A Grammar-Based Chinese to English Speech Translation System for Portable Devices

Pascale FUNG, LIU Yi, YANG Yongsheng, Yihai SHEN, and Dekai WU

Human Language Technology Center
University of Science & Technology (HKUST), Clear Water Bay, Hong Kong
{pascale,eeyliu,ysyang,shenyh,dekai}@cs.ust.hk

Abstract

Portable devices such as PDA phones and smart phones are increasingly popular. Many of these devices already have voice dialing capability. The next step is to offer more powerful personal-assistant features such as speech translation. In this paper, we propose a system that can translate speech commands in Chinese into English, in real-time, on small, portable devices with limited memory and computational power. We address the various computational and platform issues of speech recognition and translation on portable devices. We propose fixed-point computation, discrete front-end speech features, bi-phone acoustic models, grammar-based speech decoding, and unambiguous inversion transduction grammars for transfer-based translation. As a result, our speech translation system requires only 500k memory and a 200MHz CPU.

1. Introduction

Enabled by the ever-increasing hardware power of portable devices such as personal digital assistants (PDAs), PDA phones and smart phones, more and more functionalities can be incorporated into such devices. These portable devices are not merely communication and storage devices, but also information access devices [1]. Incorporating speech recognition technologies into portable devices is of increasing interest in recent years. Speech input provides a natural and easy user interface, and is faster than handwriting recognition or keypads. A portable speech translation system, moreover, will enable business and leisure travelers (who are already heavy mobile phone users) to communicate more efficiently in day-to-day situations, such as booking tickets, reserving hotel rooms, and so on.

There are several major challenges in deploying speech translation on portable devices. First, the hardware resources (e.g., memory and CPU) of portable devices are much more limited than that of desktop machines. Second, most portable devices only support fixed-point calculations rather than floating-point calculations, yet speech recognition systems need to perform a large amount of floating-point calculations for front-end signal processing and computing output likelihoods during decoding. Consequently, speech recognizers ported directly from desktop versions would have very low efficiency. The system response time increases and becomes unacceptable by users. Third, large-vocabulary continuous speech recognition (LVCSR) cannot be run on portable devices, whereas isolated word recognition is not enough for speech translation and retrieval. We are required to establish a suitable and efficient design of the speech recognizer that supports a flexible user speech input, and generates sufficient information for translation and retrieval. Finally, the challenge of translation on a portable device is the trade-off between translation quality and speed, given real-time computation with limited hardware resources. Statistical machine translation algorithms with powerful performance on servers have to be modified according to the platform specification of portable devices.

Fortunately, speech translation applications on portable devices do not require a large vocabulary, or generic grammar coverage. Words and sentence structures used in ticketing booking, hotel reservation, airline information enquiry, etc. tend to be limited. A portable speech translation device should be able to dynamically load up different lexicon, grammars, and translation memory, for different applications.

In this paper, we describe the design of a dynamic-domain speech translation and retrieval system, on portable devices, which allows flexible speech input and generates translations that are sufficiently accurate and fluent. In order to make speech translation perform smoothly and efficiently on portable devices under the constraints of limited memory size, low CPU speed and fixed-point calculation, we propose (1) using a fixed-point, front-end feature extraction method for speech signal processing, (2) using discrete HMM class-based bi-phone models instead of the original continuous HMM triphone models; (3) using state-of-the-art context free grammar (CFG) decoder for dynamic vocabulary and grammar upload; (4) using unambiguous inversion transduction grammars (ITGs) with an integrated parsing, transduction and generation process, for translation.

The architecture of our speech translation system is shown in Figure 1.

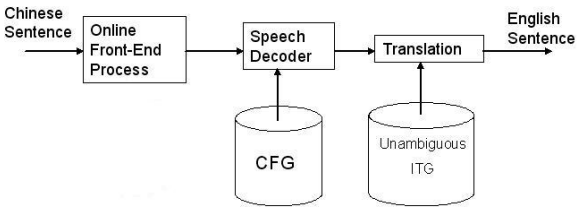

Figure 1: System architecture.

