IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-speech Translator *

Yuqing Gao, *Liang Gu*, *Bowen Zhou, Ruhi Sarikaya, Mohamed Afify, Hong-Kwang Kuo*,
Wei-zhong Zhu, *Yonggang Deng*, *Charles Prosser*, *Wei Zhang* and *Laurent Besacier*
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

ABSTRACT

In this paper, we describe the IBM MASTOR, a speech-to-speech translation system that can translate spontaneous free-form speech in real-time on both laptop and hand-held PDAs. Challenges include speech recognition and machine translation in adverse environments, lack of training data and linguistic resources for under-studied languages, and the need to rapidly develop capabilities for new languages. Another challenge is designing algorithms and building models in a scalable manner to perform well even on memory and CPU deficient hand-held computers. We describe our approaches, experience, and success in building working free-form S2S systems that can handle two language pairs (including a low-resource language).

1. INTRODUCTION

Automatic speech-to-speech (S2S) translation breaks down communication barriers between people who do not share a common language and hence enable instant oral cross-lingual communication for many critical applications such as emergency medical care. The development of an accurate, efficient and robust S2S translation system poses a lot of challenges. This is especially true for colloquial speech and resource deficient languages.

The IBM MASTOR speech-to-speech translation system has been developed for the DARPA CAST and Transtac programs whose mission is to develop technologies that enable rapid deployment of real-time S2S translation of low-resource languages on portable devices. It originated from the IBM MARS S2S system handling the air travel reservation domain described in [1], which was later significantly improved in all components, including ASR, MT and TTS, and later evolved into the MASTOR multilingual S2S system that covers much broader domains such as medical treatment and force protection [2,3]. More recently, we have further broadened our experience and efforts to very rapidly develop systems for under-studied languages, such as regional dialects of Arabic. The intent of this program is to provide language support to military, medical and humanitarian personnel during operations in foreign territories, by deciphering possibly critical language communications with a two-way real-time speech-to-speech translation system designed for specific tasks such as medical triage and force protection.

The initial data collection effort for the project has shown that the domain of force protection and medical triage is, though limited, rather broad. In fact, the definition of domain coverage is tough when the speech from responding foreign language speakers are concerned, as their responses are less constrained and may include out-of-domain words and concepts. Moreover, flexible casual or colloquial speaking style inevitably appears in the human-to-human conversational communications. Therefore, the project is a great challenge that calls for major research efforts.

Among all the challenges for speech recognition and translation for under-studied languages, there are two main issues: 1) Lack of appropriate amount of speech data that represent the domain of interest and the oral language spoken by the target speakers, resulting in difficulties in accurate estimation of statistical models for speech recognition and translation. 2) Lack of linguistic knowledge realization in spelling standards, transcriptions, lexicons and dictionaries, or annotated corpora. Therefore, various different approaches have to be explored.

Another critical challenge is to embed complicated algorithms and programs into small devices for mobile users. A hand-held computing device may have a CPU of 256MHz and 64MB memory; to fit the programs, as well as the models and data files into this memory and operate the system in real-time are tremendous challenges [4].

In this paper, we will describe the overall framework of the MASTOR system and our approaches for each major component, i.e., speech recognition and translation. Various statistical approaches [5,6,7,8] are explored and used to solve different technical challenges. We will show how we addressed the challenges that arise when building automatic speech recognition (ASR) and machine translation (MT) for colloquial Arabic on both the laptop and handheld PDA platforms.

2. SYSTEM OVERVIEW

The general framework of our speech translation system is illustrated in Figure 1. The general framework of our MASTOR system has components of ASR, MT and TTS. The cascaded approach allows us to deploy the power of the existing advanced speech and language processing techniques, while concentrating on the unique problems in speech-to-speech translation. Figure 2 illustrates the MASTOR GUI (Graphic User Interface) on laptop and PDA, respectively.

