A NEW HYBRID DURATION HIDDEN MARKOV MODEL WITH APPLICATION TO LARGE VOCABULARY TAIWANESE (MIN-NAN) WORD RECOGNITION

CHIANG Yuang-Chin[1], FANG Ren-Zhou[1], HSIEH Wen-Ping[1], LYU Ren-Yuan[2]

[1]National Tsing Hua University, Hsin-chu, Taiwan

[2]Chang Gung University, Tao-yuan, Taiwan

Email: rylyu@mail.cgu.edu.tw Http://dsp.cgu.edu.tw Tel: 886-3-3283016#5677

ABSTRACT

A new hybrid duration Hidden Markov Model (hdHMM), which combines the ideas of both the infinite duration models and finite duration models, is proposed here and applied to a large vocabulary Taiwanese speech recognition task. Such a model not only has better state duration distribution than the traditional left-to-right HMM but also is more computational efficient than the finite-duration HMM. The experiment was performed on a large vocabulary Taiwanese (Min-nan) multi-syllabic word recognition. For the speaker dependent case, the best word error rate achieved here is 7.9%. Since this paper is also one of the first papers on the speech recognition of Taiwanese speech, some basic facts about Taiwanese phonetics is also briefly introduced.

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 Southeast Asia. It was estimated that the population was more than forty-nine millions and was ranked twenty-first in the world, according to the 13th Edition of the 1996 Ethnologue. [1] In the past few decades, research on speech recognition in Taiwan focused only on the Mandarin dialect, and some commercial products have been developed. [2] Since Taiwan is a multilingual society, it is natural to study and develop computer applications for another principal languages on the island. Based on the experiences learned from the Mandarin recognition task, we choose large vocabulary multi-syllabic word recognition of Taiwanese as our pioneer study. [3][4]

The basic technology adopted here is the continuous Hidden Markov Model (CHMM) because of its success in the past decade. We use CHMM to model the Taiwanese INITIAL/FINAL phonetic units, considering both the inside- and inter-syllabic co-articulation. Promising results were achieved in our initial study, with the word error rate (WER) being 7.9% for the speaker dependent case.

With the traditional left-to-right HMM, each state has fixed probability to loop back to itself, and thus the duration distribution for each state is implicitly geometrically distributed. But that is far from the empirical distribution observed from the real data. Some researchers proposed so-called finite-duration model topology by replacing the self-looping with a finite number of replicates of the states and then estimating the transition probabilities between those replicating states individually. [5] This does improve the recognition rate, but with a cost of much higher computation time. We here propose a hybrid duration model by setting the model topology as the same as in the finite duration model, while letting the last replicated state have a self loop. For such a topology, it can be shown that the state distribution is closer to the actual one. This results in similar improvement in recognition rate as the finite-duration model, while retains the same computation efficiency as in the traditional left-to-right model case.

This paper is organized as follows. Section 2 is a brief introduction on Taiwanese, lexicon, and speech corpus. Section 3 is the front-end signal processing. Section 4 describes our baseline INITIAL/FINAL system. The discussion on duration model is in Section 5. Last section reports our experimental results.

2. TAIWANESE, LEXICON, AND SPEECH CORPUS

To facilitate the study, we need a pronunciation lexicon, phonetically balanced speech corpus for training and a set of testing speech. We begin this section with a very brief introduction on Taiwanese.

Taiwanese, like Mandarin as a member of Sino-Tibetan language family, is a monosyllabic, tonal language. According to linguistics, there are 18 INITIALs, 94 FINALs, and 7 tones. [11] They are listed in the Appendix, where the phonetic symbols we adopted is the TongYong phonetic alphabet, which is a set of general, Ascii-based, phonetic symbols suitable for the common use of the three major languages in Taiwan, i.e., Taiwanese, Hakka, and Mandarin.