2. Front End Processing

Conventional speech recognition systems use mel-frequency cepstral coefficients (MFCCs) generated by front-end signal processing as the acoustic observations. The MFCCs can be

computed using the algorithm shown in Figure 2 [2]. This front-end processing of speech signals is mathematically intensive, and requires a large amount of float-point emulation. In profiling the source code, we found that the computation time increases over 100 times (e.g., from 20ms to 2s) if we directly adopt the MFCC feature extraction process from desktop version to portable devices, which greatly increases the recognition response time. To reduce the computation time for MFCC feature extraction, we first change the float-point emulation to fixed-point emulation using scaled integers, and then simplify some computations in MFCC extraction using a polynomial approach (e.g., log function calculation) and a look-up table.

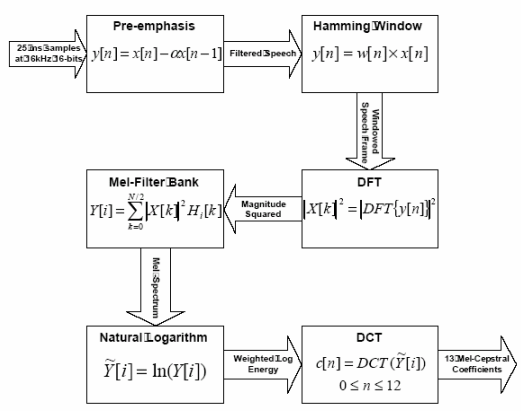

Figure 2: The MFCC computation algorithm.

The fixed-point emulation process utilizes scaled integers to perform basic mathematical functions using the existing integer hardware of portable devices. The scaling factor is either fixed at design time or determined dynamically for each specific calculation item. Fixed-point FFT in MFCC feature extraction leads to over 10 times reduction in computation time on portable devices. In order to avoid overflow problems in fixed-point emulation, we rewrite the Mel-filter bank by using the magnitude

$$Y[i] = \sum_{k=0}^{N/2} (|X[k]\sqrt{H_i[k]})^2$$

where $H_i[k] << 1$, therefore the value of the result of each multiplication is small. To further reduce the computation time, the coefficients of $\sqrt{H_i[k]}$ are precomputed and stored in a look-up table. Similarly, we also generate a look-up table stored in memory for the original floating-point calculation of sine and cosine functions in DCT. As a result, we can rapidly compute the values of Mel-filter bank and DCT through look-up in the relevant tables without complex float-point calculations. In addition, we use polynomial linear combination to estimate the calculation of log function.

From these modified and optimized approaches of front-end processing of speech signals, the computation time on portable devices is decreased to around 80ms, compared to the original time of 2100ms, resulting in a fast real-time and on-line feature extraction process.

3. Acoustic Modeling

Due to memory size and CPU speed limitations on portable devices, we propose using class-based bi-phone models. The input speech to our system is Mandarin Chinese, and the basic and natural phone-level units for Mandarin consist of 27 initials and 38 finals [3]. We first divide the initials into seven different classes according to their pronunciation similarities from phonological rules [4], and finals into five different classes according to their first vowel. Each individual initial or final is combined with a particular class of finals or initials to form a class-based bi-phone unit. For example, an independent initial unit 'b' has five distinguished class-based bi-phones, i.e., 'ba', 'bi', 'bu', 'bo' and 'be'.

With the use of class-based biphone models we achieve a good balance of recognition accuracy and memory size. The triphone models commonly used on desktop machines are capable of achieving high recognition accuracy, but need a large amount of memory (over 2M generally speaking) and a powerful CPU for on-line decoding. Such memory and CPU resources are typically impractical on portable devices. On the other hand, context-independent initial/final (i.e., mono-phone) models only require a very small memory size; however, the recognition accuracy of using such model is much lower than that of using triphone models. Experimental results showed that this level of accuracy leads to too many errors for the subsequent speech translation process. Our proposed class-based bi-phone model is at an intermediate level between triphone and mono-phone models. Compared to triphone models, the class-based bi-phone models reduce the memory size from over 2M to less than 400K while sacrificing less than 2% recognition accuracy. Compared to mono-phone models, the class-based bi-phone models only slightly inflate the model size, but achieve over 10% accuracy increase.

