Proc. ICASSP 2006
Toulouse, FR

SPEECH RECOGNITION ENGINEERING ISSUES IN SPEECH TO SPEECH TRANSLATION SYSTEM DESIGN FOR LOW RESOURCE LANGUAGES AND DOMAINS

Shrikanth Narayanan, Panayiotis G. Georgiou, Abhinav Sethy, Dagen Wang, Murtaza Bulut, Shiva Sundaram, Emil Ettelaie, Sankaranarayanan Ananthakrishnan

USC Viterbi School of Engineering

Horacio Franco, Kristin Precoda, Dimitra Vergyri, Jing Zheng, Wen Wang, Ramana Rao Gadde, Martin Graciarena, Victor Abrash, Michael Frandsen, Colleen Richey

SRI International

ABSTRACT

Engineering automatic speech recognition (ASR) for speech to speech (S2S) translation systems, especially targeting languages and domains that do not have readily available spoken language resources, is immensely challenging due to a number of reasons. In addition to contending with the conventional data-hungry speech acoustic and language modeling needs, these designs have to accommodate varying requirements imposed by the domain needs and characteristics, target device and usage modality (such as phrase-based, or spontaneous free form interactions, with or without visual feedback) and huge spoken language variability arising due to socio-linguistic and cultural differences of the users. This paper, using case studies of creating speech translation systems between English and languages such as Pashto and Farsi, describes some of the practical issues and the solutions that were developed for multilingual ASR development. These include novel acoustic and language modeling strategies such as language adaptive recognition, active-learning based language modeling, class-based language models that can better exploit resource poor language data, efficient search strategies, including N-best and confidence generation to aid multiple hypotheses translation, use of dialog information and clever interface choices to facilitate ASR, and audio interface design for meeting both usability and robustness requirements.

1. INTRODUCTION

The broader impact of developing communication augmentation systems for specific decision-making environments can be seen in their potential for facilitating multicultural efforts ranging from disaster relief, to peace and security operations to servicing diverse immigrant populations. One such application environment for speech-to-speech systems, also the domain of focus in this paper, is the medical domain. In the United States, the lack of equal medical treatment for patients with limited English speaking capability is a considerable problem. The development of cross-linguistic collaborative decision augmentation systems need to go much beyond providing simple translation of text: they need to robustly recognize speech in multiple languages, in context, and provide appropriate concept translation to move the interaction forward. The speech-to-speech systems reported in this paper hence adopt a holistic view, that goes beyond plain utterance level translation, to facilitate decision making by minimizing the cognitive mismatch among the conversational participants.

Materializing such a goal, especially for low resource languages, poses a number of design challenges. These range from dealing with data sparseness to user interface design. In this work, we describe some of the engineering issues we faced as we started towards this very ambitious role in two low resource languages - namely Persian (Farsi, mainly spoken in Iran and areas of Afghanistan) and Pashto (the main language in Afghanistan).

The work in the paper was supported by DARPA.

2. STRATEGIES FOR ADDRESSING RESOURCE CHALLENGES AND DATA SPARSENESS

Development of S2S systems requires spoken language data of various forms and amounts: several hours of acoustic speech data, adequately reflecting dialectal variants, several hundred thousand running words of domain data for language modeling as well as millions of words of parallel text in the target language pairs for machine translation.

2.1. Existing and collected data

The availability of speech/language data in any form – dictionaries, transcripts or acoustic data – was the first hurdle faced in both Pashto and Persian. On the Persian side, the only available data was the Farsdat Speech database that includes 20 read sentences from 300 speakers, diverse in age, sex, education level, and dialect, for a total of 6000 utterances (available from ELDA). The transcripts were unsatisfactory for a speech recognition application and hence had to be recreated. In order to augment these transcripts, we recruited Persian speakers from the diverse Los Angeles area and recorded read and semi-spontaneous speech data. The semi-spontaneous speech was solicited in a wizard of Oz scenario, while read speech was collected and verified by the speakers themselves with an interactive data collection tool. These data were not suitable for language modeling: For language models, in parallel to developing data mining techniques described below, an extensive data collection effort took place [1] where 300 Standardized Patient (medical student-actor patient) sessions were run at the USC campus in collaboration with the Medical School. The SP data was subsequently transcribed and translated thus resulting in over 300k words of in domain data in both English and Persian.

