SPEAKER INDEPENDENT ACOUSTIC MODELING FOR LARGE VOCABULARY BI-LINGUAL TAIWANESE/MANDARIN CONTINUOUS SPEECH RECOGNITION

Dau-Cheng Lyu [1,2] , Bo-hou Yang[1,2], Min-Siong Liang[1],
Ren-Yuan Lyu [1], Chun-Nan Hsu[2]

[1] Dept. of Electrical Engineering, Chang Gung University, Taoyuan 333, Taiwan
[2] Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
E-mail: rylyu@mail.cgu.edu.tw, Tel: 886-3-3283016ext5967

ABSTRACT: In this paper, we describe the acoustic modelling technique for a bi-lingual Taiwanese /Mandarin speech recognition system, which deals with speaker independent continuous speech based on HMMs clustered by an acoustic phonetic decision tree. A bi-lingual recogniser with a bi-lingual database of 120 people was built. The vocabulary size of this system is up to 40 thousands. Unigram, bi-gram, and tree lexicon language models have been used. In order to share the common part of the two languages, a decision-tree based clustering is adopted in inter-syllable triphone units. A 89.8% word accuracy is achieved by searching on a tree lexicon net under a context free grammar.

1. Introduction

Multi-lingual speech recognition has been widely studied in the field of speech recognition during the past years (Waibel, 2000) (Tanja, 1999). We chose Taiwanese Min-nan and Mandarin Chinese as two languages to be studied because of Mandarin Chinese is the most important language in Chinese societies, and Taiwanese Min-nan, one of the major Chinese dialects, is the mother tongue of more than 75% of the population in Taiwan. Most people in Taiwan speak at least 2 languages in their daily lives, i.e., Mandarin and Taiwanese. To make the future speech interface more friendly, the bi-lingual or even multi-lingual capability of the speech recognizer is highly desired.

In this paper, we build the system in a number of stages. First, we adopt the continuous Hidden Markov Model (CHMM) to setup a monolingual system for Mandarin and Taiwanese. Based on the transcription of the two languages into a phonetic set of symbols called Taiwan Phonetic Alphabet (TWbet), we make these two languages share the speech data with the same symbol, and then establish a baseline bi-lingual system. To obtain good recognition performance, inter-syllabic co-articulation factor is considered. However the majority of triphone contexts have only very few occurrences in the training data, hence there is not enough data for robust parameter estimation of these rarely seen triphones. In order to solve these problems, a decision tree clustering technique was used, which took advantage of acoustic/phonological knowledge and provides an effect approach of data sharing. We chose clustering in the sub-phonetic level, i.e., states in the phonetic HMM models, in which each specific state in each model has its own decision tree. A set of 126 yes/no phonological questions were constructed for each decision tree.

In language model issues, we use three types of searching nets which are a bilingual context free grammar net, a bi-gram net, and word tree net. The perplexities are also computed for all different language models for comparison.

This paper is organized as follows: section 2 introduces our bi-lingual speech database and phonetic inventory; section 3 describes the first step of speech recognition monolingual system; section 4 describes how to generate the phonetic decision tree for inter-syllable triphone units; section 5 reports the experimental results; the last section is a brief conclusion.

2. Phonetic Inventory and Database

In this research, we use a single phonetic set to cover all the sounds of Taiwanese/Mandarin speech. The phonetic transcription system called Taiwan Phonetic Alphabet (TWbet) is designed to cover all languages spoken in Taiwan. Sounds which are represented by the same TYPA symbol share one common phoneme category. We have a set of 429 base syllables for Mandarin and a set of 825 base syllables for Taiwanese. There are 183 common base syllables for both languages, and thus a set of 1049 distinct base syllables are obtained when considering both languages simultaneously. Similarly, there are 37 phones for Mandarin and 52 phones for Taiwanese. A set of 63 phones are necessary to transcribe both languages. Despite their similarity in basic syllable structure, Mandarin and Taiwanese have fairly different phonological rules and phonetic composition. For example, in Taiwanese, there are additional Nasalized Vowels and Stop Vowels, which do not exist in Mandarin. For such phones, the postfixes -P, -T,-K, -H are tailed in normal vowels, e.g., “A”, “aP”, “aT”, “aK” and “aH” (*Lyu, 2000*).