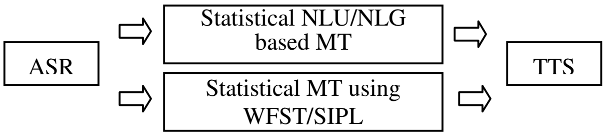
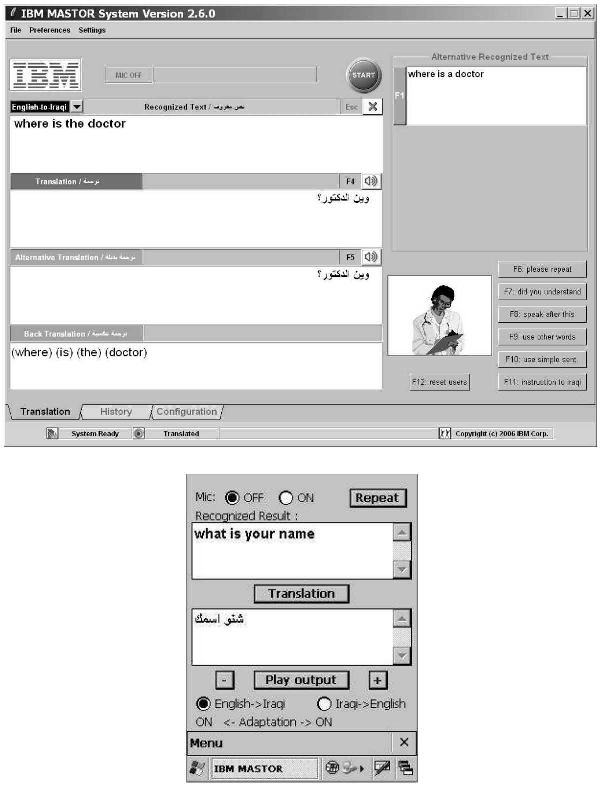


Figure 1 IBM MASTOR Speech-to-Speech Translation System

Acoustic models for English and Mandarin baseline are developed for large-vocabulary continuous speech and trained on over 200 hours of speech collected from about 2000 speakers for each language. However, the Arabic dialect speech recognizer was only trained using about 50 hours of dialectal speech. The training data for Arabic consists of about 200K short utterances. Large efforts were invested in initial cleaning and normalization of the training data because of large number of irregular dialectal words and variations in spellings. We experimented with three approaches for pronunciation and acoustic modeling: i.e. grapheme, phonetic, and context-sensitive grapheme as will be described in

* Thanks to DARPA for funding



where is the doctor

وين الدكتور؟

وين الدكتور؟

(where) (is) (the) (doctor)

where is a doctor

what is your name

شنو اسمك

Figure 2 IBM MASTOR system in Windows XP and Windows CE

section 3.A. We found that using context-sensitive pronunciation rules reduces the WER of the grapheme based acoustic model by about 3% (from 36.7% to 35.8%). Based on these results, we decided to use context-sensitive grapheme models in our system.

The Arabic language model (LM) is an interpolated model consisting of a trigram LM, a class-based LM and a morphologically processed LM, all trained from a corpus of a few hundred thousand words. We also built a compact language model for the hand-held system, where singletons are eliminated and bigram and trigram counts are pruned with increased thresholds. The LM footprint size is 10MB.

There are two approaches for translation. The concept based approach uses natural language understanding (NLU) and natural language generation models trained from an annotated corpus. Another approach is the phrase-based finite state transducer which is trained using an un-annotated parallel corpus.

A trainable, phrase-splicing and variable substitution TTS system is adopted to synthesize speech from translated sentences, which has a special ability to generate speech of mixed languages seamlessly [9]. In addition, a small footprint TTS is developed for the handheld devices using embedded concatenative TTS technologies.[10]

Next, we will describe our approaches in automatic speech recognition and machine translation in greater detail.

