Improving the Speech Recognition Performance of Beginners in Spoken Conversational Interaction for Language Learning

Hui Ye and Steve Young

Cambridge University Engineering Department
Trumpington Street, Cambridge, England, CB2 1PZ
hy216@eng.cam.ac.uk, sjy@eng.cam.ac.uk

Abstract

The provision of automatic systems that can provide conversational practice for beginners would make a valuable addition to existing aids for foreign language teaching. To achieve this goal, the SCILL (Spoken Conversational Interaction for Language Learning) project is developing a spoken dialogue system that is capable of maintaining interactive dialogues with non-native students in the target language. However, the effective realisation of the intelligent language understanding and dialogue management needed for such a system, requires robust recognition of poorly articulated non-native speech. This paper studies several popular techniques for robust acoustic modelling including HLDA, MAP and CMLLR on non-native speech data within a specific dialogue domain. In addition, a novel approach for using cross language speech data to adapt the acoustic models is described and shown to be useful when very limited non-native adaptation data is available. The experimental results provide a clear story of how to improve recognition performance on non-native speech for a specific task, and this will be of interest more generally for those developing multi-lingual spoken dialogue systems.

1. Introduction

Interest in using computers to assist language learning has grown over the past decade driven by increasing demand for foreign language teaching. The SCILL (Spoken Conversational Interaction for Language Learning) project is a collaboration between MIT CSAIL Spoken Language Systems Group and the Cambridge Machine Intelligence Laboratory. The aim of the project is to integrate state-of-the-art technologies in speech recognition, natural language processing, understanding, translation, dialogue management and text-to-speech (TTS) to provide an intelligent system to assist language learning[1, 2]. This system will have three major components:

1. an intelligent agent capable of maintaining an interactive dialogue with the student in the target language. This agent would act as the "native speaker" versed on some topic selected from a pre-defined set.
2. a second intelligent agent capable of translating within the topic domain between the student's native language and the target language. This agent will act as an on-line "tutor" giving advice on how to say required words and phrases.
3. an assessment component which provides a post-mortem analysis of the conversation and gives feedback on errors and areas to improve.

In the prototype SCILL system, the target language is Mandarin Chinese, the native language is US or UK English and the conversational domain is "the weather". The system will enable a student to participate in a dialogue with the system in Mandarin whilst simultaneously having access to a "tutor" that can tell them how to say certain phrases. For example, the topic might be about the weather in a particular city, and the bilingual "tutor" could provide the student with helpful hints on how to ask questions about the weather tomorrow and the day after. This arrangement allows the student to develop conversational skills at their own pace in a non-threatening environment. The prototype system is based on MIT's Galaxy system [3] and integrated with Cambridge's HTK/ATK speech recognition engine[4]. Although the system provides acceptable performance for native Mandarin speakers, its performance degrades badly on non-native beginners due to greatly reduced recognition accuracy. Therefore, improved performance on non-native speech is needed before the system can be put to practical use.

Previous research on non-native speech recognition has identified a number of useful techniques for improving robustness [5]. Inspired by the fact that non-native speakers tend to pronounce words differently from native speakers, Livescu and Glass proposed a lexical modelling technique [6] to improve non-native speech recognition. Similarly, Amdal and Korkmazskiy used joint pronunciation modelling to incorporate non-native pronunciations into the lexicon [7]. Both methods used data driven techniques to derive a multiple pronunciation lexicon, but the result was a relatively modest reduction in word error rate.

Based on the fact that native and non-native speech have quite different acoustic spaces, the most straightforward method to improve non-native speech recognition would be to use non-native acoustic models trained directly on non-native speech. However, to obtain sufficient non-native training data is very difficult, especially when the target group are early stage foreign language learners who have difficulty in reading prompts and who often feel inhibited when speaking aloud. When target data is limited, speaker adaptation techniques such as MAP [8] and MLLR [9] can be used to adapt acoustic models trained on native data. Tomokiyo and Waibel have studied adaptation methods for non-native speech [10] and found that substantial gains can be obtained. Wang et al. compared the effectiveness of several adaptation techniques on non-native speech [11], and consistent improvements were confirmed. In addition to using adaptation techniques to refine the acoustic models, other acoustic modelling approaches have been proposed to improve non-native speech recognition including model combination [12] and model interpolation [11].

Whilst significant gains have been obtained, the majority of the existing work on non-native speech recognition has been targetted at relatively fluent speakers for which reasonable amounts of adaptation data could be obtained. As mentioned above, however, data collection for early stage language learners is difficult. We are therefore interested in techniques which can be used to improve robustness for non-native speakers even when very limited or even no adaptation data is available.

