A LARAGE-VOCABULARY TAIWANESE (MIN-NAN) MULTI-SYLLABIC WORD RECOGNITION SYSTEM BASED UPON RIGHT-CONTEXT-DEPENDENT PHONES WITH STATE CLUSTERING BY ACOUSTIC DECISION TREE

Ren-yuan Lyu[1], *Yuang-jin Chiang*[2], *Wen-ping Hsieh*[2]

[1.] Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
[2.] Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan
Email: rylyu@mail.cgu.edu.tw, rylyu@ms1.hinet.net
Tel: 886-3-3283016#5677

ABSTRACT

In this paper, we apply context dependent phonetic modeling on the task of large vocabulary (with 20 thousand words) Taiwanese multi-syllabic word recognition. Considering the phonetic characteristics of Taiwanese, the right context dependent (RCD) phones instead of the general tri-phones are used. The RCDs are further clustered at the sub-phone or state level using a decision tree with a set of context-split questions specially designed for Taiwanese speech according to the acoustic/phonetic knowledge. For the speaker dependent case, 7.18% word error rate is achieved. A real-time prototype system implemented on a Pentium-II personal computer running MS-Windows95/NT is also shown to validate the approaches proposed here.

1. INTRODUCTION

Taiwanese, one of the major Chinese dialects, is the mother tongue of more than 75% of the population in Taiwan. It belongs to a larger Chinese dialectical family called Min-nan (or Southern-Min, Southern-Hokkian), which is also used by many overseas Chinese living in Singapore, Malaysia, Philippine, and other areas of Southern-East Asia. It was estimated that this language has more than 49 million speakers and is ranked in the 21th place in the world, according to the 1996 Ethnology. In the past few decades, scientists in Taiwan did speech recognition research on Mandarin speech. Some achievements have been achieved in recent years.[1][2][3] Since Taiwanese is another major language spoken in this land, and Taiwan is basically a multilingual society, it is non-trivial to develop a similar large-vocabulary speech recognition system for Taiwanese speech. [4] In this paper, some preliminary work about Taiwanese speech recognition has been done, including the study of Taiwanese phonetics, setting up a Taiwanese lexicon, selecting several sets of phonetically balanced words to be used in speech data collection, and recording a Taiwanese speech database.

The basic technology adopted here is the continuous Hidden Markov Model (CHMM) because of its success in speech recognition in the past decades. We adopt CHMM to model the basic Taiwanese phone units, considering both the inside- and inter-syllabic coarticulation. Since the training data here is relatively few, however, some sharing of data among the models is also attempted to try. Decision tree clustering using acoustic/phonological knowledge provides an efficient approach of data sharing. [8][9] We choose clustering in the sub-phonetic level, i.e., states in the whole phonetic models, in which each specific state in each model has its own decision tree. A set of 25 yes/no phonological questions was used to construct each decision tree. After the tree was constructed, models within each leaf of the tree were tied together, and the states of these models shared the same training data.

Experiments were conducted by choosing different state(s) to be clustered in a model. In a speaker-dependent large-vocabulary (20K word) recognition experiment adopting a N-best Viterbi beam search [7], the word error rare (WER) achieved is 7.18% for the case of choosing only the first state of a model to be clustered.

This paper is organized as follows. The Taiwanese phonetics is summarized in section 2. The scripts, lexicon, and database are described in section 3. Some basic speech technologies used here are summarized in section 4. The acoustic decision tree for Taiwanese is described in detail in section 5. In section 6, the results of the experiments are reported and then a prototype system is shown to validate these approaches. Several concluding remarks are finally given in section 7.

2. TAIWANESE PHONETICS

Taiwanese, like Mandarin or Cantonese as a member of Sino-Tibetan language family, is a tonal, monosyllabic language. Traditionally, a Taiwanese syllable is decomposed into three parts, namely an Initial, a Final and a tone. Take the syllable "dan4" (to wait) as an example, where /d/ is an Initial, /an/ is a Final, and /4/ represents a high-falling tone. There are 18 Initials (including one null Initial), 47 Finals, and 7 lexical tones in Taiwanese. [4][5][6] An Initial is equivalent to a consonant, but a Final can be further decomposed into 1 to 3 vowels plus possible consonants. The most distinct feature of Taiwanese which is different from Mandarin is that for each Final, there is a corresponding “entering-tone” Final, which is ended with an unreleased /p/, /t/, /k/ or /h/. All the phonemes are listed in <table.1>, each of which is represented by a set of specially designed phonetic alphabet called Daiim, and has a one-to-one mapping to the International Phonetic Alphabet (IPA). [4][5]

Furthermore, Taiwanese is also a tonal language with more complex tonal structures than that of Mandarin. It has 7 lexical tones, two of which are carried in syllables ending with final /p, t, k, h/ (called entering-tone). Since the task we are considering here is the recognition of multi-syllabic words, which have relatively few homonyms even when the tones are disregarded. In this initial study, we decided not to deal with the issues of tones and then reduce the 1683 phonologically allowed tonal syllables to 714 base syllables. That is, each word in the lexicon is represented as a concatenation of base syllables. The word recognition task becomes the recognition of base syllable strings.

