Fast and Accurate Recognition of Very-Large-Vocabulary Continuous Mandarin Speech for Chinese Language with Improved Segmental Probability Modeling

Jia-lin Shen[1] and Lin-shan Lee[1,2]
1. Dept. of Electrical Engineering, National Taiwan University
2. Information Science, Academia Sinica
Taipei, Taiwan, R.O.C.
xshen@speech.ee.ntu.edu.tw

Abstract

This paper presents a fast and accurate recognition of continuous Mandarin speech with very large vocabulary using an improved Segmental Probability Model(SPM) approach. In order to extensively utilize the acoustic and linguistic knowledges to further improve the recognition performance, a few special techniques are thus developed. Preliminary simulation results show that the final achievable rate for the base syllable recognition with the improved Segmental Probability Modeling is as high as 91.62%, which indicates a 18.48% error rate reduction and more than 3 times faster than the well-studied sub-syllable-based CHMM. Also, a tone recognizer and a word-based Chinese language model are included and the achieved recognition accuracy for the finally decoded Chinese characters is 92.10%.

1 Introduction

Chinese language is not alphabetic and input of Chinese characters into computers is still difficult. Although there exist almost uncountable number of words in Chinese language, a nice characteristic of the language is that each word is composed of one to several characters which are all monosyllabic, and the total number of phonologically allowed Mandarin syllables is only 1345. Also, Mandarin Chinese is a tonal language. There exist 4 lexical tones and 1 neutral tone. These 1345 Mandarin tonal syllables can be reduced to 408 base syllables disregarding the tones. Since the tones can be separately recognized using the primarily pitch contour information, fast and accurate recognition of the 408 Mandarin base syllables becomes the key problem for Mandarin speech recognition with very large vocabulary. Hidden Markov modeling (HMM) of sub-syllabic units has been found very useful in this problem [1], but here a different approach called Segmental Probability Modeling (SPM) appropriately utilizing the monosyllabic nature of Mandarin speech is investigated in detail and improved performance was obtained.

SPM was first proposed for the recognition of isolated Mandarin base syllables[2]. This model is very similar to continuous hidden Markov model (CHMM) except that the state transition probabilities are deleted and the states equally segment the syllable. In order to extend the applications of SPM to continuous speech recognition, the concatenated syllable matching(CSM) algorithm was previously developed[3], which has the following form :

$$T[y-1] = T[x-1] + \max_i [S_i(x, y-1)] \qquad (1)$$

where T[u] is the accumulated score at a point u, S_i(u,v) is the score when the SPM for the syllable i was matched with utterance section (u,v).

In the present research, a few special techniques are developed to further improve the SPM-based continuous Mandarin speech recognition with very large vocabulary. First, a modified SPM (MSPM) is proposed for better modeling the intra-syllabic and inter-syllabic acoustics and coarticulation. Secondly, the fundamental quefrency (or the first cepstrum coefficient) dips are found very useful in the detection of the syllable boundaries in the CSM algorithm. Thirdly, a non-uniform alignment(NUA) and a segmental weighting (SW) processes are developed to further improve the recognition accuracy. Finally, a syllable filter based on linguistic knowledge is applied to eliminate some impossible syllable candidates considering the linguistic admissible transition between syllables. Preliminary experimental results show that these techniques can improve the recognition accuracy step by step and the final achievable recognition rate can be as high as 91.62%, which indicates a 18.48% error rate reduction and more than 3 times faster than the well-studied sub-syllable-based CHMM. Besides, integrating the tone recognition and the linguistic processing, 92.10% recognition accuracy for the finally decoded Chinese characters is achieved.

This paper is organized as follows. In Section 2, the proposed MSPM is discussed. Then, the dips in the fundamental frequency contour for CSM algorithm is evaluated in Section 3. The NUA and SW processes for improving the MSPM's are discussed in Section 4. In Section 5, the syllable filter integrating linguistic knowledge is presented. The tone recognition and linguistic processing are described in Section 6. The experimental results are performed and analyzed in section 7. Section 8 finally makes the concluding remarks.