Table 1. The phone set of Mandarin and Taiwanese. Phones inside the shadowed area are common phones in both languages. The upper unshadowed part is belong to Mandarin only, and the lower unshowed part is for Taiwanese only.(modified TYPA)

C	GH	J	NH	rr	R	S	Y	f	y
yH		G	a	aH	b	c	d	e	g
h	i	iH	k	l	m	n	o	oH	p
r	rH	s	t	u	j	uH	N		A
AH	AP	E	EH	I	IH	M	v	O	OH
Q	z	aK	aP	aT	U	eH	eT	iK	iP
iT	oK	oP	q	uT					

A bi-lingual speech database produced by 60 male and 60 female speakers over 16k 16 bits microphone was provided in Multi-median Signal Processing Laboratory of Chang Gung University in Taiwan. All the statistics of the corpus considered here are listed in <table.2>. The training data composes of four subsets, including 11 hours of Taiwanese and 11.3 hours of Mandarin speech, uttered by 100 speakers. Totally include 92160 utterances. For testing, we chose another 20 speakers, with 34 minutes of speech of 2000 words out of a vocabulary of 40 thousands words

Table 2. Training and test data for Mandarin and Taiwanese. Set1, Set3 are tonal syllable; Set2, Set4 are words (1~5 syllables)

	Training data (100 speakers)				Test data (20 speakers)	
	Mandarin		Taiwanese		M (10)	T (10)
	Set 1	Set 2	Set 3	Set4		
Total # of syllables	1287	7790	2883	6064	2468	2547
Total # of phrases	1728	3157	2883	1913	1000	1000
Average word length	1	2.5	1	3.2	2.5	2.5
Total duration (in hours)	2.3	9.0	5.2	5.8	17 min	17 min

.

3. Baseline System

We constructed a baseline recogniser in monolingual and bilingual mode. We adopted one of the most popular speech recognition schemes based on the context dependent phonetic CHMM with a

39-dimentional feature vector composed of the following elements: 12-dimentional MFCCs, 12-dimentional delta MFCC, 12-dimentional delta delta MFCC, 1-dimentional energy, 1-dimensional delta energy, and 1-dimentional delta delta energy. Each triphone and biphone has 3 states, and 8 Gaussian mixtures. The searching space is constrained by a context free grammar that uses base-syllables as the nodes. For the purpose of comparison, the evaluation of the perplexity for each language was performed. They are 429 825, and 1049 for Mandarin, Taiwanese and the combined Bi-lingual set of 2 languages, respectively.

The results are shown in <table.3>. It shows the accuracy by using Right-Context-Dependent (RCD) phone HMMs and Tri-phone HMMs for Mandarin, Taiwanese, and Bi-lingual speech to recognize the testing speech, constrained in different language domain. We can see the Tri-phone HMMs are better than RCD HMMs, and the bi-lingual system is also better than monolingual under the same perplexity. In <Table 3>, the row 2 numbers are the number of each model.

Table 3. The accuracy rate of Mandarin syllable net (perplexity=429), Taiwanese syllable net (perplexity=825) and Bi-lingual speech (perplexity=1049), by using RCD-phone HMM and Tri-phone HMMs.

