A New Approach of Using Temporal Information in Mandarin Speech Recognition

Jyh-Her Yang, Yuan-Fu Liao*, Yih-Ru Wang and Sin-Horng Chen
Dept. of Communication Engineering, National Chiao Tung University
*Department of Electronic Engineering, National Taipei University of Technology, Taiwan
schen@mail.nctu.edu.tw

Abstract

In this paper, a new approach of using temporal information to assist in Mandarin speech recognition is discussed. It incorporates two types of temporal information into the recognition search. One is a statistical syllable duration model which considers the influences of 411 base-syllables, 5 tones, 4 position-in-word factors, and 3 position-in-sentence factors on syllable duration. Another is the timing information of modeling three types of inter-syllable boundary including intra-word, inter-word without punctuation mark (PM), and inter-word with PM. The uses of these two types of temporal information are expected to be useful for improving the segmentation accuracies in both acoustic decoding and linguistic decoding. Experimental results showed that the base-syllable/character/word recognition rates were slightly improved for both MATBN and Treebank datbase.

1. Introduction

A real-world speech signal always contains rich temporal information ranging from lower-level information, such as phone/syllable/word duration, to higher-level rhythmic information, such as the final-syllable lengthening of prosodic phrase [1]. The temporal information is known to be very helpful for human beings to understand the speech more easily. However, in automatic speech recognition (ASR), the use of temporal information is still primitive. The most basic approach is to incorporate explicit state/phone/syllable duration models or durational constraints into the recognition search for improving the recognition accuracy [2-4]. Another approach is to invoke an embedded phone/syllable/word segmentation in the recognition search process to provide additional acoustic cues to assist in the recognition [5]. But, in all those studies only lower-level durational information, such as HMM state duration or syllable/word duration, was used. No higher-level temporal information was used.

In this paper, a preliminary study of more sophisticatedly using temporal information to improve the ASR for Mandarin speech is discussed. It first extends the conventional base-syllable duration model to consider the influences of three additional affecting factors including tone, position-in-word, and position-in-sentence. With this extension, some higher-level temporal information is invoked in the recognition search. Secondly, it incorporates explicit timing information into the recognition search via constructing models for three types of inter-syllable boundary. These three types of inter-syllable boundary include intra-word, inter-word without punctuation mark (PM), and inter-word with PM. The timing information is expected to be helpful on the segmentation of correct word sequence in the recognition search.

The organization of the paper is described as follows. Section 2 presents the formulation of the new approach of using temporal information in Mandarin ASR. Experimental results are shown in Section 3. Some conclusions are given in the last section.

2. The Proposed Approach

We consider the criterion of speech recognition

$$\begin{aligned} W^*, \Upsilon^* &= \arg\max_{W,\Upsilon} p(W,\Upsilon|X_s, X_p, \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \\ &= \arg\max_{W,\Upsilon} p(X_s, X_p, W, \Upsilon|\Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \\ &= \arg\max_{W,\Upsilon} p(X_s, X_p|W, \Upsilon, \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l)\, p(W,\Upsilon|\Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \end{aligned} \quad (1)$$

where W is a word sequence candidate which is composed of words and PMs; Υ is a segmentation candidate which is composed of an HMM state sequence Φ_{state} and classes of inter-syllable boundaries Υ_b ; Λ_a, Λ_t, Λ_d and Λ_l denote respectively base-syllable (initial/final) acoustic model (AM), tone model (TM), syllable duration model (DM) and language model (LM); and X_s and X_p represent the spectral feature vector sequence and the prosodic feature vector sequence of the input utterance, respectively. In this study, we consider three classes of inter-syllable boundary including intra-word, inter-word without major PM and inter-word with major PM. Here, only major PMs belonging to {、, ，, 。, ；, ：, ？, ！} are considered. We denote them as Intra, Inter and Inter-PM, respectively. The first term in Eq.(1) is generally known as the score of acoustic decoding

and the second one is the score of segmentation and language decoding.