3. LEXICON AND DATABASE

A Taiwanese pronunciation lexicon of about 20 thousand words, each of them has a corresponding string of phonetic symbols encoded in Daiim phonetic alphabet, has been set up for this initial study. [5] In this lexicon, there are 19152 ordinarily used Taiwanese words, composed of 48318 syllables, i.e., each word containing 2.52 syllables in average.

For evaluation of the recognition system, we select several sets of words with different features:

1) R1000: 1000 randomly selected words, each of which contains 2.55 syllables in average;
2) H500: 500 highest frequently used words, each of which contains 2.12 syllables in average;
3) N407: 407 place names, each of which contains 2.08 syllables in average;
4) P396: 396 phonetically rich words, each of which contains 3.24 syllables in average.

The statistics of the evaluation set is listed in <table.2>.

Besides, a training set which contains as few words but as much phonetic variety as possible is chosen by a word selective procedure as follows: 1) Determine the phonetic unit to be used in the recognition system such as RCDs adopted here. 2) Each new selected word contains the maximal number of possible new phonetic units. 3) Include all distinct speech units, which appear in the lexicon. As a result, a set of 472 words containing all the 1029 distinct RCD phones found in the lexicon were selected. In addition, several extended sets of words, which contain as many distinct RCD phonemes as possible, were also selected to enhance the phonetic variety. Furthermore, a set of single-syllabic words, containing all phonologically possible syllables, was picked out, too. The statistics of all the sets of words used in the training session is listed in <table.3>.

The speech database used for training and evaluation were recorded by two adult speakers, including one male and one female, over a period of one month. A close-talk head-set microphone plugging in a SoundBlaster card in a Pentium-II personal computer was used. The speech waveform was sampled at 16 KHz. The statistics of the speech database is also listed in <table.2> and <table.3>.

4. SIGNAL PROCESSING, SPEECH UNITS, AND SEARCHING

The speech waveform was multiplied by a 16-ms Hamming window first. A set of 12-dimensional mel-cepstral coefficients and 1-dimentional log energy was extracted to form a 13-dimentional feature vector for each frame which shifts forward every 8 ms. A time window of 5 frames of feature vectors were used to compute the corresponding 13-dimentional delta coefficients. These 2 sequences of feature vectors and delta feature vectors were treated as statistically independent and modeled by separate Gaussian mixture densities in CHMM.

In this paper, we adopted phones, considering the right context dependency both inside a syllable and inter syllables, as the basic speech units to be modeled as CHMM. It is believed that the coarticulation effect inside a syllable is more severe than that between 2 syllables for the monosyllabic language, such as Mandarin, Cantonese or Taiwanese. So, it is natural for researchers to consider the inside-syllable coarticulation in the previous literatures. [2] In such a case, only phones that are not located in the right most position of a syllable can be right context dependent. There are thus 208 such inside-syllable RCD’ s (iRCD). However, when the speed of utterance increases, the coarticulation across 2 syllables becomes severe. In addition, for the vowel-vowel concatenation between 2 neighboring syllables, the coarticulation effect may be very severe even when the speed of utterance is slow. To alleviate such a difficulty, the inter-syllabic modeling was considered.

To deal with the problem of co-articulation across syllables, we consider using the outside-syllable RCD phones (oRCD), in which the right most phone in a syllable is dependent on the left most phone of the next syllable. From the same 20K lexicon, 1029 oRCDs are extracted. Since each new model not found in the set of iRCDs comes from certain iRCD with syllable boundary as its right context, it is reasonable to use such a kind of iRCD models as seed models to generate all its corresponding oRCD models.

However, since the number of oRCDs is relatively large in the viewpoint of the limited amount of speech data available, we have to consider some data sharing approaches to make full use of the speech data. The acoustic decision tree method is an efficient approach to achieve this purpose and will be described in detail in section 5.