3. AUTOMATIC SPEECH RECOGNITION

A. Acoustic Models

Acoustic models and the pronunciation dictionary greatly influence the ASR performance. In particular, creating an accurate pronunciation dictionary poses a major challenge when changing the language. Deriving pronunciations for resource rich languages like English or Mandarin is relatively straight forward using existing dictionaries or letter to sound models. In certain languages such as Arabic and Hebrew, the written form does not typically contain short vowels which a native speaker can infer from context. Deriving automatic phonetic transcription for speech corpora is thus difficult. This problem is even more apparent when considering colloquial Arabic, mainly due to the large number of irregular dialectal words.

One approach to overcome the absence of short vowels is to use grapheme based acoustic models. This leads to straightforward construction of pronunciation lexicons and hence facilitates model training and decoding. However, the same grapheme may lead to different phonetic sounds depending on its context. This results in less accurate acoustic models. For this reason we experimented with two other different approaches. The first is a full phonetic approach which uses short vowels, and the second uses context-sensitive graphemes for the letter "A" (Alif) where two different phonemes are used for "A" depending on its position in the word.

Using phoneme based pronunciations would require vowelization of every word. To perform vowelization, we used a mix of dictionary search and a statistical approach. The word is first searched in an existing vowelized dictionary, and if not found it is passed to the statistical vowelizer [11]. Due to the difficulties in accurately vowelizing dialectal words, our experiments have not shown any improvements using phoneme based ASR compared to grapheme based.

Speech recognition for both the laptop and hand-held systems is based on the IBM ViaVoice engine. This highly robust and efficient framework uses rank based acoustic scores [12] which are derived from tree-clustered context dependent Gaussian models. These acoustic scores together with n-gram LM probabilities are incorporated into a stack based search algorithm to yield the most probable word sequence given the input speech.

The English acoustic models use an alphabet of 52 phones. Each phone is modeled with a 3-state left-to-right hidden Markov model (HMM). The system has approximately 3,500 context-dependent states modeled using 42K Gaussian distributions and trained using 40 dimensional features. The context-dependent states are generated using a decision-tree classifier. The colloquial Arabic acoustic models use about 30 phones that essentially correspond to graphemes in the Arabic alphabet. The colloquial Arabic HMM structure is the same as that of the English model. The Arabic acoustic models are also built using 40 dimensional features. The compact model for the PDA has about 2K leaves and 28K Gaussian distributions. The laptop version has over 3K leaves and 60K Gaussians. All acoustic models are trained using discriminative training [13].

B. Language Modeling

Language modeling (LM) of the probability of various word sequences is crucial for high-performance ASR of free-style openended coversational systems. Our approaches to build statistical tri-gram LMs fall into three categories: 1) obtaining additional training material automatically; 2) interpolating domain-specific LMs with other LMs; 3) improving distribution estimation robustness and accuracy with limited in-domain resources. Automatic data collection and expansion is the most straight-forward way to achieve efficient LM, especially when little in-domain data is available. For resource-rich languages such as English and Chinese, we retrieve additional data from the World Wide Web (WWW) to enhance our limited domain specific data, which shows significant improvement [6].

In Arabic, words can take prefixes and suffixes to generate new words which are semantically related to the root form of the word (stem). As a result, the vocabulary size in Arabic can become very large even for specific domains. To alleviate this problem, we built a language model on morphologically tokenized data by applying morphological analysis and hence splitting some of the words into prefix+stem+suffix, prefix+stem or stem+suffix forms. We refer the reader to [14] to learn more about the morphological tokenization algorithm. Morphological analysis reduced the vocabulary size by about 30% without sacrificing the coverage.

More specifically, in our MASTOR system, the English language model has two components that are linearly interpolated. The first one is built using in-domain data. The second component acts as a background model and is built using a very large generic text inventory that is domain independent. The language model counts are also pruned to control the size of this background model. The colloquial Arabic language model for our laptop system is composed of three components that are linearly interpolated. The first one is the basic word tri-gram model. The second one is a class based language model with 13 classes that covers names for English and Arabic, numbers, months, days, etc. The third one is the morphological language model described above.