The score of acoustic modeling can be further simplified and expressed by

$$\begin{aligned} & p(X_s, X_p \mid W, \Upsilon, \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \\ & \quad \approx p(X_s, X_p \mid W, \Upsilon, \Lambda_a, \Lambda_t) \\ & \quad = p(X_s \mid W, \Upsilon, \Lambda_a, \Lambda_t, X_p)\, p(X_p \mid W, \Upsilon, \Lambda_a, \Lambda_t) \\ & \quad \approx p(X_s \mid W, \Upsilon, \Lambda_a, X_p)\, p(X_p \mid W, \Upsilon, \Lambda_t) \end{aligned} \tag{2}$$

Here, we assume that LM Λ_l and DM Λ_d are independent of the acoustic decoding. In Eq.(2), the first term is the score of HMM acoustic modeling using spectral features and the second term is the score of prosodic feature decoding. In the current study, we consider 411 base-syllables as the basic acoustic recognition units to further simplify $p(X_s \mid W, \Upsilon, \Lambda_a, X_p)$ as $p(X_s \mid S, \Phi_{state}, \Lambda_a)$, where S is the base-syllable sequence associated with the candidate word sequence W and Φ_{state} is a candidate of HMM state sequence associated with S.

The score of prosodic feature decoding can be further simplified and separated into two terms

$$\begin{aligned} & p(X_p \mid W, \Upsilon, \Lambda_t) \\ & \quad = p(X_t, X_b \mid W, \Upsilon, \Lambda_t) \approx p(X_t \mid T, \Lambda_t)\, p(X_b \mid \Upsilon_b) \end{aligned} \tag{3}$$

where $X_p = (X_t, X_b)$; X_t is the prosodic features for tone recognition; X_b is the prosodic features for inter-syllable boundary classification; $p(X_t \mid T, \Lambda_t)$ is the score of tone decoding; T is the tone sequence associated with W; $p(X_b \mid \Upsilon_b)$ is the score of inter-syllable boundary classification. In this study, X_t consists of 18 parameters including 9 parameters representing, respectively, $F0$ means, $F0$ slopes, and energy means of three uniformly-segmented pitch contour segments of the current syllable; 3 parameters of the last pitch contour segment of the preceding syllable; 3 parameters of the first pitch contour segment of the succeeding syllable; 2 representing pause durations preceding and following the current syllable; and one representing the duration of the current syllable. And X_b consists of the pause duration, the pitch mean and energy level jumps of the preceding and succeeding syllables, and the lengthening factor of the preceding syllable. Here, both $p(X_t \mid T, \Lambda_t)$ and $p(X_b \mid \Upsilon_b)$ are implemented by the neural network-based approach. In each case, a three-layer MLP (multi-layer perceptrons) is employed to generate output discrimination functions for all its classes. We can use these output discrimination functions to perform classification by choosing the class with maximum output as the recognized one. This can check the effectiveness of the MLP classifier. For this application, we transform them into the likelihood scores by

$$P(X \mid \text{Class } i) = \frac{P(\text{Class } i \mid X)}{\sum_k P(\text{Class } k \mid X)} \tag{4}$$

The score of segmentation and language decoding, which is the second term of Eq.(1), can also be further simplified and expressed by

$$\begin{aligned} & p(W, \Upsilon \mid \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \\ & \quad \approx p(W, \Upsilon \mid \Lambda_d, \Lambda_l) \\ & \quad \approx p(\Upsilon \mid W, \Lambda_d, \Lambda_l)\, p(W \mid \Lambda_d, \Lambda_l) \\ & \quad \approx p(X_d \mid W, \Lambda_d)\, p(W \mid \Lambda_l) \end{aligned} \tag{5}$$

where $p(X_d \mid W, \Lambda_d)$ is the score of syllable duration modeling, X_d is the syllable duration sequence derived from the segmentation information Υ, and $p(W \mid \Lambda_l)$ is the score of language decoding. Here, we assume that both acoustic model Λ_a and tone model Λ_t are independent of the segmentation and language decoding. In this study, a word-bigram model Λ_l is used.

The syllable duration model adopted is a simple multiplicative model [9] which involves 4 major affecting factors including base-syllable, tone, position-in-word, and position-in-sentence. In the model, the observed duration of syllable n is expressed by

$$X_d[n] = Z_d[n]\, \gamma_{sy_n} \gamma_{t_n} \gamma_{wp_n} \gamma_{sp_n} \tag{6}$$

where $Z_d[n]$ is the normalized (or residue) syllable duration and is modeled by a normal distribution $N(\mu, \sigma^2)$ with mean μ and variance σ^2; γ's are affecting factors; sy_n, t_n, wp_n and sp_n represent, respectively, the base-syllable, tone, word position, and sentence position of syllable n. In the study, we consider 411 base-syllables, 5 tones, 4 types of position-in-word, and 3 types of position-in-sentence. The 4 types of position-in-word are mono-syllabic word, and the beginning, intermediate and ending syllables of a word. The 3 types of position-in-sentence are the beginning, intermediate and ending syllables of a sentence which is ended with a major PM.

