ICICS-PCM 2003
15-18 December 2003
Singapore

A Bilingual Speech Recognition System for English and Tamil

C. Santhosh Kumar [1] ; Foo Say Wei [2]
[1]Amrita Vishwa Vidyapeetham, Ettimadai, Coimbatore, India
[2]Nanyang Technological University, Singapore

Abstract

This paper describes the details of a bilingual speech recognition system, *AmritaRec*, developed for English and Tamil. The performance results of the system is compared with that of a monolingual English speech recognition system adapted to Tamil using cross language transfer and cross language adaptation techniques.

1. Introduction

Over the last few decades speech recognition has evolved and matured enough to be used in commercial applications. The applications include automatic dictation software and automatic call routing to automatic transcription of the TV news.

In India, with more than 30 languages spoken across the country, it is essential to enable speech recognizers to work in a multilingual environment. A simple but practical approach is to have many monolingual systems, and select the appropriate acoustic model with the help of a language identifier[7]. This approach, however, has the disadvantage that the speech data for each of these languages under development need to be collected. The data collection process needs enormous resources, both human and financial. Also, it is not practical to collect speech data for all the languages of interest, some of these languages are spoken by less than one million people, that too dispersed all over the country (Eg. Tulu, Kongini).

In this context, other approaches need to be explored to enable speech recognizer work in a multilingual environment, by combining the acoustic models of the respective languages to make a language independent speech recognition engine or by fast adaptation to the new language with minimum amount of speech data [13].

We, in this paper, followed two approaches.

1. *Monolingual system for English:* Trained the acoustic models using the English speech (NTIMIT corpora). We used cross language transfer and cross language adaptation[14]techniques on the model to be used as a Tamil speech recognizer.
2. *Bilingual system for English and Tamil:* Trained the acoustic models for English and Tamil separately; combined the acoustic models using decision tree clustering [4] to generate the bilingual speech recognizer. The models were then adapted to Tamil or English as needed using MLLR[5] and MAP[8] adaptation techniques.

The NTIMIT corpora for English collected over telephone contained 4620 sentences for training, however our Tamil corpora contained only 400 utterances for training. With this limited Tamil speech data, it was not possible to train a triphone based speech recognition system. So, this effort to make a bilingual system to cater for both English and Tamil. We used 200 sentences each from English and Tamil databases for testing and this data set is different from the data set used for training.

2. Training the models

American English and Tamil are two languages that has very little in common. While English being a stressed language, Tamil is not. On the other hand, the number of variants of the vowels is much smaller in Tamil, while English needs a larger number of vowels for acoustic modeling. Further, Tamil has a larger number of non-vowel sounds compared to English. Yet, we see that there are acoustic similarities between many sounds in these two languages.

Table 1 shows the complete list of American English phones as well as the additional phones needed for Tamil. We have not used the stress information in English and therefore the stress markers were removed for its use in our system. For our experiments with the bilingual system, we used a super set of the monophones of the two languages.

In the first set of experiments, we used only the NTIMIT database to train the models. A triphone based system using decision tree clustering technique was developed. This system was able to recognize unseen triphones by synthesing the acoustic models from the existing set of models using the decision trees generated during the decision tree clustering. The questions used for clustering of the states are listed in Table.2. The complete triphone based system after training was tested for NTIMIT, and also for Tamil with and without adaptation[8].

	English	Additional phones for Tamil
Vowels	iy ih eh ey ae aa aw ay ah ao oy ow uh uw ux er ax ix axr	ii ee A uu U oo
Stops	b d g p t k	P tt K
Fricative	s sh z zh f th v dh m n ng em en	nn N ny
Semivowel and glides	el hh l r w y	L lzh R
Affricates	jh ch	

Table 1: Monophones used in American English and Tamil

In the second set of experiments, we trained two separate triphone based systems with one gaussian per state for all the triphones seen in the training set. The same phone spoken in two different languages could be acoustically different even if the IPA symbol representing them is the same. So, to distiguish between the triphones in the two languages, we added the language information with every triphone name. For example, *ai-p+eh?en* represents a triphone *p* in the language *en*(English) with *ai* as its left context, *eh* as its right context. Thus, *ai-p+eh?ta* would represent the same triphone for *ta*(Tamil). They were then merged together to form a single acoustic model database covering all the triphone models of the two languages.