4. Grammar-Based Decoding

Our speech translation system includes a state-of-the-art context free grammar (CFG) speech decoder. Compared to conventional isolated word decoder or continuous speech/LVCSR decoders, our CFG decoder supports dynamic vocabulary and grammars to meet the platform requirements of limited memory size and low CPU speed. A grammar generation tool is provided for the developer to write CFGs for each different application. Moreover, one important advantage in our CFG decoder is that we can attach different attributes in the grammar associated with the CFG decoder. These attributes are passed to the subsequent modules of speech translation and retrieval. As a result, the attribute values can be derived according to the exact requirements represented by the grammar. For example, our decoder is able to identify the difference of "one-two-three" as a digit string from the number "one hundred and twenty three" even if they are both represented by same character string "123". Associated with attributes, our decoder is well suited for speech translation and retrieval within the grammar constrain framework, since the focused keywords in speech translation and retrieval can be easily indexed by corresponding attributes.

Meanwhile, in order to run the CFG decoder smoothly and efficiently under limited memory size and CPU speed, we propose an approach of automatically control of searching coefficients (e.g., pruning thresholds). We first set the searching coefficients automatically according to the size of initial searching space determined by the set of confusable words. Then the pruning coefficients are dynamically adjusted/adapted according to the acoustic likelihood ratio from n-best list, during the searching process. If the initial search space is large, we will set the pruning coefficients to higher value in order to reduce the computation time. On the other hand, if the search space is small, we assign a lower value to the pruning coefficients in order to guarantee sufficient candidate paths for accurate decoding result. Our proposed method of automatically setting and adapting searching coefficients is able to greatly reduce the redundant computing cost and achieve a good tradeoff between recognition accuracy and speed under constrained hardware conditions. The output of the decoder is a Chinese sentence, but it can also produce a one-best sentence or character sequence, n-best lists, or word networks for future use.

5. ITG-based Translation

The translation component consists of (1) an inversion transduction grammar or ITG [5] in conjunction with (2) a transduction algorithm. Like other statistical MT models, our general ITG model can be time- and memory-intensive. Thus restricted models are required to meet the computational design goals for portable devices.

The ITG model was selected for its attractive balance of expressiveness and computational complexity. With respect to expressiveness, Zens & Ney [6] have shown that a variant of the ITG achieves 96.1% coverage of an automatically word-aligned version of the Canadian Hansards parallel corpus, and 96.5% coverage on the Verbmobil corpus. For our application, which is far more restricted, the translations are satisfactory with respect to both fluency and accuracy. In specialized domains like travel and ticket booking, users ask questions that the service systems are expecting, relating for instance to things like flight time, flight number, or destination city. Therefore, we expect the words and structures to be limited as well. Transduction rules can feasibly be written manually to capture the majority of common structures in the users' utterances.

With respect to complexity, efficient polynomial-time algorithms have been shown for bibracketing, biparsing, and word-alignment [5] as well as translation [7]. ITGs are computationally efficient because transductions are restricted to only **straight** and **inverted** rules, rather than arbitrary permutations [5]. A context-free transduction grammar (or **bigrammar**) is a bilingual generalization of CFGs that can be seen as a generator of string-pairs, and takes the form of a set of transduction rules (or **birules**). An ITG permits only straight and inverted birules. For example, $\mathrm{NP} \rightarrow [\mathrm{NU\ NP}]$ is an example of a birule with **straight** orientation, which means that the transduction does not change the order of the source language side symbols and produces a target language side which is still NU NP (NP and NU stand for noun phrase and noun phrase specifying number respectively). On the other hand, $\mathrm{VP} \rightarrow [\mathrm{NT\ VP}]$ is a birule with an **inverted** orientation, in which VP and NT stand for verb phrase and noun phrase specifying time respectively, VP is the left hand symbol, NT VP the source language side and VP NT is the target language side resulted from the transduction. Refer to Figure 3 for concrete examples of birules with straight orientation and inverted orientation.