4. SPEECH TRANSLATION

A. NLU/NLG-based Speech Translation

One of the translation algorithms we proposed and applied in MASTOR is the statistical translation method based on natural language understanding (NLU) and natural language generation (NLG). Statistical machine translation methods translate a sentence *W* in the source language into a sentence *A* in the target language by using a statistical model that estimates the probability of *A* given W, i.e. $p(A|W)$. Conventionally, $p(A|W)$ is optimized on a set of pairs of sentences that are translations of one another. To alleviate this data sparseness problem and, hence, enhance both the accuracy and robustness of estimating $p(A|W)$, we proposed a statistical concept-based machine translation paradigm that predicts *A* with not only *W* but also the underlying concepts embedded in *W* and/or *A*. As a result, the optimal sentence *A* is picked by first understanding the meaning of the source sentence W.

Let *C* denote the concepts in the source language and *S* denote the concepts in the target language, our proposed statistical concept-based algorithm should select a word sequence $\hat{A}$ as

$$\hat{A} = \arg\max_{A} p(A|W) = \arg\max_{A} \left\{ \sum_{S,C} p(A|S,C,W) p(S|C,W) p(C|W) \right\} ,$$

where the conditional probabilities $p(C|W)$, $p(S|C,W)$ and $p(A|S,C,W)$ are estimated by the Natural Language Understanding (NLU), Natural Concept Generation (NCG) and Natural Word Generation (NWG) procedures, respectively. The probability distributions are estimated and optimized upon a pre-annotated bilingual corpus. In our MASTOR system, $p(C|W)$ is estimated by a decision-tree based statistical semantic parser, and $p(S|C,W)$ and $p(A|S,C,W)$ are estimated by maximizing the conditional entropy as depicted in [2] and [7], respectively.

We are currently developing a new translation method that unifies statistical phrase-based translation models and the above NLU/NLG based approach. We will discuss this work in future publications.

B. Fast and Memory Efficient Machine Translation Using SIPL

Another translation method we proposed in MASTOR is based on the Weighted Finite-State Transducer (WFST). In particular, we developed a novel phrase-based translation framework using WFSTs that achieves both memory efficiency and fast speed, which is suitable for real time speech-to-speech translation on scalable computational platforms. In the proposed framework [15] which we refer to as Statistical Integrated Phrase Lattices (SIPLs), we statically construct a single optimized WFST encoding the entire translation model. In addition, we introduce a Viterbi decoder that can combine the translation model and language model FSTs with the input lattice efficiently, resulting in translation speeds of up to thousands of words per second on a PC and hundred words per second on a PDA device. This WFST-based approach is well-suited to devices with limited computation and memory. We achieve this efficiency by using methods that allow us to perform more composition and graph optimization offline (such as, the determinization of the phrase segmentation transducer **P**) than in previous work, and by utilizing a specialized decoder involving multilayer search.

During the offline training, we separate the entire translation lattice *H* into two pieces: the language model *L* and the translation model *M*:

$$M = Min\ (Min\ (Det\ (P) \circ T) \circ W)$$

where $\circ$ is the composition operator, Min denotes the minimization operation, and Det denotes the determinization operation; *T* is the phrase translation transducer, and *W* is the phrase-to-word transducer. Due to the determinizability of *P*, *M* can be computed offline using a moderate amount of memory.

The translation problem can be framed as finding the best path in the full search lattice given an input sentence/automaton *I*. To address the problem of efficiently computing $I \circ M \circ L$, we have developed a multilayer search algorithm.

Specifically, we have one layer for each of the input FSM's: *I*, *L*, and *M*. At each layer, the search process is performed via a state traversal procedure starting from the start state $\vec{s}_0$, and consuming an input word in each step in a left-to-right manner.

We represent each state **s** in the search space using the following 7-tuple: s_I, s_M, s_L, c_M, c_L, $\bar{h}$, s_{prev}, where s_I, s_M, and s_L record the current state in each input FSM; c_M and c_L record the accumulated cost in *L* and *M* in the best path up to this point; $\bar{h}$ records the target word sequence labeling the best path up to this point; and s_{prev} records the best previous state.