On the Pashto side, the only preexisting corpus of recorded Pashto we could find was a series of untranscribed Voice of America (VOA) Pashto service broadcasts, recorded by the Linguistic Data Consortium from the broadcasts. However, the data were not well matched for the target task: the broadcasts were dominated by fewer than a dozen speakers, showed only a fraction of the dialect diversity of Pashto, were not of very high audio quality, and did not reflect dialog speech style. We therefore recruited approximately 80 regionally diverse Pashtuns from a local émigré community, and asked each to record 100-200 spontaneously generated utterances, including answers to questions. These recordings totaled about 7 hours of speech. We also transcribed 5 hours of the VOA data. The acoustic model training data consisted of these 12 hours of speech, or approximately 100,000 running words.

For language modeling we used the above transcriptions plus additional in-domain text which was obtained by translating English role-played dialogs, yielding a total of 270,000 words. The vocabulary totaled 6680 different forms. The same data were used for translation purposes. The challenges associated with the collection and transcription of the corpus used for the Pashto system are described in more detail in the overview paper for the translation system in [2].

2.2. Transcriptions & Lexicons

The single greatest challenge to all aspects of both the Pashto and Persian languages in this project has been transcription schemes of the languages especially those that can accommodate machine spoken language processing. Pashto lacks a consistent, standardized writing system or spelling norms and as a consequence one word can be written in multiple ways, and different words the same way. To avoid the non-standardization of the native orthography we tried to transcribe acoustic data directly into a phonemic representation. We found however that awareness of the phonemes was quite difficult for transcribers to achieve, and made still more difficult by the fact that our phonological analysis was intended to cover a broad range of Pashto dialects and thus was not descriptive of any one speaker's inventory. We therefore used native (Arabic-based) script instead and found it to be more reliable despite its difficulties. The word list in native script orthography was then phonemically transcribed. In many cases multiple phonemic representations were associated with each script form. These representations may reflect widely differing pronunciations, or entirely different words written the same way, or variant but equivalent syntactic forms. Likewise, any given phonemic representation may be associated with one or several script forms, with one or several meanings intended. Speech recognition processing used an isomorphism of the native script, so that texts could be used for language modeling. This isomorphism was creating ambiguity both at the input and at the output of the recognizer. Each "isomorphic class", which was a "word" for all the practical purposes of recognizer training and testing, had multiple pronunciations, sometimes quite a large number of them, which impacted acoustic model training and the accuracy of the search. The output of the recognizer, on the other hand, corresponded to a sequence of word classes and was not disambiguating among word meaning for the same class, but passed this difficult task onto the translation engine.

Persian does have a standardized writing system, but one that is ill-conditioned for direct use in a speech to speech translation system. The Modern Persian language employs a borrowed alphabet from Arabic, modified by the addition of 4 letters and modification of two character shapes. Despite the ability to mark vowel sounds in the written script, this procedure is extremely rarely followed in the Persian transcribed language, thus generating a lossy encoding of the appropriate orthography (this problem incidentally is common, at varying degrees, to all languages that use the Arabic script). The solution was to create three different encoding schemes: A one-to-one mapping from the Arabic script to the Latin alphabet (USCPers), an augmented version (USCPers+) encoding the additional vowel information contained in spoken Persian, and a phonetic transcription scheme (USCPron) enabling the creation of the ASR and TTS components [3]. Given the framework of the above transcription schemes the next step was the collection of sufficient data for the creation of dictionary containing mappings along the three schemes. Clearly the generation of pronunciation from the Arabic script is an ill posed problem and thus dictionaries need to be generated by humans. During the initial data collection, transcribers & transliterators converted clean English utterances into these multiple formats thus simultaneously generating pronunciation dictionaries. The availability of some initial data allowed for some automation of the process using statistical learning methods [4].