0-7803-3192-3/96 $5.00©1996 IEEE

2 Modified SPM (MSPM)

Conventionally each Mandarin syllable is decomposed into an INITIAL/FINAL format. Here INITIAL means the initial consonant of a syllable and FINAL means the vowel (or diphthong) part but including possible media and nasal ending. There exist 22 context independent(CI) INITIAL's and 41 context independent(CI) FINAL's. The 22 CI INITIAL's can be further expanded to 113 context dependent (CD) INITIAL's considering the beginning phoneme of the following FINAL's. It has been found that these 113 CD INITIAL's and 41 CI FINAL's give very good results for Mandarin speech recognition[1]. A segment sharing concept for SPM has also been proposed before[4] and found very useful in the present problem, in which the first few segments of the SPM's for the syllables having the same CD INITIAL's share similar characteristics thus can share the same segments of the models and so do the remaining segments of the models for the syllables having the same CI FINAL's.

Furthermore, in order to include a "transition segment" modeling the transition from the FINAL of a syllable to the INITIAL of the next syllable, a total of 41*(22+1)=943 transition segments will be needed if all syllable transitions are considered. Instead, here all the possible ending phonemes of FINAL's can be classified into 12 categories as shown in Table 1.(a), while the INITIAL's can be classified into 7 or 11 classes as shown in Table 1.(b) and (c). In this way, the number of transition segments is reduced to 12×(7+1)=96 or 12×(11+1)=144 respectively. These segment-shared SPM's plus the transition segments constitute the modified SPM (MSPM) proposed here in this paper, for better modeling the intra-syllabic and inter-syllabic acoustics and coarticulation, as shown in Fig.1. The first few segments are used to model the 113 CD INITIAL's, the following several segments for the 41 CI FINAL's and the last segment is the transition segment.

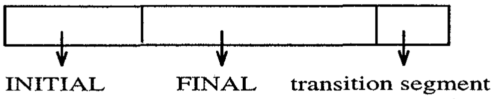

Figure 1: the structure of the modified SPM (MSPM).

3 Fundamental Quefrency Dips

In the CSM algorithm, the dips in the energy contour are first used to predict the possible syllable beginning frames in an utterance[1][3]. However, the syllable boundary detection in continuous speech using energy dips is usually unstable and coarse because of the co-articulation effect in continuous speech. In this paper, the dips in fundamental quefrency contour are found very useful in the detection of syllable boundaries which can be used in place of the energy dips. This is because of the INITIAL/FINAL characteristics of Mandarin Chinese which will be discussed as belows.

(a)

Ψ	Ψ^*	a	o	e	i	eh	u	yu	n	ng	er

(b)

category		
plosives		b,p,d,t,g,k
nasals or liquids		m,n,l
affricates		tz,ts,j,ch,ji,chi
fricatives		f,s,sh,h,r,shi
NULL INITIAL's (followed by)	front vowels	i,e,yu
	mid vowels	eh
	back vowels	a,o,u

(c)

category		
bilabiel		b,p,m
labio-dental		f
dental		d,t
alveolar		n,l
velar		g,k,h
palatal		ji.chi,shi
retroflex		j,ch,sh,r
dental and alveolar		tz,ts,s
NULL INITIAL's (followed by)	front vowels	i,e,yu
	mid vowels	eh
	back vowels	a,o,u

Table 1: (a)The 12 classes of the ending phonemes of FINAL's, and (b), (c) the classification of the INITIAL's into 7 and 11 classes.