In this paper, we firstly assess the use of current techniques for robust speech recognition and adaptation in this particularly difficult application domain. We then describe a cross language adaptation technique which uses enrolment data in the speaker's native language to adapt the target language models. The results demonstrate that a combination of techniques can significantly improve performance. For example, in the experiments reported here on early stage learners of Mandarin Chinese, the character error rates were reduced by more than 50%.

The remainder of the paper is organised as follows. In section 2, an evaluation database of bilingual speech collected in the weather domain is described. Then in section 3, the baseline bilingual speech recognition system is presented together with its performance on non-native speech. In section 4, a number of model refinement techniques are introduced and evaluated. Then in section 5, a cross-lingual adaptation technique is described which can be used when there is no target language adaptation data available. Finally conclusions are presented in section 6.

2. Bilingual Corpus

A bilingual weather corpus was created for the SCILL project. The corpus contains speech data from 40 speakers (20 native Chinese and 20 native English), of which 15 are female and 25 are male. Each speaker was assigned a different set of 80 sentences to read in the weather domain of which half are in English and half are in Chinese. The organisation of the sentence sets is as follows:

1-10	pseudo-utterances in English
11-20	pseudo-utterances in Chinese
21-30	natural utterances in English
31-40	natural utterances in Chinese
41-50	utterances with mixed Chinese and English
51-65	natural utterances in English
66-80	natural utterances in Chinese

Sentences 1-50 are provided for testing and sentences 51-80 are provided for adaptation. All speech was read from prompts. The "pseudo-utterances" denote prompts generated by a handcrafted finite state grammar and the natural utterances denote prompts extracted from real conversations. Although not covered in this paper, this separation allows a comparison to be made between grammar-constrained recognition and N-gram based recognition.

Since the target application environment will be extremely variable in terms of microphone, speaker and background noise, the corpus was collected under realistic conditions such as in classrooms during language classes. It thus covers variations in background noise, microphones, volumes, ages of speakers (from 14 to 65), and speaker fluency. The speech data is recorded in 16k 16bits mono format.

The main purpose of this paper is to evaluate the speech recognition performance of native-English speakers speaking Chinese, hence the bilingual weather corpus has been partitioned into six sets:

Native: Chinese sentences 11-20 and 31-40 spoken by all 20 Chinese speakers, 400 utterances in total.

NonNativeAB: Chinese sentences 11-20 and 31-40 spoken by all 20 English speakers, 400 utterances in total.

NonNativeA: Chinese sentences 11-20 and 31-40 spoken by the first 10 native English speakers, 200 utterances in total.

AdaptA: Chinese sentences 66-80 spoken by the first 10 native English speakers, 150 utterances in total.

AdaptB: Chinese sentences 66-80 spoken by the last 10 native English speakers, 150 utterances in total.

EnrolA: for each speaker in NonNativeA, English sentences 51-65 are used as cross-language enrolment data.

Note that in the above naming convention, the first 10 English speakers are denoted by A, and the second 10 English speakers by B.

3. Baseline System

The speech recognition engine used in our system is ATK 1.4 [4]. ATK is an application toolkit for HTK which allows real-time recognisers built using HTK derived models. The acoustic model set and language model set used in ATK are entirely compatible with those trained by HTK. Moreover, ATK supports multiple recognisers running simultaneously. Thus in the SCILL system, bilingual recognition can be implemented with two ATK recognisers, one in English and the other in Mandarin Chinese. However, in this paper, we are only concerned with the Mandarin recogniser.

The Mandarin acoustic model in our baseline system consists of a word internal triphone hmm set with the standard 39 dimensional MFCC features and cepstral mean normalization (CMN), and 4 mixture components for each tied state. The model is trained on Microsoft's Mandarin speech toolbox corpus [13]. This corpus contains read speech from 100 male speakers for a total of 19,688 utterances. The baseline language model is a domain specific word class language model interpolated with a standard bigram model trained from a text corpus in weather domain. The vocabulary size is around 1000. Testing this baseline system results in a 13.45% character error rate (CER) for native speakers (Native), and 40.22% for non-native speakers (NonNativeAB). Clearly there is a large gap in CER between native and non-native speakers, and indeed, for most practical purposes the non-native performance is useless. The next section discusses a number of ways in which this performance gap can be reduced.

4. Model Refinement and Adaptation

4.1. Choice of front end

The effectiveness of different choices of front end was investigated by comparing MFCC and PLP features, and the use of a HLDA transform [14]. As for the MFCC front-end, the PLP front end includes 13 PLP coefficients and their delta and acceleration coefficients to give in total 39 coefficients per frame. The HLDA front-end, applies a 52×39 dimensional HLDA transform to the standard 39 PLP features augmented with 13 extra tertiary coefficients. Table 1 shows the recognition results in terms of CER. As can be seen, the MFCC front end has

similar performance when compared to PLP, but the use of the PLP+HLDA transform reduces the CER by around 2% absolute.