To reduce the search space, two active search states are merged whenever they have identical s_I, s_M, and s_L values; the remaining state components are inherited from the state with lower cost. In addition, two pruning methods, histogram pruning and threshold or beam pruning, are used to achieve the desired balance between translation accuracy and speed.

To provide the decoder for the PDA devices as well that lacks a floating-point processor, the search algorithm is implemented using fixed-point arithmetic.

5. CONCLUSION

We described the framework of the IBM MASTOR system, the various technologies used in building major components for languages with different levels of data resources. The technologies have shown successes in building real-time S2S systems on both laptop and small computation resource platforms for two language pairs, English-Mandarin Chinese, and English-Arabic dialect. In the latter case, we also developed approaches which lead to very rapid (in the matter of 3-4 months) development of systems using very limited language and domain resources. We are working on improving spontaneous speech recognition accuracy and more naturally integrating two translation approaches.

6. ACKNOWLEDGEMENT

The authors sincerely thank Drs. Yoshinori Tahara, Fu-hua Liu, Yongxing Li, Etienne Marcheret, Raimo Bakis, Ellen Eide, Burn Lewis, Tony Lee, Ossama Emam, and Lubos Ures for their help and contributions to the MASTOR S2S system.

7. REFERENCES

[1] Y. Gao et al, "*MARS*: A Statistical Semantic Parsing and Generation Based *Multilingual Automatic tRanslation System*," *Machine Translation*, vol. 17, pp.185-212, 2004.

[2] L. Gu et al, "Improving Statistical Natural Concept Generation in Interlingua-based Speech-to-Speech Translation," in *Proc. Eurospeech'2003*, pp.2769-2772.

[3] F.-H. Liu, "Robustness in Speech-to-Speech Translation," in *Proc. Eurospeech'2003*, pp.2797-2800.

[4] B. Zhou et al, "Two-way speech-to-speech translation on handheld devices," in Proc. *ICSLP'04*, South Korea, Oct, 2004.

[5] H. Erdogan et al, "Using Semantic Analysis to Improve Speech Recognition Performance," *Computer Speech and Language*, vol.19, pp.321-343, 2005.

[6] R. Sarikaya, et al, "Rapid Language Model Development Using External Resources for New Spoken Dialog Domains," in *Proc. ICASSP'05*, Philadelphia, PA, Mar, 2005.

[7] L. Gu et al, "Concept-based Speech-to-Speech Translation using Maximum Entropy Models for Statistical Natural Concept Generation," *IEEE Trans. Speech and Audio Processing*, vol.14, no.2, pp.377-392, March, 2006.

[8] B. Zhou et al, "Constrained phrase-based translation using weighted finite-state transducers," in *Proc. ICASSP'05*, Philadelphia, Mar, 2005.

[9] E. Eide et al, "Recent Improvements to the IBM Trainable Speech Synthesis System," in *Proc. ICASSP*, Hong Kong, China, 2003.

[10]Dan Chazan et al, "Reducing the Footprint of the IBM Trainable Speech Synthesis System," in *ICSLP-2002*, pp.2381-2384

[11]R. Sarikaya et al, "Maximum Entropy Based Vowelization of Arabic," Interspeech2006 (submitted for publication).

[12]L.R. Bahl, et al, "Robust methods for using context-dependent features and models in a continuous speech recognizer," in *Proc. ICASSP*, 1994

[13]D. Povey & P.C. Woodland, "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," In *Proc. ICASSP*, Orlando, 2002.

[14]M. Afify et.al, "On the Use of Morphological Analysis for Dialectal Arabic Speech Recognition," Interspeech 2006 (submitted for publication).

[15]B. Zhou, S. Chen, and Y. Gao, "Fast Machine Translation Using Statistical Integrated Phrase Lattices," submitted to COLING/ACL'2006.