After combining the acoustic models of English and Tamil, the states are clustered in a way to minimize the loss of entropy due to the clustering of similar states by the decision tree clustering[3] techniques. In decision tree clustering, a language, a left context or a right context, related question is applied and the one which results in the minimum decrease of entropy when merged is chosen for merging at that step. The set of these states are combined to form a single physical state. This process is continued until the loss of entropy by any further merge goes above a chosen threshold. Next, a list of triphones unseen in the training set but occur in the test set are synthesised by selecting an appropritate state in the decision tree generated during the clustering . This makes the acoustic model complete with all possible triphone models. In the combined acoustic models after the decision tree clustering, we observe that many of the states in Tamil and English were shared by the same physical state indicating that there are some similarities even in two diverse languages such as American English and Tamil.

Stop	b d g p t k P tt K
Nasal	em en m n ng nn N ny
Fricative	ch dh f jh s sh th v z zh nn N ny
Liquid	el hh l r w y L lzh R
Vowel	aa ae ah ao aw ax axr ay eh er ey ih ix iy ow oy uh uw ii ee A uu U oo
Front	ae b eh em f ih ix iy m N p v w ii ee
Central	ah ao axr d dh el en er l n ny r s t th z zh A
Back	aa ax ch g hh jh k ng ow sh uh uw y uu U oo
Front Vowel	ae eh ih ix iy ii ee
Cental Vowel	aa ah ao axr er A
Back Vowel	ax ow uh uw uu oo U
Long Vowel	ao aw el em en iy ow uw uu oo U
Short Vowel	aa ah ax ay eh ey ih ix oy uh U
Diphthong	aw axr ay el em en er ey oy
Front Start	aw axr er ey A ee
Fronting Vowel	ay ey oy
High Vowel	ih ix iy uh uw ii uu U
Medium Vowel	ae ah ax axr eh el em en er ey ow ee oo
Low Vowel	aa ae ah ao aw ay oy A
Rounded	ao ow oy uh uw w uu oo
Unrounded	aa ae ah aw ax axr ay eh el em en er ey hh ih ix iy l r y R L lzh ii ee A
Reduced	ax axr ix
IVowel	ih ix iy ii
AVowel	aa ae aw axr ay er A
OVowel	ao ow oy oo
UVowel	ah ax el em en uh uw uu
Language	Tamil

Table 2: Questions to form state clusters

3. Results

In the first set of experiments, we used the monlingual system trained using the NTIMIT corpora. The acoustic models were tested on NTIMIT and Tamil speech. For Tamil we tested the models with and without adaptation. The adaptation was done in two steps, first MLLR and then MAP adaptation. We have noticed that adapting the models in this order resulted in a better recognition accuracy. The results are tabulated in Table. 3. As expected, the performance of this system without adaptation on Tamil speech data was very poor, while the adaptation has improved the performance significantly. The Tamil lexicon was adapted to American English phone set using their linguistic similarities. It may be noted that in some cases more than one Tamil phone had to be mapped to a single English phone. For example, *l, lzh, L* were mapped to *l* leading to inacuu-

racy in the modeling of these sounds.

In the second set of experiments, the bilingual speech recognition system trained using the combined NTIMIT and the limited Tamil speech corpora was evaluated for NTIMIT and Tamil corpora separately. We observe that due to the wide difference in the speech signal characteristics of these two languages, the acoustic models tried to become general enough to cater for both the languages. As a result, individual recognition performance of the system on NTIMIT data was not as high as the monolingual system. However, the recognition accuracy for Tamil has increased substantially in the bilingual system.

In these experiments, we mapped the Tamil lexicon using the American English phonetic symbol set to the extent possible based on their lingistic similarities and IPA symbol. In addition we have included a set of phonetic symbols to cater for Tamil. Eg. *lzh, L, T*. These phones are not present in American English. Thus, in the bilingual modeling approach, we have a better acoustic model. Even sounds sharing the same symbol across the two languages were treated differently as the language name was tagged to the triphone name. In this case the algorithm was allowed to merge the closest phone in the two languages within the contraints of the decision questions listed in Table. 2 subject to minimum loss of entropy. Table 4 presents the test results of the system for NTIMIT and Tamil speech corpora separately with and without adaptation.