Table 1: *Recognition results with different choices of front end*

CER (%)	MFCC	PLP	PLP+HLDA
Native	13.45	14.01	13.03
NonNativeAB	40.22	40.07	38.06

4.2. Word internal vs cross-word models

Another aspect concerning model complexity that we have investigated is the comparison between word internal (WI) triphones and cross word (XW) triphones. Table 2 indicates that XW triphones out-perform WI triphone on native speech data, but degrade on non-native speech data. The reason for this is probably because the fluency of the non-native speakers is poor with many short pauses between words. Hence cross word effects are small and the silence context assumed at word ends by word internal models dominates. Therefore, even though cross-word triphones are commonly used in most speech recognition systems, word-internal triphones are more appropriate for this application and are used in all further experiments reported below.

Table 2: *Recognition results using WI and XW triphones. Both model sets use PLP+HLDA front-end*

CER (%)	Word Internal (WI)	Cross Word (XW)
Native	13.03	10.87
NonNativeAB	38.06	41.16

4.3. Number of mixture components

It is well known that enhancing the model complexity by increasing the number of mixture components of the hmm set will lead to steady improvement on recognition accuracy, as long as there is sufficient training data to prevent overtraining. Table 3 shows a steady CER reduction on native speech when increasing the component number from 4 to 8 per state, however, on the non-native speech a small degradation is observed. This may be because "sharpening" the acoustic models on native training data moves them further away from the non-native speakers. However, since the degradation on non-native speech is very small, only 0.52%, and the improvement on native speech is more significant, about 1.6%, 8 mixture components per state are used in our system.

Table 3: *Recognition results over different number of mixture components.*

CER (%)	4mix	6mix	8mix
Native	13.03	11.77	11.48
NonNativeAB	38.06	39.51	38.58

4.4. Adaptation using CMLLR and MAP

Constrained MLLR [9] and MAP [8] have been widely used for speaker adaptation and have proved to be very effective. In the SCILL application domain, adaptation data in the target language is not available for individual speakers because it is usually too difficult for them to produce. However, adaptation can be used to transform a "native" model set into a model set tuned for non-native speakers.

Table 4 presents the recognition results of using CMLLR and MAP adaptation to transform the native speaker models to adapted non-native speaker models. The baseline native model is the 8 mixture PLP+HLDA word internal triphone set. With 150 utterances in the non-overlapping adaptation set AdaptB, 41 CMLLR transforms were estimated. The results show that using a pool of non-native adaptation data collected in advance, can substantially improve the performance on new non-native speakers. Furthermore, it is clear that CMLLR is more effective than MAP giving a 19% absolute reduction in CER compared to 13% absolute reduction for MAP. Finally, note that adapting the model set towards non-native speakers does not seriously impact on the performance for native speakers.

Table 4: *Recognition results of CMLLR and MAP adaptation using the AdaptB data set.*

CER (%)	Baseline	CMLLR	MAP
Native	11.48	11.60	13.89
NonNativeA	39.86	20.66	26.83

5. Cross language speaker adaptation

In the last section, adaptation of native Mandarin Chinese models towards a non-native speaker set improved recognition for new non-natives outside of the adaptation set. Speaker dependent adaptation might be expected to provide further improvement, however as noted above, it is very difficult in practice for early stage learners to provide enrolment data in a new language for which they have great difficulty articulating. They would of course be easily able to produce enrolment data in their own native language.

When someone is learning a foreign language, they tend to use their native phoneme set to pronounce words in the foreign language, and this motivates the idea of cross language adaptation, which in our case is to use English data to adapt Mandarin Chinese acoustic models. To do this, we first constructed a mapping by-hand between the phoneme sets of the two languages. Since the two languages are very different, the mapping is not necessarily one to one, neither is it a full mapping that finds counterparts in Chinese for all the phonemes in English. A fragment of this mapping is shown in Table 5. Then for each non-native speaker in test set NonNativeA, we use his/her corresponding 15 English utterances in EnrolA to conduct cross language adaptation as follows.

For each English enrolment utterance:

1. force align the utterance using an English acoustic model set so that the phoneme boundaries are marked.
2. for each English phone x spanning segment t_1 to t_2, if there exists a phoneme mapping in Table 5 $x \rightarrow y$, use Viterbi recognition to find which of all the Mandarin triphones with base phone y matches the acoustic segment t_1 to t_2 the best.