The tones are traditionally classified into 7 different types, which can be further clustered into 2 groups, i.e., entering tones and non-entering tones. In fact, the syllables carrying entering-tones and non-entering tones are different base-syllables because of their different FINALs. The issue about tone sandhi is more complicated than other tonal languages, such as Mandarin, and is beyond the scope of this paper.[9][10]

The pronunciation lexicon was extracted from the one used in the Daiim Input Method, which is a software for inputting Taiwanese into the computer by the keyboard.[8] There are 19152 multi-syllabic words, containing 48318 syllables. Of the lexicon, there are 1683 different tonal syllables and 714 different base syllables disregarding the tones. In this paper the recognition task does not include the recognition of tones, and thus the word recognition task becomes the recognition of base-syllable strings. From the near 20k lexicon, we extract several sets of words that contain all appropriate speech units, including base-syllables, context-dependent INITIAL/FINALs and context-dependent phonemes. We also have another set of single-syllabic words consisting of all phonologically possible syllables. These sets are used for recording our training speech data. A male and a female speaker recorded all the scripts. The testing speech data contains of 407 place names in Taiwan, a set of 396 phonetically balanced words, 500 highest frequently used words and 1000 randomly selected words out of the near 20k lexicon. Again, the same male and female speaker recorded the evaluation speech. The statistics of speech corpus is summarized in <table 1 >.

3. FRONT-END SIGNAL PROCESSING

The input speech waveform was pre-emphasized with a coefficient 0.975 and is then multiplied by a 16-ms Hamming window. A set of 12 dimensional Mel-cepstrum and 1-dimensional log energy was extracted to form a 13-dimensional feature vector for each frame which is shifted forward every 8 ms. A time window of 5 frames of feature vectors is used to compute the corresponding delta cepstral coefficients. These 2 sequences of feature vectors are treated as statistically independent and modeled by separate Gaussian mixture densities in CHMM.

4. THE BASELINE INITIAL/FINAL SYSTEM

CHMM is adopted to model the INITIAL and FINAL of each of the 714 Taiwanese base syllables. However, co-articulation within a syllable and between syllables need to be considered.

4.1. Inside-Syllabic Co-articulation Modeling

Initially, the context dependent (CD) INITIAL and context dependent FINAL was used as the speech units. Of the 714 base syllables, there are 147 CD INITIALs and 77 CI FINALs. A silence model is also added. The model topology is a standard left-to-right model without skipping states, with the number of states twice the number of the phonemes in that speech unit and each state having 2 Gaussian mixtures. The parameters of these models are estimated after several iterations of Baum-Welch re-estimation procedure from the training data. And then the Viterbi beam search in the lexicon tree is used to find N best candidates of the recognized words.

4.2. Inter-Syllabic Co-articulation Modeling

Examining the recognition results of some preliminary experiments, one finds that many recognition errors occur in words whose syllables start or end with a vowel (or nasal) and form consecutive vowels (or nasals) with previous or next syllable. For an example, the word “leather-shoes” is pronounced as a 2-syllable word as /pue-e/, which have a 2 consecutive vowels across the syllable boundary. Such words are often mis-recognized. To alleviate such problems suggests the need to include models for those inter-syllable co-articulation. From the 20k pronunciation lexicon, we extracted a set of 105 inter-syllable right context dependent (ISRCD) phones to be used as the speech units. As will be seen in the experimental results shown in section 6, the ISRCD phone modeling reduces WER as expected.

5. A NEW HYBRID DURATION MODELING

In this section, we will discuss the left-to-right duration model, the finite duration model and two hybrid duration models. The results in recognition rate as well as computation time will be summarized in the next section.

5.1. Infinite Duration (ID) modeling

The traditional left-to-right topology of a single state is shown in <fig.1>. The state transition probabilities are constant and this implies, if unconstrained, the number of consecutive frames generated from a state would have a geometric distribution. The duration distribution of a typical state is thus:

$$f(d) = (1 - a_{self}) \times (a_{self})^{d-1} \quad , where\ d \in N \ .$$

After estimating a_{self} from the training speech data, the probability density function (pdf) of the state duration for a particular state is depicted in <fig.2>

along with the actual observed distribution. We see that the actual observed duration is distributed more like a gamma function. The discrepancy between modeled and actual distribution suggests more refined modeling for the duration distribution.