2.3. Acoustic modeling

For both Persian and Pashto systems we used front-ends with 16 kHz sampling rate, 10 ms frame advance rate, 12 mel frequency cepstral coefficients plus normalized energy and first- and second-order differences (39 features). The phonetic set for Persian has 34 units (29 phonemes, silence, br, ls, ga, and lg), while for Pashto it is 43 (41 phones, silence, reject). 3-state triphone hidden Markov models (HMMS) were trained with state clustering. For the Persian 4207 clustered states where used with on average 14 Gaussians per state. The Pashto system used a much smaller model size to fit the limitations of the anticipated small footprint platform (129 phone-state Gaussian clusters with 32 Gaussians each), trained using discriminative maximum mutual information estimation (MMIE, [5]). The SONIC [6] speech recognition engine was used for the Persian system, while the Pashto system was based on DYNASPEAK [7].

Both systems used an English phoneme mapping into the target language as an initialization step for the models, which were subsequently adapted or re-trained using the limited amount of data available. For Pashto only a knowledge based (linguistic) phone mapping was used, while for the Persian models we investigated three different methods: a knowledge based one, a data driven phoneme mapping, and a data driven state mapping method. The data driven techniques employed the *Earth Movers Distance* (EMD) method, which tries to minimize the amount of work needed to change one GMM into another. By using EMD at the sub-phoneme level, we achieved an improvement of 2.7% phoneme level recognition over the usage of only the Persian speech data. This advantage, although promising for extremely limited data, is insignificant once the data size increases. In that case the models derived from cross-lingual phonetic alignment were used only for alignment purposes. We intend to further exploit the cross-lingual knowledge as we move our translation system into more resource-poor languages. Currently we have speech recognition engines in multiple languages (English, Persian, Arabic, Greek etc) thus the pool of potential GMM mixtures is increasing and we expect the potential benefits will be larger and provide a procedure of quick language portability into new languages and dialects.

2.4. Language modeling

The amount of data required for language modeling is orders of magnitude higher than for acoustic modeling to provide adequate surface form coverage for these 2-way S2S systems. However, as is the practice, written text can be used as a approximation to spoken language transcripts (although easy access to text data may be diffcult e.g., Persian or even not possible for some of the target languages e.g., Pashto). In the creation of appropriate language models for the Persian-English and Pashto-English translation systems we faced two major hurdles. The first was the lack of (medical) domain data, and the second was the lack of any general background data in Persian and Pashto for bootstrap.

Let us consider the case of the Persian-English system where there was some publicly available textual resources (medical domain data in English, and some Persian text, but relatively smaller amounts). In generating appropriate domain data multiple parallel approaches were followed: The first step was identifying medical domain text in existence, and for this purpose we employed medical phrase books, paraphrasing, wizard of Oz data collections etc. This material clearly is very limited and was used as a seed to mine web-data. The web provides an abundance of text and even some transcribed material but is also very difficult to identify and automatically filter the appropriate in-domain material. Our initial attempts were based on a bag-of-words approach [8], while subsequently significantly more advanced algorithms were developed. The current method [9] is based on an iterative web crawling approach which uses a competitive set of adaptive models comprised of a generic topic independent background language model, a noise model representing spurious text encountered in web based data (Webdata), and a topic specific model to generate query strings using a relative entropy based approach for WWW search engines and to weight the downloaded Webdata appropriately for building topic specific language models. This method resulted in a 14% improvement when compared to the results with a generic model built using only 5K words of in domain data as a seed corpus. In addition to providing in domain English data, the simplified bag-

of-words webdata approach was also used to mine Persian text, that enabled us to create a background language model in Persian. This effort did not have the same level of impact due to the lack of relevant web resources in Persian.

In parallel to developing data mining techniques, an extensive data collection effort took place [10] where 300 Standardized Patient (medical student-actor patient) sessions were run at the USC campus in collaboration with the Medical School. The SP data was subsequently transcribed and translated thus resulting in over 300K words of in domain data. The resulting language models were based on layers of information, most notably the Medical phrasebooks & paraphrases (English & Persian), in domain manually collected data (E&P), in domain web-data (E), generic web-data (P) and existing generic models (E). The resulting models represent a vocabulary of over 21,000 English words and over 8000 Persian words. Furthermore, both the English and Persian models are class based enabling us to employ human knowledge to augment important classes such as medication names, greetings, relationships, several types of named entities etc.