On conclusion of this process, all phone segments of the enrolment data which have entries in the phone mapping table will have an associated Mandarin triphone model. Treating these

Table 5: *Phoneme mapping from English to Chinese.*

English	Chinese	English	Chinese
ay	ai	aa	a
ao	o	aw	ao
ax	e	ah	a
b	b	d	d
ey	ei	f	f
er	er	g	g
hh	h	ih	i
iy	i	k	k
l	l	m	m
n	n	ow	ou
uw	ou	p	p
t	t	s	s
w	w	y	y
...	...	...	...

segments and their associated models as adaptation data, CMLLR adaptation can be performed. Due to the limited amount of data, this adaptation is limited to a single global transform.

In Table 6, the result of cross language speaker adaptation is shown. Even though there is a relatively small amount of data (15 utterances) and the data is in the wrong language, a 5.8% absolute reduction in CER is obtained. This shows that cross language adaptation is feasible and effective when within language adaptation data is unavailable. In addition, if we pool the cross language data "EnrolA" together with the within language data "AdaptB" to carry out CMLLR adaptation, further improvement (19.38% CER) is obtained. This result is comparable with the conventional speaker adaptation result (19.72%) which uses speaker dependent data in "AdaptA" to do CMLLR adaptation for each speaker.

Table 6: *Recognition results of cross language adaptation.*

CER (%)	baseline	EnrolA	+AdaptB	AdaptA
NonNativeA	39.86	34.07	19.38	19.72

6. Conclusions

In the framework of developing a practical spoken dialogue system for language learning, this paper has focused on solving one of the main obstacles - robust recognition of the speech of early stage language learners. Various acoustic modelling and adaptation techniques have been investigated to optimise speech recognition performance. These experiments showed that PLP+HLDA provides the most robust front-end and CMLLR is a very effective adaptation technique for making native acoustic models more robust to non-natives. It was also shown that the use of cross-word triphones is not appropriate for this type of data.

Experience has shown that early stage learners of Mandarin find it very difficult to provide enrolment data, hence conventional speaker enrolment is not possible. However, it is possible to collect native English enrolment data and by using the proposed cross-language adaptation procedure worthwhile gains can be obtained.

Overall, with the techniques implemented, the final character error rate on non-native speakers has been halved from 40% to 19.4%. Although this is still higher than the comparable native speaker performance of 11.4%, by suitably constraining the language model, we believe that this will nevertheless allow us to build and deply a practical dialogue system which can provide useful conversational practice for early stage language learners.

7. Acknowledgements

This work is supported by a grant from the Cambridge-MIT Institute (CMI). The authors thank the volunteers for their participation and contribution to the bilingual speech database collection.

8. References

[1] S. Seneff, C. Wang, and J. Zhang, "Spoken Conversational Interaction for Language Learning", Proc InStil Workshop, Venice, June, 2004

[2] S. Seneff, C. Wang, M. Peabody, and V. Zue, "Second Language Acquisition through Human Computer Dialogue"m Proc ISCSLP 04, Hong Kong, December, 2004.

[3] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "GALAXY-II: A Reference Architecture for Conversational System Development", Proc. ICSLP 98, Sydney, Australia, November 1998.

[4] Young, S. "ATK - An Application Toolkit for HTK", Cambridge University Engineering Department, 2004.

[5] L.M. Tomokiyo, "Handling Non-native Speech in LVCSR: A Preliminary Study", Proc. Incorporating Speech Technology in Language Learning, 2000.

[6] Livescu, K. and Glass, J. "Lexical modeling of non-native speech for automatic speech recognition", Proc. ICASSP 2000.

[7] I. Amdal, F. Korkmazskiy and A.C. Surendran, "Joint pronunciation modeling of non-native speakers using data-driven methods", Proc. ICSLP00, Beijing, China, 2000.

[8] Gauvain, J.-L. and C.-H. Lee, "Maximum a Posteriori Estimation of Multivariate Gaussian Mixture Observations of Markov Chains." IEEE Trans Speech and Audio Processing 2(2): 291-298, 1994.

[9] Gales, M.J.F., "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition". Computer Speech and Language, vol. 12, 1998 .

[10] Tomokiyo, L.-M. and Waibel, A. "Adaptation methods for non-native speech", Proceedings of Multilinguality in Spoken Language Processing, 2001.

[11] Z. Wang, T. Schultz and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech", Proc. ICASSP 2003.

[12] Witt, S. and Young, S. "Bilingual Model Combination for Non-Native Speech Recognition." Proc Institute of Acoustics Conf Speech and Hearing, Windermere, England, 1998.

[13] Chang, E., Shi, Y., Zhou, J. and Huang, C. "Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research," Eurospeech 2001, Aalborg, Denmark, 2001.

[14] Kumar, N. and A. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition." Speech Communication 26: 283-297, 1998