An iterative sequential optimization procedure is employed to train the syllable duration model. It first initializes the training by estimating all affecting factors independently, i.e.,

$$\gamma = \frac{\sum_{n=1}^{N} X_d[n]\delta(\gamma_n, \gamma)}{\mu \sum_{n=1}^{N} \delta(\gamma_n, \gamma)} \quad (7)$$

for $\gamma = \gamma_{sy_n}$, γ_{t_n}, γ_{wp_n}, or γ_{sp_n},

$$\mu = \frac{\sum_{n=1}^{N} \frac{X_d[n]}{\gamma_{sy_n}\gamma_{t_n}\gamma_{wp_n}\gamma_{sp_n}}}{N} \quad (8)$$

and

$$\sigma^2 = \frac{\sum_{n=1}^{N} \left(\frac{X_d[n]}{\gamma_{sy_n}\gamma_{t_n}\gamma_{wp_n}\gamma_{sp_n}} - \mu\right)^2}{N}, \quad (9)$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta function and N is the total number of training syllables. It then sequentially estimate the four types of affecting factors, γ_{sy_n}, γ_{t_n}, γ_{wp_n} and γ_{sp_n}, one-by-one based on the ML (maximum likelihood) criterion with objective function

$$L = \sum_{n=1}^{N} \log f(X_d[n]) \quad (10)$$

where

$$f(X_d[n]) = N\left(\frac{X_d[n]}{\gamma_{sy_n}\gamma_{t_n}\gamma_{wp_n}\gamma_{sp_n}}; \mu, \sigma^2\right) \quad (11)$$

μ and σ^2 are also updated using Eqs.(8) and (9), respectively. The sequential optimization step is iteratively executed until a convergence is reached.

To reduce the computational complexity, a two-stage recognition search is adopted in this study. In the first stage, a word-lattice which consists of top-10 candidate words is constructed by using only the acoustic model Λ_a and the word-bigram LM Λ_l. Then, in the second stage the best word sequence is determined from the word-lattice by using the criterion shown in Eq.(1). The two-stage recognition search is realized using the HTK toolkit [8].

3. Experimental Results

Performance of the proposed approach was examined by simulations using two databases. One was the Anchor set of MATBN (Mandarin Chinese Broadcast News Corpus) [6]. It was uttered by 4 anchors in fast speaking styles and is composed of 175,194 training syllables and 14,906 testing syllables. The acoustic models consisted of 100 3-state right-*final*-dependent (RFD) *Initial* HMM models, 40 5-state context-independent (CI) *Final* models, 19 3-state Particle models, one 3-state Breath model, one 3-state Silence model, one 1-state Short Pause model (tied with the middle state of Silence model), and 3 3-state Garbage models. Another database was the read-speech database of Sinica Treebank [7]. It was uttered by a single female announcer in a normal speed. It was composed of 380 utterances with 52,192 syllables. The acoustic models consisted of 100 RFD *Initial* models and 40 CI *Final* models.

For LM, a general bigram LM was first trained using the following three corpora: (1) Sinorama: a news magazine with 9.87 million words; (2) NTCIR: an IR test bench consisting of several domain with 124.4 million words; and (3) Sinica Corpus: general text corpus collected for the language analysis with 4.8 million words. Here a 60,000-word lexicon was used. For the recognition of MATBN, the general LM was adapted using the texts of MATBN which was composed of 1.31 million words with 23,314 particles and 90,052 breathes.

We first examined the syllable duration model. Table 1 shows some affecting factors (AF) for the two databases. They include: (1) mono-syllabic word (MW), and the beginning (BW), intermediate (IW) and ending syllables (EW) of a word for position-in-word; (2) 5 tones; and (3) the beginning (BS), intermediate (IS) and ending syllables (ES) of a sentence for position-in-sentence. It can be found from the table that IW in position-in-word, Tone 5, and BS in position-in-sentence are much shorter; while ES in position-in-sentence is very long.

Table 1: Some affecting factors (AF) of the syllable duration model for the two databases.

AF / Database	Position-in-word				Position-in-sentence		
	MW	BW	IW	EW	BS	IS	ES
MATBN anchor	1.05	0.97	0.84	1.02	0.85	0.98	1.34
Sinica Treebank	1.05	0.96	0.88	1.03	0.90	0.99	1.20
	Tone						
	T1	T2	T3	T4	T5		
MATBN anchor	1.01	1.05	0.98	1.02	0.73		
Sinica Treebank	1.03	1.07	1.00	1.02	0.72		

We then examined the performances of the MLP tone classifier (see Table 2) and MLP inter-syllable boundary classifier (see Table 3). It can be found from Table 2 that tone recognition rates of 75.1 and 85% were achieved for MATBN and Treebank, respectively. Both Tone 1 and Tone 4 were easier to be recognized while Tone 3 and Tone 5 were not. It can also be found from Table 3 that accuracy rates of 58.8% and 69.1% were achieved in the inter-syllable boundary classifications for MATBN and Treebank databases, respectively. The class of inter-word with PM was easier to be correctly detected.