Because of the morphological complexity of Pashto and the relatively smaller amount of available training data, language modeling posed an even more serious challenge. We addressed the problem by adapting the algorithm presented in [7], and built a language model that had more fine-grained backoff layers than a traditional word n-gram language model. To achieve this, we first generated a clustering tree for the vocabulary with the root of the tree representing the whole vocabulary and every node representing a class that includes all words in its descendant nodes. The tree is generated using the minimum discriminative information clustering algorithm using a similarity metric based on the left and right contexts of a word. When estimating the conditional probability of a word based on its n-gram prefix, we first back off to its context with the most distant word replaced by its class, from the most specific to the most general, and if none of these backoffs could guarantee a minimum number of occurrences then back off to the normal lower-order (n-1)-gram prefix. The resulting language model achieves a relative perplexity reduction of over 10% and a significant word error rate reduction of 11% relative on in domain data.

3. ENABLING ROBUST SPOKEN INTERACTIONS

The next step in the design is to integrate and test the acoustic and language models within the recognizer implementation. Let us consider the case of the SRI DynaSpeak system to highlight the recognizer design especially to make it handheld compatible. The Persian-English recognizer used the Colorado Sonic system [6], and followed similar design steps. The other design issue relates to the user interface. We describe both these aspects below.

3.1. Recognition engine features

An important challenge that the recognition system had to meet for field use, was the use of a handheld computer as the hardware platform. While this platform has the benefit of a long battery life, that would allow autonomous use in the field for extended periods of time, its drawbacks are the lack of hardware floating point computation, slower speed, and more limited memory than standard PCs. While for previous simpler phrase translation systems [11] the DynaSpeak$^{\text{TM}}$ engine proved appropriate, the task of limited-domain spontaneous speech-to-speech translation required additional features.

The original DynaSpeak$^{\text{TM}}$ engine used a hierarchical and dynamic search strategy [7]; while this strategy is memory efficient it is expensive in computation, and is mainly designed for handling small rule-based grammars. To achieve the necessary speed for this task, where a medium vocabulary statistical language model is applied, we developed a new flat search on an optimized state-level decoding graph. The state-graph is generated via a determinization and minimization algorithm [12] in a special form of weighted finite state acceptor, in which symbols are attached with state instead of arcs, and word symbols are left at word ending states. An efficient Viterbi algorithm is implemented to perform decoding, which made it several times faster than the previous hierarchical search strategy. To generate alternative hypotheses for the later translation step, we also implemented an efficient n-best search algorithm. In the decoding graph, we mark the states where hypotheses recombination is likely to happen. At these states, which we called lattice states, we record all incoming theories during the search. A lattice representing alternative hypotheses can be constructed based on this information. An N-best list is extracted from the lattice using an efficient A* algorithm. Given the likelihood scores of n-best hypotheses, a confidence score can be evaluated for each hypothesis based on computing the posterior probability of each hypothesis and using an appropriate log probability scaling to smooth the posterior distribution.

3.2. Noise Robustness

The levels of environmental noise often found in real-world S2S applications present a significant challenge to speech recognition systems. Without specific noise-robust processing, even state-of-the-art speech recognition degrades rapidly under decreasing signal-to-noise ratios (SNR). Our approach to improve robustness of the ASR to noise is based on the simultaneous application of two complementary methods: (1) acoustic model training in noise and (2) feature compensation. For model training we combine clean speech with noisy speech which is obtained by adding a set of noise samples using a range of SNRs up to a minimum of 15 dB. By limiting the SNR in training to be at least 15 dB we ensured that there was almost no degradation when dealing with clean speech. The chosen noises are those likely to be found in the target environment, they are both stationary and non stationary. The relative error reduction achieved by this approach 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 [13], which is a piecewise linear transformation of a noisy feature space into a clean feature space. Recently, we extended the method to allow multi noise training and better performance on non stationary noises [14]. During recognition, different POF mapping sets can be dynamically selected based on real time estimates of the SNR of the current condition. In our prototype translation system we apply the POF feature compensation when the measured SNR is lower than 15 dB. For SNRs in the range of 5 to 15 dB we have achieved relative error reductions between 30% to 75% depending on the type of noise.

