Dealing with Multilinguality in a Spoken Language Query Translator

Pascale Fung, Bertram Shi, Dekai Wu, Lam Wai Bun, Wong Shuen Kong
Dept. of Electrical and Electronic Engineering, Dept. of Computer Science
University of Science & Technology (HKUST)
Clear Water Bay, Hong Kong
{pascale,eebert}@ee.ust.hk,dekai@cs.ust.hk

Abstract

Robustness is an important issue for multilingual speech interfaces for spoken language translation systems. We have studied three aspects of robustness in such a system: accent differences, mixed language input, and the use of common feature sets for HMM-based speech recognizers for English and Cantonese. The results of our preliminary experiments show that accent differences cause recognizer performance to degrade. A rather surprising finding is that for mixed language input, a straight forward implementation of a mixed language model-based speech recognizer performs less well than the concatenation of pure language recognizers. Our experimental results also show that a common feature set, parameter set, and common algorithm lead to different performance output for Cantonese and English speech recognition modules.

1 Introduction

In the past few decades, automatic speech recognition (ASR) and machine translation (MT) have both undergone rapid technical progress. Spoken language translation has emerged as a new field combining the advances in ASR and MT(Levin et al., 1995; Mayfield et al., 1995; Lavie et al., 1995; Vilar et al., 1996). Robustness is a critical issue which must be addressed for this technology to be useful in real applications. There are several robustness issues arising from the multilingual characteristics of many spoken language translation systems which have not studied by the speech recognition community since the latter tends to focus on monolingual recognition systems.

One problem in a multilingual system is accent variability. It is frequently assumed that the speakers using a system are native speakers belonging to the same accent group. However, this is not generally true. For example, in Hong Kong, although many people can speak English, one encounters a large variety of different accents since in addition to Hong Kong's large population of Cantonese speakers, there are also many Mandarin speakers and many Indian, British, American and Australian Hong Kong residents.

Another problem with multilinguality is mixed language recognition. Although the official languages of Hong Kong are English, spoken Cantonese and written Mandarin, most Hong Kongers speak a hybrid of English and Cantonese. In fact, since many native Cantonese speakers do not know the Chinese translations of many English terms, forcing them to speak in pure Cantonese is impractical and unrealistic.

A third problem is the complexity of the design of recognizers for multiple languages. Many large multilingual spoken language translation systems such as JANUS (Lavie et al., 1995) and the C-STAR Consortium decouple the development of speech recognition interfaces for different languages. However, for developers of a multilingual system at one single site, it would be more efficient if the speech interfaces for the different languages shared a common engine with one set of features, one set of parameters, one recognition algorithm and one system architecture, but differed in the parameter values used.

We are studying the issues raised above in the domain of a traveling business-person's query translation system (Figure 1). This translator is a symmetrical query/response system. Both ends of the system recognize input speech from human through a common recognition engine comprising of either a concatenated or a mixed language recognizer. After the speech is decoded into text, the translator converts one language to another. Both ends of the system have a speech synthesizer for output speech. The domain of our system is restricted to points of interest to a traveling business-person, such as names and directions of business districts, conference centers, hotels, money exchange, restaurants. We are currently implementing such a system with Cantonese and English as the main languages. We use HMM-based, isolated word recognition system as the recognition engine, and a statistical translator for the translation engine.

2 Does accent affect speech recognizer performance?

We have performed a set of experiments to compare the effect of different accents. We train two sets of models: an English model using native American English speakers as reference and a Cantonese model using native Cantonese speakers as references. Word models of 34 (17 English and 17 Cantonese) simple commands were trained using 6 utterances of each command per speaker. The models were evaluated using a separate set of native Cantonese and native American English speakers. The recognition results are shown in Figure 2.

Our experimental results support the claim that recognition accuracy degrades in the presence of an unmodelled accent. In order to bring the recognizer performance for the non-native speaker to that of the native speaker, we need to improve the models in the recognizer. An obvious solution seems to train the model on different accents. However, it is quite a daunting task to train every language with every type of accent. One approximation is to train the system with a mixture of separate languages so that the model parameters would capture the spectral characteristics of more than one language. A mechanism for gradual accent adaptation might potentially increase recognition accuracies of the speech recognizers of both source and target languages.

3 How to deal with mixed language recognition?

Consider two possible ways to implement a mixed language recognizer—(1) Use two pure monolingual recognizers to recognize different parts of the mixed language separately; (2) Use a single mixed language model where the word network allows words in both languages. Method (1) requires some sort of language identification to switch between two recognizers whereas method (2) seems to be more flexible and efficient.

We compared the recognition accuracies of a pure language recognizer with a mixed language recognizer. In the pure language recognizer, the word candidates are all from a single language dictionary, whereas the mixed language dictionary contains words from two dictionaries. See Figure 3. In the concatenation model, we assume *a priori* knowledge (possibly from a language identifier) of the language ID of the words. The *expected* recognition rate of the concatenation model is the product of the accuracies of the pure language model.

From this preliminary experiment, we discover that although a *mixed* language model offers greater flexibility to the speaker, it has a considerably lower performance than that of the *concatenation* of two pure language models. The reason for such a performance degradation of a mixed model is not difficult to deduce—the dictionary of a mixed model has more candidates. Consequently, the search result is less accurate. If the recognizer knows *a priori* which dictionary (English or Chinese) it should search for a particular word, it would make less error.