Models Test data & perp.	RCD-phone HMMs			Tri-phone HMMs		
	For Mandarin (178)	For Taiwanese (374)	For Bi-lingual (452)	For Mandarin (558)	For Taiwanese (1005)	For Bi-lingual (1280)
Mandarin (429)	51.98%		60.18%	61.57%		64.50%
Taiwanese (825)		54.17%	58.69%		58.18%	62.69%
Bi-lingual (1049)			50.49%			54.70%

4. State Clustering by Acoustic Decision Tree

Context dependent acoustic model based on decision tree clustering has been widely adopted in LVCSR application (Sheng, 2000) (Liang, 1998). In those papers we use phonetic decision tree state clustering method to solve two problems: one is that the large number of inter-syllable context dependent models are trained limited by amount of training data, another is that data should be shared across languages. A phonetic decision tree is a binary tree in which a yes/no question about phonetic context is attached to each node, which is then split with respect to the answer. In general, there are a number of left and/or right phonetic questions that could be asked at each parent node. After evaluation function, chose the “best” question for splitting. It will be stoped until some values satisfy the criterion, then the states in the same node will share data. Therefore, there are three issues in decision tree technology as described in the following sub-sections.

4.1 Evaluation function

A proper criterion to decide the "best" question for splitting a tree node is an evaluation function as follows:

$$L_s = \sum_{m \in S} N_{mi} \mid \vec{\mu}_{mi} - \vec{c} \mid^2$$

where N_{mi} is the occupancy of state i of model m, $\vec{\mu}_{mi}$ is the mean vector of state i of a model m in the tree node S, and $\vec{c}$ is the centre vector of all $\vec{\mu}_{mi}$ in node S. L_s is then the measure of the dissimilar of all models in the tree node S. When the parent node S splits into 2 children nodes S_1 and S_2, we want to maximize

$$\Delta L = L_s - (L_{s_1} + L_{s_2}),$$

that is, to increase the similarity of models inside a node.

A threshold of ΔL is set by experiments to stop the process of splitting the tree. After the tree is constructed, states within each leaf node are tied together and the key states of these models share the same training data. Figure 1 shows the clustering examples for a decision tree.

4.2 The Question Set

By taking many possible factors into consideration, we designed 4 sets of linguistic questions to be asked to make judgments about right/left contexts. A total 63 questions were obtained, including 10 language-dependent questions, 11 common questions, 28 consonant questions, and 14 vowel questions. The four set questions are describes in more detail as following:

(1). Language dependent questions: Unlike those question sets used in Western language, the question set used here had to include Mandarin and Taiwanese several major different phonological rules and phonetic composition. Therefore, we design several questions to discriminate or merge the two languages.

(2). Common questions: In this set we include general characteristics in usual conversation.

(3)(4). Consonant/vowel questions: Consonant and vowel types are another important factors in our experiments. Thus we have a set of questions to discriminate them.

4.3 Stopping Criteria

The acoustic decision tree can be applied to a state or a model level for node unit. However, since each state has its own effect on the model, and the results of (Liang, 1998), we choose state as the units to be clustered, although we also use model as a node unit before. Clustering quality is sensitive to the combination of thresholds used, and determining appropriate values requires time-consuming construction and comparison of several sets of tree. Pruning-based methods for state clustering are used in this paper because no arbitrary stopping thresholds need be specified appropriate values are learned automatically during the pruning process. The best results are selected empirically.