3.3. User interface (UI) aspects

3.3.1. Endpointer

In real world environments, accurate utterance endpointing is a difficult task. To increase robustness both of the translation devices employ a push-to-talk (PTT) procedure and a voice activity detection within the continuously acquired audio. This means that the PTT beginning and end signals are used only to cue likely start and endpoints, so the system is robust to user synchronization errors that would otherwise result in the speech being truncated. The explicit visual indication of the voice activity to user was found to be a useful feature.

3.3.2. Repair techniques, dialog management and user interface

The goal of the S2S translation system is to help the two participants communicate. Often, a simple gesture is worth a lot more than several correctly translated utterances. To this end the translation device needs to aid users in employing non-verbal methods and in translating common concepts accurately and fast. For example, if a user asks whether a certain body part hurts, it is a lot more reliable to ask if it hurts “here” and employ a gesture. Additionally it is often beneficial if the users are given translations that are extremely accurate in the translation.

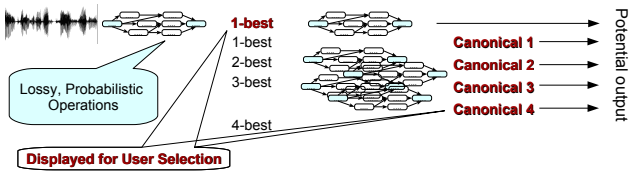

Fig. 1. As seen from the figure, the user is given up to 4 choices - the 4-top class canonical forms - that can be accurately translated by prior knowledge ("I can definitely translate these" section), while one choice is shown where the user is not guaranteed perfect translation ("I can try to translate these" section).

To explain some of the UI choices, consider first the Persian-English (Transonics) system design. To enable a more robust interaction, it employed a concept translator in addition to the statistical machine translation (Fig. 1). The 1-best choice from the speech recognizer is translated using a statistical translation approach, while the top four best choices are mapped using a competitive language model approach into preprogrammed concepts. By providing this classifier approach the system also allows for helping the trained user lower the risk of mistranslation. For example the system will attempt to use phrases that contain deictic information such as "point where it hurts" reduce risk of mistranslation. This device also requires much less data for training so it is feasible to enable concept augmentation by the end user. In addition build-in thresholds of translation and ASR confidence allow the automatic translation - pass-through mode - (if confidence is very high), or explicit confirmation of the utterance to be translated, or choice mode where the best choices are displayed for the user to select appropriately. Furthermore, to enable error reduction users are allowed to type in phrases or correct wrongly recognized input.

The user interface of the Transonics system went through several incarnations, of a range of combinations from mouse control to keyboard control, from extreme customization to no options to the user. The current interface has been the result of user studies into evaluating several of the previous interfaces, and soliciting users feedback and suggestions. It currently has two distinct versions: One version allows no customization thus being appropriate for the novice user and a second version allows for a range of options such as thresholds for automatic translation, for fallback to confirmation and fallback to choice modes, restricting the domain of conversation (along two axes. eg. greetings-viral infections or diagnosis-cancer etc). A planned third version will allow for inclusion of custom hot-buttons. The interface, which is organized in 4 sections, displays the dialog history pane on the left, current options (1-best + 3 canonical) in the middle and hot-buttons on the right. The middle-top represents the control & monitor area of the device. In addition, the device can be operated in multiple modes: It can be operated by an external keypad-like device where each of the buttons corresponds to an action such as "doctor", "patient", hot buttons "repeat", "rephrase" etc, or it can be operated using a pointing device. Additionally it can be operated using the keyboard by specified key-combinations.