5.2. Finite Duration (FD) Modeling

In order to model the duration more accurately, a finite duration model topology for HMM has been proposed. [5] This is done by replacing the self-looping state with a string of P replicated states, as illustrated in <fig.3>. The number P is determined by

$$P = E(S) + K \times stddev(S),$$

where *S* is number of frames that have been mapped to the state obtained from the infinite duration experiment, *E(S)* is expected value of *S*, *stddev(S)* is standard deviation of *S*, and *K* is a small fixed integer to be determined experimentally. The duration distribution of a particular state is then

$$f(d) = a_d \prod_{i=1}^{d-1}(1-a_i), \quad where\ a_i = 0\ if\ i \geq P$$

Experiments with *K* values being 2, 4 or 6 were conducted here. The resulting estimated duration distributions and the actual observed duration from the infinite duration case for a particular state are sketched in <fig.4>. We see that as *K* grows, the distribution is closer to the actual one, and the WER is also reduced.

Unfortunately, the improvement in WER comes with a price in computation time. Since the number of replicated states is large due to the large variance of duration, this leads to a relatively large parameter space and search space, and makes it impractical to be used in a real time system. To alleviate such a deficiency, we try to combine the finite duration modeling with traditional infinite duration modeling in hope to improve the WER with little computations.

5.3 Hybrid Duration (HD) Modeling

First Hybrid Duration Model

By making the last replicated state have self loop in the finite duration model, we have our first hybrid duration model, as illustrated in <fig.5a>. The distribution of the hybrid duration model is as follows:

$$f(d) = a_d \prod_{i=1}^{d-1}(1-a_i), \quad where\ a_i = 0\ if\ i > P$$

Under this model, the recognition result is comparable to the finite duration case, while the computational efficiency is improved.

Second Hybrid Duration Model

After several trials, we come up with a somewhat more complicated model as illustrated in <fig.5b>. The distribution for the second model can be shown as

$$f(d) = \begin{cases} a_1 & if\ d = 1 \\ a_2 \times (1-a_{self}) & if\ d = 2 \\ f(d-1) \times a_{self} + (1-a_1-a_2) \times \left(\prod_{i=3}^{d-1}(1-a_i) \right) \times a_d \times (1-a_{self}) & if\ 3 < d \leq P+1 \\ f(d-1) \times a_{self} & if\ d > P+1 \end{cases}$$

For the case $a_1 = 0$ and $a_1 > 0$, the estimated distribution and the actual one are depicted in <fig.6>. Although the shapes of the estimated densities between $a_1 = 0$ and $a_1 > 0$ do not differ much, both the accuracy and computational efficiency for $a_1 = 0$ are improved.

Further examination of the experimental results reveals that many recognition errors occur because that some states have short duration. This suggests a model with minimum duration $m < P$. This can be easily done by setting $a_1 = \cdots = a_m$. The case that $a_1 = 0$ corresponds to $m = 1$.

By choosing m and P carefully, both the recognition rate and computation time are improved. Table 2 reports the best case with $P = E(S) - 2$ and $m = E(S)/3$. The improvement in computation time comes somewhat surprisingly, since the parameter space of the second hybrid duration model is larger than that in the ID case. The reason may be that by using a better duration model, the parameters are more accurately estimated. And thus the beam search in the recognition stage can better differentiate between confusing states. In <fig.7> we compare the estimated densities for $m = E(S)/2 \times 3$, $E(S) - 2,$, and $E(S)$.

6. EXPERIMENTAL RESULTS

Our experimental results are summarized in <table.2>. We give a brief conclusion of the experiments on word recognition as the following:
For the baseline study the WER we achieved in average is 11.4% for inside-syllable modeling and 9.4% for inter-syllable modeling. The speed for each case is approximately the same, about 3 times of the real time on a popular UltraSparc machine.

Applying finite duration model for the inter-syllable setup, the WER decreases to 8.1%. The average time needed to recognize one second of

speech grows from 3.20 sec to 12.22 sec.