This is therefore a potentially interesting problem. Should we incorporate a language identifier in parallel to the recognizers or should we accept the loss in recognition rate but enjoy the flexibility of a mixed language recognizer? We will implement a language identifier and carry out more experiments to compare the output from the recognizers.

4 Can the source and target languages share the same recognition engine?

One important issue for multilinguality in a spoken language translator is the complexity of implementing more than one recognizer in the system. An efficient approach is to use recognizers which are identical except for parameter values. Will this enable robust recognizers?

The word-based HMM recognizers for English and Cantonese use identical features (Nine MFCCs and nine delta MFCCs.) The same microphone was used to record both languages. The same initialization procedure was used to initialize the recognizer for both languages. For English, the number of HMM states is deduced from spectrograms. For Cantonese, it is deduced from phoneme numbers for each word. The recognizers were evaluated using native English and Cantonese speakers who were not in the training set.

In general, the English recognizer is more robust than our Cantonese recognizer even though identical parameter set, training and testing mechanisms are used. Rather than jumping to the conclusion that a different feature set is needed for Cantonese, we would like to find out what other factors could cause a lower performance of the Cantonese recognizer. For example, we would like to perform experiments on a larger number of speakers to determine whether training and test speaker mismatch caused such a performance degradation.

5 Conclusion and future work

In this paper, we have examined three issues concerning the robustness of multilingual speech interfaces for spoken language translation systems: accent differences, mixed language input, and the use of common feature sets for HMM-based speech recognizers for English and Cantonese. From the re-

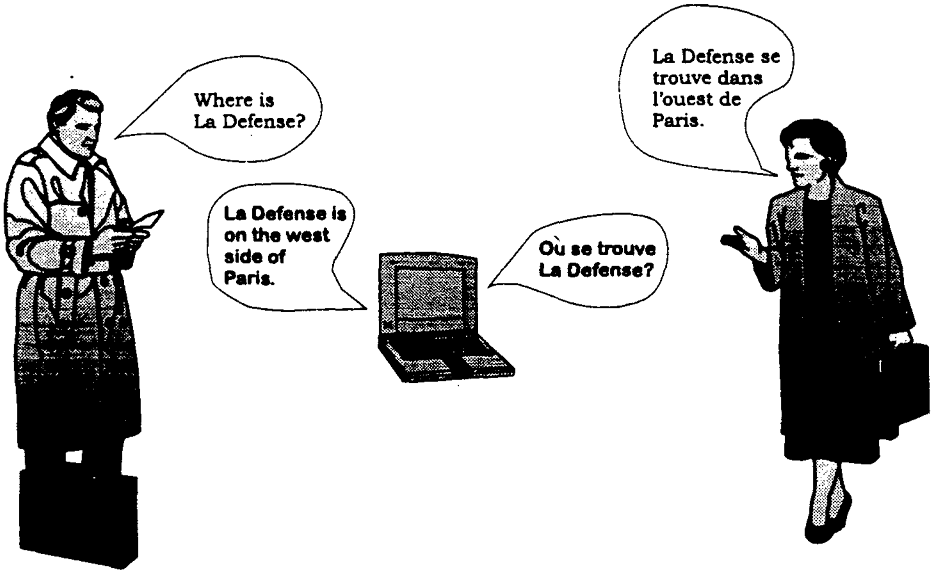

Figure 1: A symmetrical system as traveling business-person's query translator

Figure 2: Speech recognizers perform better on native speakers

Native speaker	English model	Cantonese model
English	94%	77%
Cantonese	86%	90%
Average	90%	83%

Figure 3: Speech recognizers perform better with concatenated pure language model than with mixed language model

Native speaker	Speech	Mixed model	English only	Cantonese only	Concatenate (expected)
English	English	92%	94%	-	-
	Cantonese	59%	-	77%	-
	Mixed	64%	-	-	66%(72%)
Cantonese	English	86%	86%	-	-
	Cantonese	75%	-	90%	-
	Mixed	68%	-	-	78%(77%)

sults of our preliminary experiments, we find that accent difference causes recognizers performance to degrade. For mixed language input, we found out that a straight forward implementation of a mixed language model-based speech recognizer performs less well than the concatenation of pure language recognizers due to the increase in recognition candidate numbers. Finally, our experimental results show that the Cantonese recognizer has a lower recognition rate on the average than the English recognizer despite a common feature set, parameter set, and common algorithm. We will perform more expriments using larger training and test sets to verify our results.

References

A. Lavie, L. Levin, A. Waibel, D. Gates, M. Gavalda, and L. Mayfield. 1995. JANUS: Multi-lingual translation of spontaneous speech in a limited domain. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, pages 252–255, Montreal, Quebec, October.

L. Levin, O. Glickman, Y. Qu, C. P. Rose, D. Gates, A. Lavie, A. Waibel, and C. Van Ess-Dykema. 1995. Using context in machine translation of spoken language. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 173–187, Leuven, Belgium, July.

L. J. Mayfield, M. Gavalda, Y.-H. Seo, B. Suhm, W. Ward, and A. Waibel. 1995. Concept-based parsing for speech translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 196–187, Leuven, Belgium, July.

J. M. Vilar, A. Castellanos, J. M Jimenez, J. A. Sanchez, E. Vidal, J. Oncina, and H. Rulot. 1996. Spoken-language machine translation in limited domains: Can it be achieved by finite-state models? In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 326–333, Leuven, Belgium, July.