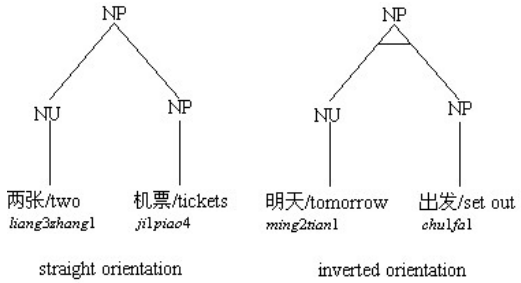

Figure 3: Examples of birules with straight and inverted orientation.

Although the transduction algorithm is actually a single integrated process, for conceptual purposes it can be described in terms of the three steps of traditional transfer MT as follows:

- *Parsing*. The input Chinese sentence from the decoder is parsed using the source language side of the transduction rules, which yields a parse tree of the input. For example, Figure 4 shows the parse tree of the Chinese input sentence "我明天从香港出发去南京".

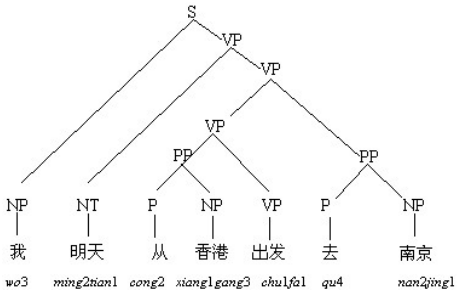

Figure 4: Example parse tree for a Chinese input sentence.

- *Transfer*. Via the transduction rules, the parse tree of the Chinese input is mapped to an English tree that generates the output translation. In other words, by looking not only at the source language side of the transduction rules but also their target language side, we now see the Chinese parse tree as a Chinese-English biparse tree. For example, Figure 5 shows the Chinese-English biparse tree of the Chinese input "我明天从香港出发去南京". Notice the node in the parse tree labeled as VP with two children labeled as NT and VP, which is an instance of the transduction rule $\mathrm{VP} \rightarrow \langle \mathrm{NT\ VP} \rangle$. This means that in the corresponding constituent for the generation side, the children of the VP node are VP and NT.

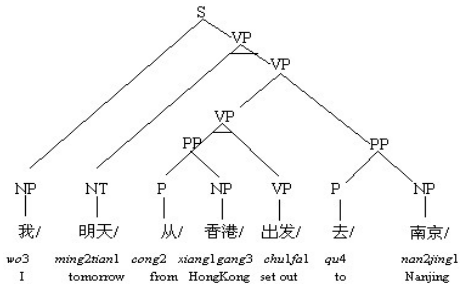

Figure 5: Example biparse tree for a Chinese input sentence.

- *Generation*. The leaf nodes of the English side of the biparse tree are output in left-to-right order. For example, the English string generated by the tree in Figure 6 is: "I set out from HongKong to Nanjing tomorrow".

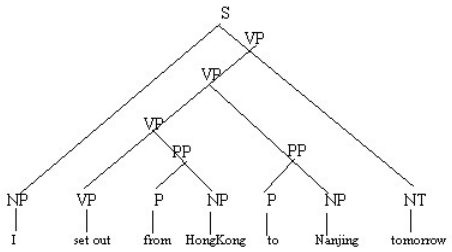

Figure 6: Example of the output of generation from the biparse tree.

In order to meet the design goals, we further constrain the ITG to be an **unambiguous** ITG, thus facilitating a lightweight transduction algorithm suitable for small devices. The rules are constrained such that the set of rules in the bigrammar is unambiguous on both the parsing and transduction levels, that is, the parsing step and the transduction step of the translation processes can be carried out without fear of ambiguity. By building the system we have found empirically that this strong constraint is feasible for limited domains like the Ticketing Booking Service, although it is doubtless infeasible for more general domains.