To alleviate such a deficiency, we combine the finite duration modeling with traditional infinite duration modeling to a hybrid duration model. The same WER reduction can be obtained at less computation time. By choosing the minimum duration m carefully, the WER drops to 7.9%, and the CPU time decreases to 2.01 sec for each second of speech data.

7. REFERENCE

[1] http://www.sil.org/ethnologue/top100.html

[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, pp57-60

[3] Ren-yuan Lyu, Yuang-chin Chiang, Ren-Zhou Fang and Wen-Ping Hsieh, "A Large-Vocabulary Taiwanese (Min-nan) Speech Recognition System Based on Inter-syllabic INITIAL-FINAL Modeling and Lexicon-Tree Search", Rocling98, Aug, 1998, Hsin-chu, Taiwan

[4] Ren-yuan Lyu, Yuang-chin Chiang, Wen-ping Hsieh, Ren-zhou Fang 'A Large-Vocabulary Taiwanese (Min-nan) Multi-syllabic Word Recognition System Based upon Right-Context-Dependent Phones with State Clustering by Acoustic Decision Tree', ***International Conference on Spoken Language Processing***, Nov. 1998, Sydney, Australia

[5] Pincone, J. "Continuous Speech Recognition Using Hidden Markov Models", IEEE Signal Processing Magazine, July, 1990, vol. 7, no. 3, pp26-41

[6] Qan, GokRin, YuangChin Chiang, "The Report of Study on the Frequently Used Taiwanese Spoken Words" (" x f ` ", in Taiwanese.) Institutes of Statistics, National Tsing-Hua University , Taiwan, 1995.

[7] Chiang, YuangChin, "Some Principles of Selecting Taiwanese Written Script" (" r ", (In Taiwanese) Newsletter of Center for Taiwan Study, 5-6, 1995, pp40-69.

[8] Chiang YuangChin, "Daiim Input Method version4.1", Institutes of Statistics, National Tsing-Hua University , Taiwan, 1997

[9] Chiang YuangChin, " x O ", ((In Taiwanese.) Newsletter of Center for Taiwan Study, Oct, 1998. (To appear.)

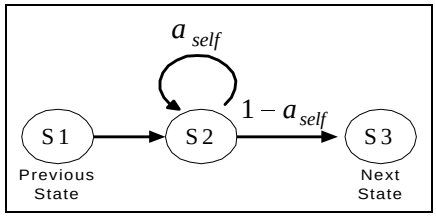
[10] Ko SeKai, Chiang YuangChin et al., " f gah ", (In Taiwanese.) Taiwanese World, Vol.12, 1997.

[11]Ong IokDik, " x ";` 1957.

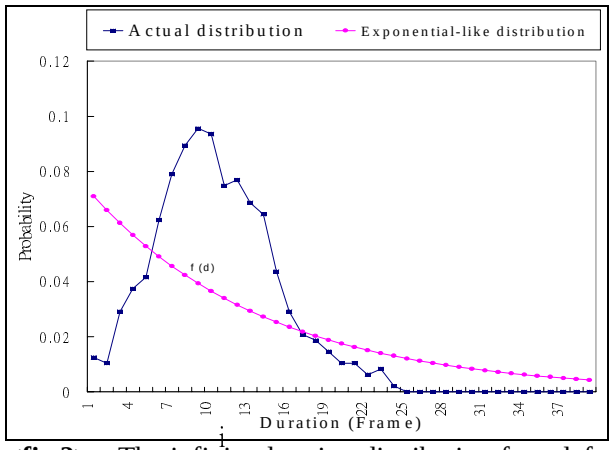
8. TABLES AND FIGURES

	Number of items	Quantity of Recorded speech signal (sec)	Number of Syllables per Word
Lexicon	19,152		2.52
Syllables (Training)	5,271	2,569	1
Words (Training)	3,787	3,705	3.04
Training Data	9,149	6,273	3.04
Testing Data	2,303	1,874	2.49