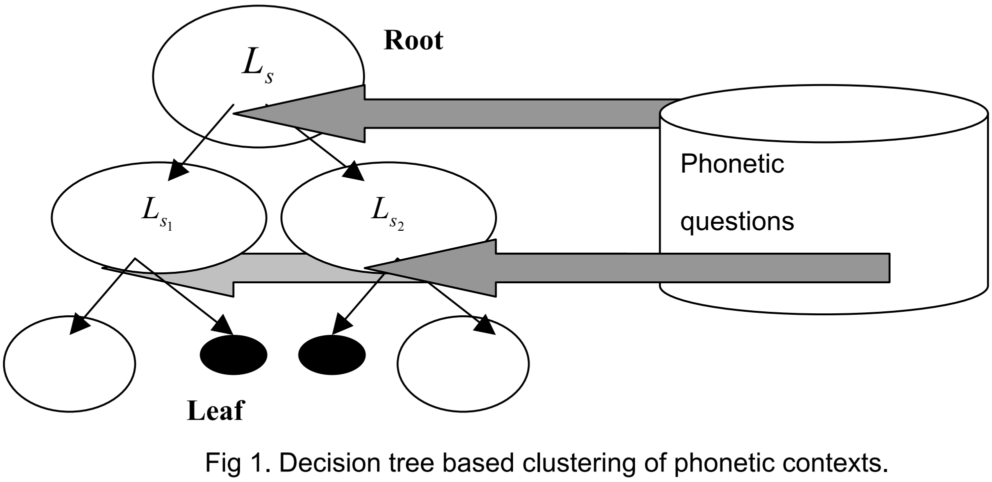

Fig 1. Decision tree based clustering of phonetic contexts.

5. Bi-lingual Recognition System

In order to solve the appearance of co-articulation in continuous speech, we use inter-syllable triphone models. Because of the acoustic model numbers grow so huge (from 1280 to 7394), the occurrence of training data among models aren't balanced, some states "allot" 20 frames, some others "allot" more than ten thousands frames. To break the equal mixture numbers, the CHMM mixture numbers were adjusted according to the amount of training data. The more the training data are, the more of the mixture numbers are. Finally the Gaussian mixtures in a state are from 2 to 16. In Figure 2 it was listed the statistics of the acoustic phonetic labels in the training data. They are Mandarin, Taiwanese, bi-lingual intra-syllable triphones and bi-lingual inter-syllable triphones. In this picture the horizontal and vertical axis are the model's index and the count of the phonetic units, also we can see the counts of bi-lingual inter-syllable triphones of almost 40% models are below 10. So the change of the mixture depending on the amount of training data is much better than a uniform mixture number for all models. We can see the results in <table 4> row 3 and row4, adjust mixture numbers according to training data is better than fixed mixture numbers, even the total number of mixtures are lesser than row 3. They are relative 10.5%, 12.1% and 9.2% improve syllable accuracy rate for uni-gram, bi-gram and tree net.

In the improvement of decision tree task, we do lots of experiments in tuning the best performance. Therefore, we listed the attribute results at row 5-7 of <table 4>. In those experiments, we set different thresholds for stopping criteria. In (*Sanker, 1998*) was shown to us "the better performance is achieved by reducing the number of HMMs state clusters and increasing the number of Gaussians per state cluster ". So that we follow the direction of this advisement, do DT1, DT2 and DT3 experiments by decreasing the number of state, and then increase the mixture numbers per state. In other words, we increase the lower bound of the minimum training data per cluster, or the value of ΔL, make sure that the every leaf node in a tree can get more training data comparatively. Even though DT1 and DT2 the mixture numbers are much lesser than row 4 in <table 4> (almost half size), but the performance is better. However, DT3, we get the worse result relative DT2, the reason we thought is over-pruning as a result get side effect. At last, compare with the baseline, we achieve the syllable accuracy rate improvement 17.18% under uni-gram, and 20.54%, 8.69% under bi-gram and tree net language model condition

The decision tree based tying algorithms are used to improve the bi-lingual system performance. In decision-tree based approach, if a question is more frequently asked than the others, or its dissimilar value is also bigger than the others, then the associated phonetic classification is considered more important. In our experiments, the principal question set is language dependent questions set, then is the vowel set and consonant set, last is common questions set. It means in our questions the language set is the dominant set influence the decision tree construction.