The 20K-word lexicon is organized in terms of the chosen speech units as a tree data structure to be used as the search space. There are about 58K nodes in the lexicon tree, with each node containing one chosen speech unit. A widely used Viterbi beam search is then used to find N best paths and then the N candidates of the recognized words. [7]

5. STATE CLUSTERING BY ACOUSTIC DECISION TREE

It is known that increasing the specificity of models may decrease the trainability because of a limited amount of training data. To get the optimal trade-off between specificity and trainability, data should be shared among similar models. A method called "Acoustic Decision Tree" is known to deal with this problem quite well. [8][9]

A decision tree is a binary tree in which a yes/no question is attached to each node, which is then split with respect to the answer. Those questions are designed from knowledge of phonetics. A set of 25 questions specially designed for Taiwanese speech was proposed here. Some of them are listed in <table.4> For each node, the “best” question is chosen to be asked and the node is split according to the answer to the best question.

The criterion to decide the best question for splitting a tree node is computed as follows:

$$L_S = \sum_{m \in S} N_{mi} \left| \vec{m}_{mi} - \vec{c} \right|^2$$

where N_{mi} is the occupancy of state i of model m, $\vec{m}_{mi}$ is the mean vector of state *i* of a model *m* in the tree node *S*, and $\vec{c}$ is the center vector of all $\vec{m}_{mi}$ in node *S*. L_S is then the measure of the dissimilarity of all models in the tree node *S*. Each splitting should maximize the decrease of the measure of the dissimilarity , i.e., the increase of similarity, when the parent node *S* splits into 2 children nodes S_1 and S_2.

That is, to maximize

$$\Delta L = L_S - (L_{S_1} + L_{S_2})\,.$$

A threshold of ΔL is set by experiments to stop 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.

We construct decision trees for each state of each kernel phone. Totally 25 questions are asked to make judgments about right contexts. The concept of acoustic decision tree is shown in <fig.1>, where the objectives for clustering are states of models with /a/ as their kernel unit. There are totally 33 models with /a/ as their kernel unit, so the node of the first state of models a+"S" contains 33 mean vectors for classification. As the splitting is going, these 33 mean vectors are finally grouped into about 5 groups.

The advantage of tree classification is that it can generate a vocabulary-independent system. For each new RCD unit, we can ask the same 25 questions in a hierarchical manner and classify it into a leaf node. Whether or not this unit has been shown in the training corpus, it shares the model of that leaf node and thus can be recognized.

The Acoustic Decision Tree can be applied to any level of speech unit. However, since each state has its own effect on the model, we choose state as the units to be clustered. In order to get better trade-off between number of states and recognition accuracy, experiments of clustering different states in a model have been tried and all the results are reported in the following section.

6. EXPERIMENTAL RESULTS

To validate the approaches proposed here, several experiments have been done. The baseline system is based on 208 iRCD models. Each model is a left-to-right Hidden Markov model containing two states, and each state contains two mixtures. The word error rate (WER) obtained is 13.57%/4.94% for top1/top5 candidate. Since the task is for speaker dependent case in this initial study, we try other model prototypes, i.e., increase the state number to 3 but decrease the mixture number to 1. From our experiments, we obtain better results (WER=12.00%/3.19% for top1/top5 candidate) and thus use it as the baseline system compared with the following experiments.

The 1029 oRCD models were tried, too. Since they model the coarticulation effect across syllables better, the accuracy is also better. The WER for top1 candidate becomes 7.89%, a significantly improvement. However, the total state number increases from 624 to 3,087, and thus the computational load increase nearly 3,087/624!

State clustering by acoustic decision tree is performed then. Here 3 different cases have been tried, i.e., clustering the first state only (S1), clustering both state 1 and state 2 (S12), and then clustering all 3 states (S123). The WER we obtained here is 7.18%, 7.22%, and 8.54%. The total state number for 3 cases are 2,501, 1,902, and 1,368, respectively.

All the experimental results are summarized in <table.5>.

To validate the approaches proposed in this paper, a prototype system was implemented on a Pentium-II personal computer running MS-Windows95/NT, by using the configuration of N3M1/oRCD/S1 as described in <table.5>, considering the trade-off between speed and accuracy. The graphic user interface (GUI) is shown as in <fig.2>, where one can see the top 5 recognized candidates for each utterance.

7. DISCUSSIONS AND CONCLUSIONS

In this paper, we have presented a speaker dependent large vocabulary recognition system. It is an initial research for Taiwanese speech recognition. We set up the basic structure for recognition and adopt several kinds of phone units to model speech sounds. For the basic structure, we combine beam search with tree lexicon to reduce the searching time. We use inside-syllable RCD phone units as our baseline system. The WER is 12.00%. After adding inter-syllable RCD models to consider co-articulation effects, it reaches WER of 7.89%. The error reduction is about 34.25% at the cost of 3087/624 increase of computational load.