Suppose an all-pole filter H(z) of order P is used to represent the system transfer function of speech, which has the form,

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} = \frac{1}{\prod_{k=1}^{P}(1 - z_k z^{-1})} \qquad (2)$$

where $\{a_k\}$, k=1...P, are the linear prediction coefficients while $\{z_k\}$, k=1...P, are the poles corresponding to the modes of the linear system of speech. As a consequence, the relationship between $\{a_k\}$ and $\{z_k\}$ can be derived [5] :

$$a_k = z_1^k + z_2^k + ... + z_P^k \qquad k = 1...P \qquad (3)$$

The fundamental quefrency c_1 is the first term of the lpc-derived cepstrum and can be ,thus, obtained,

$$c_1 = a_1 = z_1 + z_2 + ... + z_P \qquad (4)$$

Because $\{a_k\}$ are all real, the poles $\{z_k\}$ are either real or complex conjugate pairs. Therefore, c_1 can be reduced to the sum of the projection on the real axis for all poles in Z-domain. It is noted that when the poles occur in the right half plane of Z-domain, higher value of c_1 can be obtained, while the poles in left half plane imply lower value of c_1. Since the spectral peaks for FINAL's usually locate in the low frequency part while those for INITIAL's usually locate in high frequency part, some falling gaps in the fundamental quefrency contour will occur in the transition from FINAL to INITIAL. In addition, since the INITIAL's are much shorter in the whole syllable such that the majority part of INITIAL's is co-articulated with the following FINAL's, the fundamental quefrency contour will immediately rise from INITIAL to FINAL. In other words, there exist some fundamental quefrency dips in the inter-syllable transition boundaries. The dips in the fundamental quefrency contour are believed to be more accurate and stable than that in the energy contour due to the INITIAL/FINAL characteristics of Mandarin speech.

4 Non-uniform Alignment and Segmental Weighting

In SPM's, the stochastic state transition behavior in HMM's is replaced by a deterministic process, i.e., uniform segmentation. However, the INITIAL part and the transition segment in a syllable are usually much shorter than the FINAL part, but very important for the recognition of Mandarin syllables. As a result, a non-uniform alignment(NUA) process instead of uniform segmentation is developed to divide the syllable section into segments in an utterance using a nonlinear function. This non-linear function is designed such that the INITIAL and transition parts occupy less length of the whole syllable, which has the form :

$$g(n) = \frac{n}{N+\alpha(n)} \times L \qquad n = 1, ..., N \quad (5)$$

where g(n) is the ending point for segment n, N is the total number of segments, and L is the length of this syllable section. Also, α(n) is a non-negative mono-decreasing function except that $\alpha(N-1)$ = -α(1). Apparently, $\alpha(N)$ equals to zero such that g(N) = L.

Then, a nonlinear segmental weighting (SW) function is used to emphasize the likelihood score of the most discriminative parts, i.e. the INITIAL and transition parts. This segmental weighting function is composed of N elements, i.e. $\{w_1, w_2, ..., w_N\}$, where each w_i is a constant positive value. As a consequence, the syllable section score $S_i(u, v)$ for syllable i in eq.(1) integrating the NUA and SW processes can be expressed as :

$$S_i(u,v) = K \sum_{j=1}^{N} w_j d_j(u + g(j-1), u + g(j); \lambda_{i,j}) \quad (6)$$

where $K = \frac{g(N)}{\sum_{j=1}^{N} w_j(g(j)-g(j-1))}$ is a normalization factor, w_j is the weighting factor for segment j and $d_j(a, b; \lambda_{i,j})$ represents the segmental probability of segment j between the section (a,b) when matching with the model $\lambda_{i,j}$.

5 Syllable Filter

In order to integrate some linguistic knowledge to further improve the recognition performance, a syllable filter is finally applied to eliminate some illegal syllable candidates such that a more accurate syllable path can be obtained. In addition, the syllable filter can increase the number of correct candidates to improve the linguistic processing. Here a syllable bigram is derived to describe the linguistic admissible transition from one syllable to another. The syllable bigram used here is trained from a large Chinese text corpus which consists of a total of 4.2M characters (2.7M words) collected from daily newspapers. Therefore, the syllable pairs with higher probability in the syllable bigram imply stronger linguistically connection between them. Therefore, eq.(1) can be replaced by the following form:

$$T_k[y-1] = \max_{i,1 \le l \le m}{}^{k}[T_l(x-1) + S_i(x, y-1) + \eta P_{h_{x-1}(l),i}](7)$$

where max^k means the k-th highest score in the 408×m accumulated probabilities, $h_{x-1}(l)$ is the top l syllable candidate in the frame point x-1 and $P_{h_{x-1}(l),i}$ is the transition probability from syllable $h_{x-1}(l)$ to i. Also, η is a weighting factor to emphasize the syllable filter probabilities. In other words, for each possible ending point, the top m candidates must be calculated. It is clear that the recognition process integrating these syllable transition information is time-consuming. Instead, the syllable filter can be added in the post processing in which comparable improvement in the recognition accuracy can be achieved with much higher recognition speed.

6 Tone Recognition and Linguistic Processing

From eq(1), for each syllable section in an utterance, not only the base syllable recognition is evaluated, but the tones can be recognized in the same phase[1]. Therefore, the final output in the acoustic processing is the tonal syllable recognition result. Here the CHMM is used as the tone recognizer with a total of 5 models each for one tone. Combining the base syllable and tone recognition results, a tonal syllable lattice with 10 candidates is first constructed for the linguistic processing. Then the tonal syllable lattice is transformed into a word lattice via a lexical access process. Finally the word-based Chinese language model trained from the Chinese text corpus mentioned previously is used to find out the most possible characters, words and sentences.

7 Experiments and Discussion

The speech database used here for speaker dependent task was produced by two male and two female speakers. Each speaker produced 3 sets of all the 1345 isolated Mandarin tonal syllables and 2 continuous utterances each for 352 phonetically balanced sentences (with a total of 2701 syllables covering all the 1345 Mandarin syllables). Also, 3 paragraphs randomly selected from daily newspapers covering the economics, politics and societ y news separately were produced in continuous mode which is composed of totally 106 sentences or 1215 syllables. For base syllable recognition, Cepstral coefficients of order 14 and the corresponding 14 delta cepstral coefficients are derived from the LPC coefficients and used as feature parameters. Instead, the pitch and energy together with their first and second order delta coefficients are used to form a feature vector with dimension 6 in the tone recognition.

In the following experiments, the 3 sets of 1345 isolated syllables are used in training initial models, the

2×352 phonetically balanced continuous sentences are used in re-estimating the continuous model parameters, and the rest of 3 articles are used in testing. The recognition rates are evaluated as the percentages of correctly recognized syllables minus insertion rates and deletion rates. Moreover, the results here are average of the four speakers.

As shown in Table 2, the experimental result using syllable-based SPM and CSM algorithm provides an accuracy of 61.27% only, but it can be significantly increased to 73.91% with the segment sharing concept. Also, the recognition speed is improved by exactly 4 times as compared to the syllable-based SPM. Furthermore, the modified SPM (MSPM) with the transition segment representing the inter-syllable transitions is performed in experiments 3 and 4. The experiments with 7 and 11 classes of INITIAL's are tested respectively where the error rates are further reduced by 10.43% and 7.82% in comparison with the segment shared SPM with slightly increased recognition time. The above experimental results are obtained using the energy dips in the CSM algorithm. However, when the full search mode is applied, i.e., every frame is the possible beginning syllable point, the recognition rates can be immediately increased from 76.63% to 83.70% with more than 12 times of recognition time as also listed in Table 2. Now when the energy dips are replaced by the fundamental quefrency dips, the recognition complexity is almost unchanged but the accuracy can be significantly improved to 83.59%. Then, the proposed non-uniform alignment(NUA), the segmental weighting (SW) function and syllable filter are added in the segment shared SPM using CSM algorithm. It can be found that the recognition rates can be improved step by step and finally to as high as 89.53% and the time needed to recognize a syllable is 0.38 sec in the last row of Table 2.