We employed a variant of the transduction algorithm for translating using ITGs as described by Wu [9] and Wu & Wong [7]. The algorithm can be characterized as a bilingual generalization of Earley parsing [8]. However, unlike [9], we utilized a linguistic bigrammar with real constituent structures, as opposed to simply a BTG (bracketing ITG) containing only generic rules such as $A \rightarrow [AA]$ or $A \rightarrow \langle AA \rangle$. Moreover, the linguistic bigrammar is formulated for precise parsing of the input language, as opposed to robust parsing biased toward the output language as in [7]. This has the consequence of permitting us to dispense with the language model, reducing time and space complexity by an order of magnitude. Because the bigrammar is unambiguous, the time and space complexity is reduced from the Earley upper bound of $O(n^3)$ to being linear in the length n of the input sentence.

Writing transfer rules tends to be both difficult and time consuming when the system is extended to cover a wider range of user utterances or to translate in more general domains; the system, therefore, should be easily adaptable to stochastic formulations to increase disambiguation power. The ITG model is easily made stochastic by assigning probabilities to the transduction rules, which may be estimated via estimation-maximization (EM) [10].

6. Conclusions

Compared to a continuous speech recognizer, our CFG decoder has very small memory size requirements. Unlike isolated word decoders, our CFG decoder supports flexible user speech input ranging from isolated words to short phases and full utterances, which provides sufficient information for speech translation. The translation component takes the output of the Chinese speech recognizer in the form of a sentence, or word lattices, and decodes the sentence into English in linear time and space, as dictated by an unambiguous inversion transduction grammar.

7. Acknowledgements

This work is partly supported by grant S/P584/03A of the Innovation & Technology Commission, and grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09 of the Research Grants Council, both of the Hong Kong government. We thank ZHAI Lufeng for his help in editing this paper.

8. References

[1] Eric CHANG, Frank SEIDE, Helen M. MENG, Zhuoran CHEN, YU Shi, and Yuk-Chi LI. A system for spoken query information retrieval on mobile devices. *IEEE Transactions on Speech and Audio Processing* 10(8):531-541, Nov 2002.

[2] Brian DELANEY, Nikil JAYANT, Mat HANS, Tajana SIMUNIC and Andrea ACQUAVIVA. *A low-power, fixed-point, front-end feature extraction for a distributed speech recognition system*. HP Laboratories Technical Report, HPL-2001-252. 9 Oct 2001.

[3] J.H. HUANG. Chinese Dialects, Xiamen University Press, Xiamen, 1987 (Chinese version).

[4] LIU Yi and Pascale FUNG. Partial change accent models for accented Mandarin speech recognition. *Proceedings of the IEEE Automatic Speech Recognition and Understanding*. St. Thomas, U.S. Virgin Islands: Dec 2003.

[5] Dekai WU. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics* 23(3):377-404. Sep 1997.

[6] Richard ZENS and Hermann NEY. A Comparative Study on Reordering Constraints in Statistical Machine Translation. *ACL-03: 41st Annual Meeting of the Assoc. for Computational Linguistics*. Sapporo, Japan: Jul 2003.

[7] Dekai WU and Hongsing WONG. Machine translation with a stochastic grammatical channel. *COLING-ACL'98*. Montreal: Aug 1998.

[8] Jay EARLEY. *An Efficient Context-Free Parsing Algorithm*, Ph.D. Thesis, Computer Science Dept., Carnegie Mellon University. Pittsburg, PA. 1968.

[9] Dekai WU. A polynomial-time algorithm for statistical machine translation. *ACL-96: 34th Annual Meeting of the Assoc. for Computational Linguistics*. Santa Cruz, CA: Jun. 1996.

[10] Dekai WU. Trainable coarse bilingual grammars for parallel text bracketing. *WVLC-3*: *3rd Annual Workshop on Very Large Corpora,* 69-82. Cambridge, MA: Jun 1995.