<table.1> Lexicon and Speech data

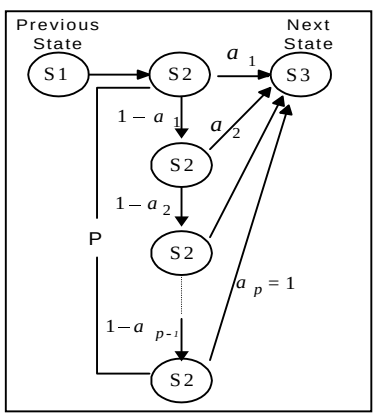


<fig.1> The topology of a typical state in a left-to-right model with infinite state duration distribution

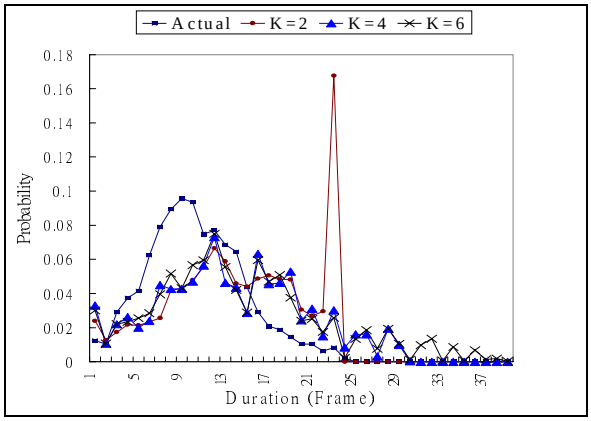


<fig.2> The infinite duration distribution for a left-to-right model and the actual observed finite state duration distribution

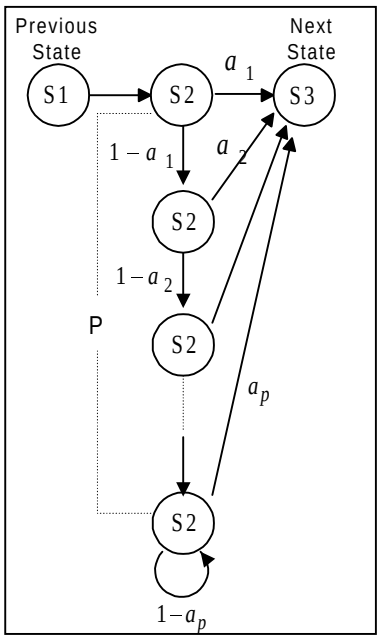
x



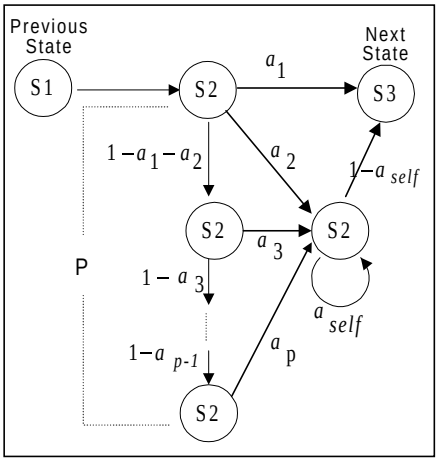
<fig.3> The topology of a particular state in the finite duration model



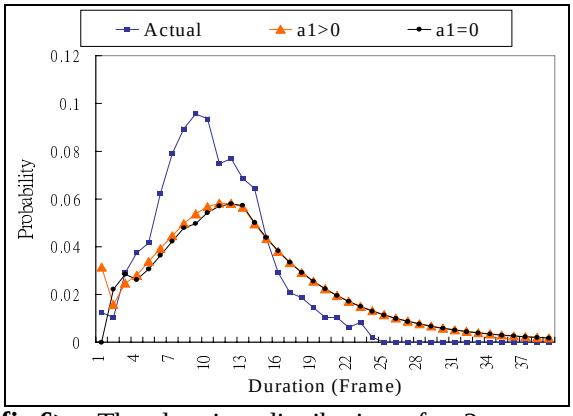
<fig.4> The duration distribution for finite-duration modelin with different *K* values