As a comparison, three types of CHMM's are also tested in Table 3. First, the sub-syllable-based CHMM, with exactly the same 113 CD INITIAL's and 41 CI FINAL's as basic units are tested in experiment 10 [1]. The syllable duration limitation is then considered and the syllable filter is finally added in experiments 11 and 12. It is noteworthy that comparing experiment 9 with 12, 18.48% error rate reduction can be obtained at more than 3 times of recognition speed using the proposed improved SPM techniques.

Finally, we combine a CHMM-based tone recognizer and a word-based Chinese language model with the base syllable recognition to find out the output characters. As shown in Table 4, it can be found that the results for tones are 86.67% and the achieved top 1 and top 10 recognition rates for tonal syllables are 81.10% and 98.97% respectively. The final results for character accuracy with 10 syllable candidates included in the tonal syllable lattice are as high as 92.10%.

8 Conclusion

In this paper, we applied an improved Segmental Probability Model (SPM) to continuous Mandarin speech recognition with very large vocabulary and achieve very good performance both in accuracy and speed. A few special techniques are developed to further improve the recognition performance step by step by making use of the acoustic and linguistic knowledges. The final achievable rate is as high as 91.62%, which indicates a 18.48% error rate reduction and more than 3 times faster than the well-studied sub-syllable-based CHMM. Adding a CHMM-based tone recognizer and a word-based Chinese language model, the achieved recognition accuracy for the finally decoded Chinese characters is 92.10%.

References

[1] Hsin-min Wang, Jia-lin Shen and Lin-shan Lee, "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data", *ICASSP*, pp.61-64, 1995.

[2] Lin-shan Lee, *et.al.* "Golden Mandarin(II) - An Improved Single-Chip Real-time Mandarin Dictation Machine for Chinese Language with very Large Vocabulary", *ICASSP*, pp.503-506, 1993.

[3] Jia-lin Shen, Hsin-min Wang and Lin-shan Lee, "An Initial Study on A Segmental Probability Model Approach to Large-Vocabulary Continuous Mandarin Speech Recognition", *ICASSP*, pp.133-136, 1994.

[4] Jia-lin Shen, Hsin-min Wang, Renyuan Lyu, Lin-shan Lee. "Incremental Speaker Adaptation Using Phonetically Balanced Training Sentences for Mandarin Syllable Recognition Based on Segmental Probability Models", *ICSLP*, pp. 443-446, 1994.

[5] M. Schroeder, "Direct (nonrecursive) Relationship Between Cepstrum and Linear Predictor Coefficients", *IEEE ASSP*, pp. 297-311, Apr. 1994.

experiments	rate	ins.	dels.	time
1. syllable based SPM	61.27	0.78	0.16	1.12
2.segment shared SPM	73.91	0.0	0.16	0.28
3. MSPM (7 classes)	76.63	0.0	0.16	0.34
4. MSPM (11 classes)	75.95	0.0	0.16	0.36
5. using full search (7 classes)	83.70	0.16	0.16	4.23
6. using fundamental quefrency (7 classes)	83.59	0.0	0.16	0.35
7. plus NUA	85.69	0.0	0.16	0.35
8. plus SW	86.27	0.0	0.16	0.35
9. plus syllable filter	91.62	0.0	0.16	0.38

Table 2: The recognition results for experiments 1-8 using various types of SPM's. The recognition rate is evaluated as $\frac{\mathit{correctly\ recognized}-\mathit{ins.}-\mathit{dels.}}{\mathit{total\ syllables}}\times 100\%$, while the time (sec/syllable) needed is on Sun SPARC20.

experiments	rate	ins.	dels.	time
10.sub-syllabic CHMM	80.39	4.20	0.16	1.14
11. plus duration limitation	84.27	0.08	0.82	1.15
12. plus syllable filter	89.72	0.08	0.58	1.35

Table 3: The recognition results for various types of CHMM's

	tone	tonal syllable		character
		(top 1)	(top 10)	
accuracy	86.67	81.10	98.97	92.10

Table 4: The recognition results for tones, tonal syllables and characters