It may be noted that in all the tests above we used a context free grammar without any language model as the aim of this experiment was only to compare the effectiveness of the acoustic modeling in a multilingual system compared with the monolingual system adapted using cross language transfer and cross language adaptation techniques.

	% accuracy
NTIMIT	76.55
Tamil	19.92
Tamil with adaptation	55.57

Table 3: Word recognition accuracy for English monolingual system

	% accuracy
NTIMIT w/o adaptation	58.62
NTIMIT with adaptation	66.93
Tamil w/o adaptation	61.61
Tamil with adaptation	64.42

Table 4: Word recognition accuracy for the bilingual system

4. Conclusions

In this paper, we investigated the effect of sharing the acoustic models across two languages for effectively modeling the acoustic space of these languages, without having to model each of these languages separately. Though we have used two languages,American English and Tamil, that has very little similarity, the experiments demonstrate that the acoustic modeling can be done efficiently for more than one language. This has the effect of reducing the computational cost on the search engine as we need to use only one acoustic model for many languages.

Encouraged by the initial results, we are currently developing a system to cater for Indian accented English, Tamil, and Hindi. The work in this direction is in progress and the results will be reported in due course of time.

Further, mapping of the monophones across the two languages was done manually in this set of experiments. In a truly multi-lingual setup, when we are to handle many languages, this could be quite tedious and time consuming. We are also working on an automatic clustering algorithm to group the similar monophones across languages.

5. Acknowledgements

The first author would like to thank Mr. Udhay Kumar, final year B.E student, for helping with the Tamil speech data preparation and verification; Dr. Harini Jayaraman for her help in preparing the Tamil lexicon.

References

[1] X. Huang, et al, Spoken Language Processing, Prentice Hall PTR, NJ, 2001

[2] L.R. Rabiner, et al, Fundamentals of Speech Recognition, Prentice Hall Inc., 1993.

[3] S.J. Young, et al, "Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems", Cambridge University Technical Report.

[4] J.J.Odell, "The Use of Context in Large Vocabulary Speech Recognition", Ph.D Thesis, Cambridge University Engineering Dept., 1995.

[5] C.J.Legetter, et al, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, pp. 171-185, 1995.

[6] M.J.F.Gales, et al, "Mean and Variance Adaptation with the MLLR Framework", Computer Speech and Language, pp. 249-264, 1996.

[7] J. Navaratil, "'Spoken Language Recognition-A Step Toward Multilinguality in Speech Processing"', IEEE Trans. on Speech and Audio Processing, Sept. 2001, pp 678-685.

[8] J.L.Gauvain, et al, "Maximum a Posteriori Estimation of Multivariate Gaussians Mixture Observations of Markov Chains", IEEE Trans. on Speech and Audio Processing, pp. 291-298, 1994.

[9] Bojan Imperl, et al, "Agglomerative vs. Tree-based Clustering for the Definition of Multilingual Triphones", Proc. EuroSpeech, Budapest, Hungary, 1999

[10] T. Schultz, et al, "Language Independent and Language Adaptive Large Vocabulary Speech Recognition", Proc. EuroSpeech, Budapest, Hungary, 1999.

[11] P. Bonaventura, et al, "Multilingual Speech Recognition for Flexible Vocabularies", Proc. EuroSpeech, Greece, 1997.

[12] J. Billa, et al, "Multilingual Speech Recognition: The 1996 BYBLOS Callhome System", Proc. EuroSpeech, Greece, 1997.

[13] A. Contantinescu, et al, "On Cross-language Experiments and Data-driven Units for ALISP", Proc. Automatic Speech Recognition and Understanding, Santa Barbara, 1997.

[14] A. Waibel, "'Multilinguality in Speech and Language Systems"', Proc. of the IEEE, Aug. 2000, pp 1297-1311.