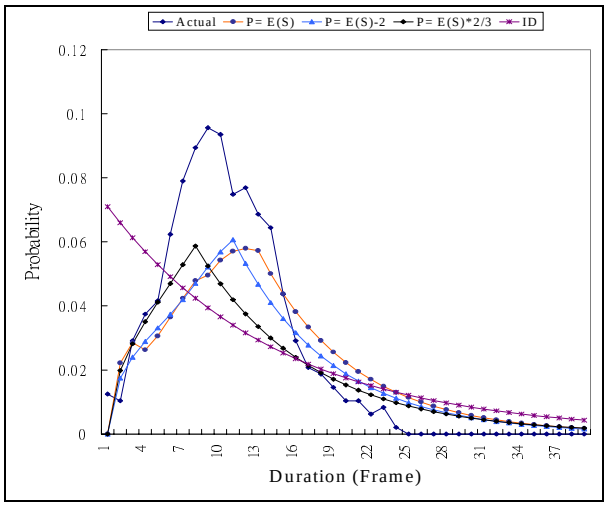
<fig.5a> The topologies of a particular state in First Hybrid Duration Modeling



<fig.5b> The topologies of a particular state in Second Hybrid Duration modeling



<fig.6> The duration distributions for 2 types of hybrid duration models



<fig.7> The duration distribution for hybrid duration models plus minimal duration *m*

System Configurations		WER %	Num. of states	Computation Cost
Conventional HMM Modeling	Inside-syllable	11.4	745	2.77
	Inter-syllable	9.4	745	3.20
Finite Duration Modeling	P = E(S) + 2*stddev(S)	9.5	6,329	10.16
	P = E(S) + 4*stddev(S)	8.1	8,793	12.22
First Hybrid Duration Model	P = E(S)/2*3	8.4	2,706	4.53
Second Hybrid Duration Model	P = E(S) – 2	8.5	3,149	5.21
Plus Minimum Duration	P = E(S) – 2 m = 1	8.2	3,149	3.25
	P = E(S) – 2 m = E(S) / 3	7.9	3,149	2.01

<table.2> Summary performance of different duration modeling

APPENDIX

TAIWANESE INITIALS

	Chinese Character	INITIAL		Chinese Character	INITIAL
1.		b	10.		k
2.		p	11.		nq
3.		m	12.		q
4.		v	13.		z
5.	M	d	14.		c
6.		t	15.		s
7.		n	16.	p	r
8.		l	17.	n	h
9.		g	18.		null initial

TAIWANESE FINALS

	Chinese Character	non-entering FINAL		Chinese Character	entering FINAL
1.		a	48	n	ah
2.		e	49		eh
3.		i	50		ih
4.		o	51		oh
5.		or	52		orh
6.		u	53		uh
7.		ai	54		aih
8.		au	55		auh
9.		ia	56		iah
10		ior	57		iorh
11		iu	58		iuh
12		iau	59		iauh
13		ua	60		uah
14		ue	61		ueh
15		ui	62		uih
16	n	uai	63		uaih
17		ann	64		annh
18		enn	65		ennh
19		inn	66		innh
20	c	onn	67		onnh
21		ainn	68		ainnh
22		aunn	69		aunnh
23		iann	70		iannh
24		ionn	71		ionnh
25.		iunn	72		iunnh
26.		iaunn	73		iaunnh
27.		uann	74		uannh
28.		uenn	75		uennh
29.		uinn	76		uinnh
30.		uainn	77		uainnh
31.		m	78		mh
32.		ng	79		ngh
33.		am	80		ap
34.	w	an	81		at
35.		ang	82	U	ak
36.		om	83		op
37.	L	ong	84	c	ok
38.		im	85		ip
39.	]	in	86	@	it
40.		ing	87		ik
41.		iam	88		iap
42.		en	89		et
43.		iang	90		iak
44.		iong	91	|	iok
45.		un	92		ut
46.		uan	93		uat
47.		uang	94		uak

Note: Each Chinese character contains the corresponding INITIAL or FINAL in each cell of the tables. For those FINALs which have no commonly agreed written scripts, we use the symbol " " to represent them.