Table 2: Performance of the tone recognizers. (unit: %)

	T1	T2	T3	T4	T5	average
MATBN Anchor	77.7	74.3	66.3	83.4	42.0	75.1
Sinica Treebank	88.0	84.4	70.8	92.6	74.9	85.0

Table 3: Experimental results of the inter-syllable boundary recognizer. (unit: %)

	Intra	Inter	Inter-PM	average
MATBN Anchor	51.0	64.0	70.6	58.8
Sinica Treebank	78.8	57.3	81.9	69.1

Lastly, we examined the performance of the proposed method of using temporal information in Mandarin speech recognition. Table 4 displays the experimental results. It can be found from Table 4 that the baseline system which used the acoustic and language models, Λ_a and Λ_l performs well. Base-syllable/character/word recognition rates were 93.49/91.04/86.29 and 94.01/84.99/75.43 for the MATBN anchor and Sinica Treebank databases, respectively.

Table 4: The experimental results of the proposed method for Mandarin ASR. (unit: %)

		Syllable Recogniti on. rate	Character Recogniti on. rate	Word Recog. rate
MATBN Anchor	Baseline	93.49	91.04	86.29
	Baseline +tone recogn.	93.59	91.15	86.51
	Proposed	93.66	91.23	86.62
Sinica Treebank	Baseline	94.01	84.99	75.43
	Baseline +tone recogn.	93.89	85.41	75.73
	Proposed	94.00	85.55	75.93

It is noted that both character and word recognition rates for Treebank were relatively low as compared with those of MATBN because Treebank contained much more proper nouns and DM compound words which were treated as individual characters rather than words. The performances were slightly improved as we incorporated the tone recognizer. The performances were further improved for the proposed method as we used the temporal information in the recognition search.

4. Conclusions

A new approach of using a statistical syllable duration model and an inter-syllable boundary model to assist in Mandarin ASR has been discussed in this paper. Experimental results showed that it slightly outperformed the baseline system. Further studies include an analysis of its effectiveness on different type of pronunciation conditions, the use of more sophisticated temporal models, and so on.

ACKNOWLEDGEMENTS

This work was supported in part by MOE under contract EX-94-E-FA06-4-4 and in part by NSC under contract NSC94-2213-E009-020. The authors want to thank the MediaTek for supporting our research and the Academia Sinica, Taiwan for providing the Sinica Treebank Corpus.

REFERENCES

[1] Tseng, Chiu-yu and Lee, Yeh-lin (2004). "Speech rate and prosody units: Evidence of interaction from Mandarin Chinese," *Proceedings of the International Conference on Speech Prosody 2004*, (Mar. 23-26, 2004), Nara, Japan, 251-254.

[2] L. Rabiner and B.-H. Juang. Fundamentals of Speech Recognition. Prentice-Hall, 1993, pp 384-385.

[3] A. Anastasakos, R. Schwartz, and H. Shu, "Duration modeling in large vocabulary speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 628–631.

[4] C. Mitchell, M. Harper, L. Jamieson, and R. Helzermam, "A parallel implementation of a hidden Markov model with duration modeling for speech recognition," in *Digital Signal Process.*, vol. 5, 1995, pp. 43–57.

[5] W. J. Wang, Y. F. Liao and S. H. Chen, "RNN-based Prosodic Modeling for Mandarin Speech and Its Application to Speech-to-Text Conversion", Speech Communication, 36 (2002), pp.247-265.

[6] Hsin-min Wang, "MATBN 2002: A Mandarin Chinese Broadcast News Corpus" ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003).

[7] C. R. Huang, K. J. Chen, F. Y. Chen, Z. M. Gao and K. Y. Chen. 2000, "Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface", Proceedings of 2nd Chinese Language Processing Workshop 2000, Hong Kong, pp. 29-37.

[8] Hidden Markov Model Toolkit (HTK) , http://htk.eng.cam.ac.uk

[9] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, " A New Duration Modeling Approach for Mandarin Speech," IEEE Trans. On Speech and Audio Processing, vol. 11, no. 4, July 2003.