To save memory, computational time, and get more robust models, we use Acoustic Decision Tree to do state clustering. We designed a set of 25 context-splitting questions for Taiwanese and set a criterion for maximizing the increase of a measure of similarity. This method has reduced the total number of states to 2501/3087 of inter-syllable RCD models and further reduce the WER to 7.18%.

The strategies used for our recognition system are standard approaches. For further researches, we can refine our phone units to contain more context complexity. As to state clustering criterions, we can try an entropy based criterion to classify phone units, too.

8. REFERENCE

[1] Lin-shan Lee, etc, 'Golden Mandarin (II) - An Improved Single-chip Real-time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary', ***ICASSP-93***, Apr. 1993, pp. II-503-506

[2] Ren-Yuan Lyu, et al. "Golden Mandarin (III)-User-Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary", ICASSP-95,May. 1995 pp57-60

[3] Hsin-min Wang, et al., "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data", IEEE Trans. on Speech and Audio Processing, vol. 5, no. 2, pp.195-200, March 1997.

[4] Ren-yuan Lyu, Yung-jin Chiang, Ren-jou Fang, Wen-ping Hsieh, "A Large-Vocabulary Taiwanese (Min-nan) Speech Recognition System Based on Inter-syllabic Initial-Final Modeling and Lexicon-Tree Search", ***ROCLING XI Conference***, Aug. 1998, Hsinchu

[5] Yuang-jin Chiang, " The Daiim Taiwanese Input System, Ver4.1", Inst. Of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan, 1994

[6] Robert L. Cheng, "Taiwanese and Mandarin Structures and Their Developmental Trends in Taiwan--I: Taiwanese Phonology and Morphology", 1997

[7] C.H. Lee, etc, " A frame-synchronous network search algorithm for connected Word recognition", IEEE Trans.

ASSP, pp. 1649-1658, Nov. 1989

[8] J.J. Odell, “ The Use of Context in Large Vocabulary Speech Recognition”, PHD Dissertation of Cambridge University, 1995

[9] H.W. Hon, “Vocabulary-Independent Speech Recognition: The VOCIND System”, PHD Dissertation of Carnegie Mellon University, 1992

<table.1> A List of Phonemes in Taiwanese

b	p	m	v	d	t	n	l	g	k	nq
q	z	c	s	r	H	-p	-t	-k	-h	
a	i	u	e	o	Or	A	I	U	E	O

<table.2> Statistics of the Testing Word Sets

		Nword	NSyl	Tspeech	
				Male	Female
Testing Word Sets	R1000	1,000	2.55	826	656
	H500	500	2.12	361	397
	N407	407	2.08	304	311
	P396	396	3.24	385	256
	The whole	2,303	2.49	1,876	1,620
Lexicon		19,152	2.52	N/A	N/A

<table.3> Statistics of the Training Word Sets

		Nword	NRCD	TSpeech	
				Male	Female
Training Word Sets	Min_word	472	1,029	459	445
	Ext_word	1,045	1,029	965	981
	Single_syl	2,874	213	1,417	1,486
	The whole	4,391	1,029	2,841	2,912
Lexicon		19,152	1,029	N/A	N/A

Nword: Number of words totally
NSyl: Number of syllables per word
TSpeech: Speech Length in seconds
NRCD: Number of distinct RCD phonemes

<table.4> Some Examples of 25 Questions for Acoustic Decision Tree

	Questions	Members
Q1	Is its right context a vowel?	a, A, e, E, i, I, o, O, or, u, U
Q2	Is its right context a right nasal_vowel?	A, E, I, O, U
Q3	Is its right context a glottal stop? or Unreleased p, t, k?	-p, -t, -k, -h
Q4	Is its right context a voiced consonant?	m, n, -m, -n, -ng, b, d, g, z, l, q, v, r
Q5	Is its right context a nasal?	m, n, ng , -m, -n, -ng
Q6	Is its right context a bilabial?	b, p, m, v
…	…	…

<table.5> The Experimental Results

	WER%	NState
N2M2/iRCD	13.57	416
N3M1/iRCD	12.00	624
N3M1/oRCD	7.89	3,087
N3M1/oRCD/S1	**7.18**	**2,501**
N3M1/oRCD/S12	7.22	1,902
N3M1/oRCD/S123	8.54	1,368

WER: Word Error Rate
NState: Number of the total states to be calculated
N2M2: 2 states, 2 mixtures
N3M1: 3 states, 1mixtures
iRCD: inside syllable RCD
oRCD: outside syllable RCD
S1: clustering state 1
S12: clustering both state 1 and state 2
S123: clustering all 3 states