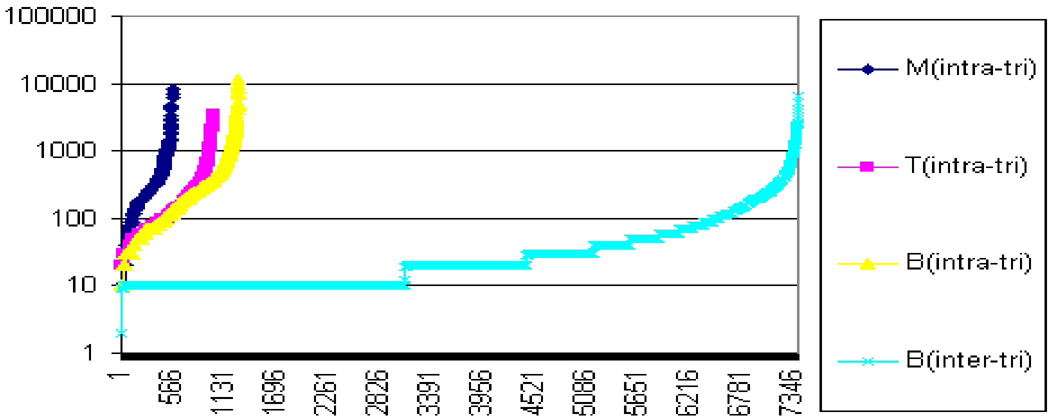

Fig 2. The statistics of the occurrence to corresponding model in training data.

6. Conclusion

In this paper, we investigated the phonetic decision-tree state clustering sharing acoustic models for bi-lingual speech recognition. From the baseline system, after expending models; mixture's numbers adjust to fit for the training data, and the decision-tree states clustering, the final results were shown in the <Table 4>. In this table we can obviously see the advancement in each step that we exert. In future experiments, we will do more research about different ways of optimizing the unconstrained bi-lingual language model, and new approaches for sharing acoustic models across languages. More languages issues will be included in the system.

Table 4. The experimental results using inter-syllable phone unit, and based on decision tree tying algorithms. The row 1 value is language model relative perplexity (unit syllable accuracy rate %)

	# of state	# of mixture	Uni-gram (1049)	Bi-gram (664)	Tree (98)
Bi-lingual base line	1280	30720	54.70%	57.79%	82.62%
Inter-uniform	22182	177456	52.8%	56.90%	68.78%
Inter-adjust	22182	136176	58.32%	63.77%	75.09%
DT1	11091	68002	62.57%	65.98%	85.06%
DT2	5546	33894	64.10%	69.66%	89.80%
DT3	3773	30184	62.61%	66.26%	87.33%

REFERENCES

Ananth Sanker (1998). Experiments with a Gaussian Merging-Splitting Algorithm for HMM Training for Speech Recognition, Proceeding of Broadcast News Transcription and Understanding Workshop

P.C. Woodland C.J. Leggetter, J.J. Odell, V. Valtchev and S.J. Young (1995) The 1994 HTK Large Vocabulary Speech Recognition System, Proceedings ICASSP'95, Detroit, 1995.

Po-yu Liang Jia-lin Shen and Lin-shan Lee, (1998) Decision tree clustering for acoustic modelling in speaker-independent mandarin telephone speech recognition, Proceeding of the first International Symposium on Chinese Spoken Languages Processing (ISCSLP'98), Singapore, 1998

Ren-yuan Lyu. Chi-yu Chen, Yuang-chin Chiang and Min-shung Liang (2000) Bi-lingual Mandarin/Taiwanese(Min-nan), Large Vocabulary, Continuous Speech Recognition System Based on the Yong-yong Phonetic Alphabet., ICSLP2000,Oct. 2000, Beijing, China

Sheng Gao; Tan Lee; Wong, Y.W.; Bo Xu; Ching, P.C. and Taiyi Huang (2000) Acoustic Modeling for Chinese Speech Recognition: A Comparative Study of Mandarin and Cantonese, Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on, pp. 1261 -1264

Tanja Schultz and Alex Waibel (1999) Language Adaptive LVCSR Through Polyphone Decision Tree Specialization. Workshop on Multi-lingual Interoperability in Speech Technology (MIST '99). Leusden, The Netherlands

Waibel, A. (2000) Multilinguality in Speech and Spoken Language Systems. Geutner, P.; Tomokiyo, L.M.; Schultz, T.; Woszczyna, M. Proceedings of the IEEE, , Volume: 88 Issue: 8, pp. 1297 -1313