Similar to the Persian system, the Pashto device [15] allows the user the choice between the pass-through, confirmation, and the choice modes, which adds reliability to critical communications or when recognition accuracy may be degraded because of environmental conditions. In addition, the Pashto system provides the choice that the operator can select a given hypothesis from the list and hit a button that plays "Did you say ?" concatenated with a synthesized version of the selected hypothesis. This feature is particularly important for languages like Pashto where there is no widespread use of written language, making visual confirmation methods difficult to use. The confidence scores can be used to help determine when the recognition accuracy may be low, and the confirmation mechanism engaged.

4. CONCLUSIONS

Design of speech to speech systems requires a holistic approach to enable robust interactions, including and beyond component level optimizations. One of the key problems is dealing with resource constraints. Another important aspect is the user interface design. Using examples drawn from SRI's English-Pashto system design and the USC-HRL English-Persian system, this paper reported on some of the design issues and engineering solutions.

5. REFERENCES

[1] Robert Belvin, Emil Ettelaie, Sudeep Gandhe, Panayiotis Georgiou, Kevin Knight, Daniel Marcu, Scott Millward, Shrikanth Narayanan, Howard Neely, and David Traum, "Transonics: A practical speech-to-speech translator for english-farsi medical dialogs," in *Proc. of The Association of Computational Linguistics*, Univ. of Michigan, Ann Arbor, June 2005.

[2] A. Kathol, K. Precoda, D. Vergyri, W. Wang, and S. Riehemann, "Speech translation for low-resource languages: The case of pashto," in *Eurospeech*, 2005.

[3] Shadi Ganjavi, Panayiotis G. Georgiou, and Shrikanth Narayanan, "A transcription scheme for languages employing the arabic script motivated by speech processing application," in *Proceedings of the Workshop Computational Approaches to Arabic Script-based Languages. 20th International Conference on Computational Linguistics (Coling 2004)*, 2004.

[4] Panayiotis G. Georgiou, Hooman Shirani-Mehr, and Shrikanth Narayanan, "Context dependent statistical augmentation of persian transcripts," in *Proceedings of the International Conference on Spoken Language Processing*, Jeju, Korea, October 2004.

[5] J. Zheng, J. Butzberger, H. Franco, and A. Stolcke, "Improved maximum mutual information estimation training of continuous density HMMs," in *Proc. EUROSPEECH*, 2001.

[6] Bryan Pellom, "SONIC: The university of colorado continuous speech recognizer, technical report TR-CSLR-2001-01," Tech. Rep., University of Colorado, March 2001.

[7] H. Franco, J. Zheng, J. Butzberger, F. Cesari, M. Frandsen, J. Arnold, R. Rao, A. Stolcke, and V. Abrash, "DynaSpeak: SRI's scalable speech recognizer for embedded and mobile systems," in *Proc. Human Language Technology Conference*, 2002.

[8] S. Narayanan, P. G. Georgiou, et al., "Transonics: A speech to speech system for English-Persian interactions," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, 2003.

[9] A. Sethy, P. G. Georgiou, and S. Narayanan, "Building topic specific language models from webdata using competetive models," in *Eurospeech*, 2005.

[10] Robert Belvin, Win May, Shrikanth Narayanan, Panayiotis Georgiou, and Shadi Ganjavi, "Creation of a doctor-patient dialogue corpus using standardized patients," in *Proc. LREC*, Lisbon, Portugal, 2004.

[11] H. Franco, J. Zheng, K. Precoda, F. Cesari, V. Abrash, D. Vergyri, A. Venkataraman, H. Bratt, C. Richey, and A. Sarich, "Development of phrase translation systems for handheld computers: From concept to field," in *Proc. of Eurospeech*, 2003.

[12] M. Mohri, F. Perreira, and M. Riley, "Weighted finite state transducers in speech recognition," in *In ISCA ITRW ASR*, 2000.

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

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

[15] K. Precoda, H. Franco, A. Dost, M. Frandsen, J. Fry, A. Kathol, C. Richey, S. Riehemann, D. Vergyri, J. Zheng, and C. Culy, "Limited-domain speech-to-speech translation between English and Pashto," in *Proc. HLT/NAACL 2004 Demonstrations*, Boston, MA, January 2004.