table 1. List of data sources and uses. (FSG -finite state grammar, LM - Language Model, CBMT - classifier based machine translation, DM - dialog manager).

In flow-through mode, whatever the best interpretation and translation produced is, it will be spoken by the system to the other language speaker. On the other extreme, the system will produce (either in speech or through a visual display) its interpretation, and require the speaker to confirm (either verbally or by pressing a "go ahead" button) before producing the translation for the other speaker. There are also two intermediate modes: one in which the interpretation is presented, but user "go ahead" is not required (but the user may "stop" the translation, if wrong), and the second mode in which the system will decide whether to translate or require more feedback, based on confidence scores. The decision will be trained based on collected data, but thresholds will also be user-adjustable.

At this point, only a rudimentary dialog manager has been implemented, using the flow-through choice.

5. ASR

To recognize speech utterances we employ separate English and Persian ASRs built using HTK 3.1. The speech features used are 12 MFFCs and the zeroth cepstral coefficient and their Δ and $\Delta\Delta$ derivatives, employing Hamming windows of 25 ms with a feature vector calculated every 10 ms. The baseline English models are created from the train subset of TIMIT database after downsampling it to 8 kHz. We use triphone models in the English ASR. All English triphone models have 3 states with 16 GMMs per state.

Unlike the English language, there is a lack of adequate speech data in Persian. To overcome this drawback we adopted a data driven language-adaptive technique. We borrowed acoustic data from English to compensate for the lack of data in Persian. The key requirement in enabling the use of English data for Persian ASR is the development of a phoneme mapping between the two. We used a novel Earth Movers Distance based sub-phonetic/phonetic mapping [6]. Due to lack of data we were restricted to using monophone models for the Persian ASR. Future research will concentrate on building triphone Persian ASR using data from Persian and English. Additionally, for adaptation/re-training, the Marine Acoustics data was translated into Persian and read by 18 native Persian speakers (9 females and 7 males). We compared Persian adapted/re-trained ASRs using seed models from (i) sparse Persian speech data (FARSDAT), (ii) knowledge based English phonemes, (iii) data driven phonetic models and (iv) data driven sub-phonetic models as shown on Table 2. The results we obtained are very encouraging, illustrating that it is possible to make use of acoustic data even between diverse languages like English and Persian to improve the performance of ASRs in languages constrained by sparse data.

We also observe that our proposed technique while having better performance when the re-training is used does not perform as well when only adaptation is used. A possible reason for this is that the adaptation scheme used, MLLR, is restricted to only linear transformations, which may not be sufficient to model differences in phonemes between different languages, where phoneme contexts play an important role.

Seed Models	Phoneme Error Rate	
	Re-training	Adaptation
FARSDAT	20.35%	38.95%
Knowledge based	20.00%	39.87%
Phonetic mapping	20.13%	57.03%
Sub-phonetic mapping	19.80%	51.48%

Table 2. Phoneme error rates obtained for different approaches. Observe that sub-phonetic mapping ASR achieved the best recognition performance when re-training was used.

Due to the lack of available language (text) data we have yet to deploy a full LM based ASR for Persian, although we are in the process of mining data from Persian news sources for this purpose. Table 3 shows an example of the data processed where the generation of USCPers is automated to a large degree. The data are subsequencly processed by our team of transliterators to create the USCPers+ script of the same text, while at the same time minor modifications may be made to reflect predefined classes.

The LM generated from the above data as well as our existing English language LM will be interpolated with the ones we expect to create from the limited amount of medical data available, such as the standardized patient examination data and the USC/HRL collection effort. The FSG based recognizers are using all the available data as described earlier and listed in Table 1.

6. MT

The approach we employ for the Machine Translation unit is twofold. A classifier is applied as the main translator unit of the system because of its faster and more accurate performance, while a *statistical machine translator* (SMT) is kept as the backup unit for the cases when the classifier response is not within an acceptable confidence margin. These cases should be relatively infrequent if significantly large number of classes are chosen for the classifier based MT.

As a first step in building a classifier, the proper set of standard questions and answers that covered the context was selected. Every standard question or answer was chosen as a representation of a class. Following each input utterance, the system is expected to classify it in one of the predefined classes and generate the pre-stored translation. This requires training data to create the classes that are represented by each standard question or answer. To collect this data we created an online tool where a sentence was presented and users were asked to paraphrase it, thus expanding the coverage area of our training corpus. In addition, and in order to model expected errors introduced from the ASR module, we collected acoustic paraphrasing data that were not cleaned (*i.e.,* the recognized transcript may not match the uttered speech).

The resulting dataset was used to train a naive Bayesian classifier with uniform prior probabilities. The test set was gathered from an ASR and consists of both standard questions and paraphrased data. Since the test phase paraphrasing is separate and additional to the training set paraphrasing, there is significant test phrases that are new for the system.

From the collected data we have established a monotonically increasing performance in the MT classification as the paraphrasing increases. With the available paraphrasing the performance is roughly linearly increasing with over 1-2% per paraphrase round. We are continuously collecting more data to improve our classification quality, and additionally we are introducing more original phrases to increase our domain coverage.

A more sophisticated classification scheme that consists of a lattice of finite state transducers has been under development. A group of FST's model each main block in the real system. Thus, the ASR is modeled by a phoneme corrupter FST followed by a phoneme-to-word transducer. A set of unigram based FST's associated with each class followed by a bigram filter and a word-to-phoneme FST forms the speaker model. For every utterance from the (real) ASR a detection procedure, *e.g.,* Viterbi algorithm is performed to get the corresponding class. Early experiments with this system show 3.4% increase over the accuracy of the naive Bayesian classifier. Better ways of using the training data to build these FST's are under investigation.

The second method, to be used in the case of poor classification confidence, is the Stochastic MT method. SMT is based on word-to-word translation and can generate the translation for every input sentence. However, accuracy of the SMT, although a function of the training corpus, is in general expected to be lower than the accuracy of the classifier. The best performance for SMT can be achieved by employing a large amount of bilingual parallel text for training. This training corpus can be used to build a language model for the target language along with a statistical translation table, which relates words in the source language and their counterparts in the target language. However, due to the lack of any significant amount of English-Persian parallel text, we are following the approach of using the initial and target language models, and combining these with a dictionary approach for transition between the two languages. The LMs used are the same as has already been discussed in the ASR section.

Finally, another consideration in favor of the classifier system is the high computational demands of the SMT algorithms. In a S2S system where latency is crucial, the SMT system would always be kept as a backup choice after the utilization of the faster classifier system.

7. TTS

We rely on a hybrid unit selection based speech synthesis. In the default case, when the output is chosen from a classifier-

Persian	مدیر عامل سازمان پارکها گفت تاکنون در تهران ۱۸ هزار هکتار (معادل ۱۸۰ میلیون متر مربع) به فضای سبز شهر تهران در ۶ سال گذشته افزوده شده واین شهر در حال حاضر ۶۳۰ بوستان، گلستان و باغ شهری بزرگ و کوچك دارد
USCPers	mdyr?Aml sAzmAn pArkhA gft tAknvn dr thrAn hyJdh hzAr hktAr m?Adl $d v hStAd mylyvn mtr mrb? bh f2Ay sbz Shr thrAn dr SS sAl g#Sth Afzvdh Sdh v Ayn Shr dr HAl HA2r SS$d v sy bvstAn glstAn v bAQ Shry bzrg v kvCk dArd
USCPers+	modyr?Amel sAzmAn pArkhA goft tAkonvn dar tehrAn hyJdah hezAr hektAr mo?Adel $ad va haStAd myliyvn metr moraba? beh fa2Ay sabz Sahr tehrAn dar SeS sAl go#aSteh Afzvdeh Sodeh va Ayn Sahr dar HAl HA2er SeS$ad va sy bvstAn golestAn va bAQ Sahry bozorg va kvCak dArad
USCPron	modir?Amele sAzmAne pArkhA goft tAkonun dar tehrAn heJdah hezAr hektAr mo?Adele sad o haStAd milyun metre moraba? be fazAye sabze Sahre tehrAn dar SeS sAle gozaSte afzude Sode va in Sahr dar hAle hAzer SeSsad o si bustAne golestAn va bAqe Sahrie bozorg va kuCak dArad
Gloss	The managing director of Parks and Recreation Services said that so far since last year, 18,000 acre, which is equivalent to one hundred eighty million square foot, of green area has been added to Tehran in the past six years and this city currently has 630 parks, orchards and small and large garden.

Table 3. An example from the online newspaper, Hamshahri (September 16, 1996, Fourth Year, Number 1068), where the Persian (Arabic Script) transcription is converted into the forms of USCPers, USCPers+ and USCPron for the purposes of creating a Language Model, dictionary and lexicon. The Gloss is also provided here although not generated in our collection.

based MT, the generated phrases are known a priori. Hence, our first system release enabled us to use a prompt based system for spoken output. The other end of the unit selection possibility is through diphone concatenation. We have implemented such a synthesizer based on Festival [7] for English and Persian. Note that there are 29 sounds in the Persian language (6 vowels & 23 consonants), which results in the theoretical number of 900 diphones, fewer than needed for English that has a larger vowel inventory.

8. CONCLUSIONS

In this paper we have presented an overview of our S2S translation system and described the issues we are currently addressing especially related to data needs. We have described the interaction as well as work currently undertaken in each of the main sub-systems. Finally, we provide an outline of the path we are following in creating the English-Persian translation system. As we are currently focusing on the data collection aspects of the project, we aim to provide evaluation results at a later stage.

9. ACKNOWLEDGMENTS

This work was supported by the DARPA Babylon program, contract N66001-02-C-6023.

10. REFERENCES

[1] Manny Rayner, David Carter, Pierrette Bouillon, Vassilis Digalakis, and Mats Wirén, Eds., *The Spoken Language Translator*, Cambridge University Press; 1st edition, January 2000.

[2] W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, 2000.

[3] A. Lavie et al., “A multi-perspective evaluation of the NESPOLE! speech-to-speech translation system,” in *Proc. of ACL 2002 workshop on Speech-to-speech Translation: Algorithms and Systems*, Philadelphia, PA, July 2002.

[4] “The DARPA Babylon program,” `http://darpa-babylon.mitre.org`.

[5] Shadi Ganjavi, Panayiotis G. Georgiou, and Shrikanth Narayanan, “ASCII based transcription schemes for languages with the Arabic script: The case of Persian,” in *ASRU*, 2003.

[6] Naveen Srinivasamurthy and Shrikanth Narayanan, “Language-adaptive Persian speech recognition,” in *Eurospeech*, 2003.

[7] “The Festival Speech Synthesis System,” `http://www.cstr.ed.ac.uk/projects/festival/`.