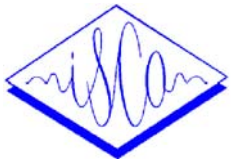

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Prosodic Modeling of Mandarin Speech and Its Application to Lexical Decoding

Wern-Jun Wang , Yuan-Fu Liao and Sin-Horng Chen

Department of Communication Engineering, National Chiao Tung University

Applied Research Laboratory, Chunghwa Telecommunication Laboratories
Tel: +886-3-5731822, Fax: +886-3-5710116, Email: schen@cc.nctu.edu.tw
Tel:+886-3-4244536, Fax:+886-3-4244167, Email: wjwang@ms.chttl.com.tw

10.21437/Eurospeech.1999-180

ABSTRACT

In this paper, a new RNN-based prosodic modeling method for Mandarin speech recognition is proposed. It is performed in the post-processing stage of the acoustic decoding aiming at detecting word boundaries for assisting in the lexical decoding. It employs a simple RNN to learn the relationship between input prosodic features, extracted from the input utterance with syllable boundaries provided by the preceding acoustic decoding, and output information related to word boundaries. Simulations on a large single-speaker database were performed to evaluate the proposed method. Experimental results showed that 71.9% of word tags and 95.3% of punctuation mark (PM) tags could be correctly detected. By incorporating the prosodic model into an HMM-based continuous Mandarin speech recognition system, the character recognition rate increased from 73.6% to 74.7% with a reduction of 17% on the computational complexity. So the proposed prosodic modeling method is helpful for speech recognition.

1. INTRODUCTION

Prosody is an inherent supra-segmental feature of human's speech. It controls the stress, intonation pattern, and timing structure of continuous speech which, in turn, decide the naturalness of the speech [1]. In the past, prosodic information was rarely used in speech recognition. But it became an interesting research issue in recent years. One approach of prosodic modeling is to use a statistical model, such as decision-tree or hidden Markov model (HMM), to detect prosodic phrasal boundaries and word prominence [1][2]. The model can be trained using a large speech database with properly tagging to provide major and minor breaks of prosodic phrases as well as prominence levels of words. A well-known prosody tagging system is the Tones and Break Indices (TOBI) system [3] that labels prosodic phrasal boundaries by a seven-level scale. Two problems of this approach can be found. One is that the labor-intensive tagging work for preparing a large training database must be done by linguistic experts. Besides, the consistency in labeling is difficult to maintain over the whole database. The other difficulty is that it needs to further explore the relationship between the detected prosodic phrasal information and the syntactic structure in order to properly incorporate it into the following linguistic decoding. Due to the difficulty of solving these two problems, another approach which directly uses syntactical features of the text associated with the input utterance as the output targets to model was proposed recently [4]. One problem of this approach is that the syntactical phrase structure is not completely matched with the prosodic phrase structure. It may result in the inaccuracy of prosodic modeling and hence degrade its usability in the following linguistic decoding.

Prosodic modeling is even more important for Mandarin speech recognition because Mandarin Chinese is a monosyllabic tonal language. Each character is pronounced as a syllable with a tone associated to it. Although syllable is the basic pronunciation unit, word, which consists of one to several characters, is the smallest meaningful unit. Due to the fact that no special marks (except punctuation marks) are used to delimit word boundaries, disambiguate word boundaries is a problem in Mandarin speech recognition. This problem is conventionally solved in the lexical decoding using the acoustic decoding results and a statistical model-based language model. But, due to the fact that human beings rely mainly on the prosodic information in their word perception, prosodic modeling may provide more useful cues to solve this problem. In the past, there were very few studies related to the prosodic modeling for Mandarin speech recognition [5].

In this paper, a new RNN-based prosodic modeling approach for Mandarin speech recognition is proposed. It is performed in the post-processing stage of the acoustic decoding aiming at detecting word boundaries for assisting in the following lexical decoding. Compared with the previous studies, the proposed method has a distinct property of using word-level linguistic features as the targets to model instead of using the conventional multi-level prosodic indices like the TOBI system. This leads to several advantages of the proposed method. First, it is easier to incorporate the prosodic model into the lexical decoding by directly taking its outputs as additional scores or by using its outputs to set some path constraints in the lexical decoding search. Second, no complicated syntactic analyses are needed. Third, it is very easy to prepare a large training database without the help of linguistic experts.

2. THE RNN PROSODIC MODEL

The proposed prosodic model uses an RNN to detect information related to word boundaries from the input prosodic features extracted from the vicinity of the current syllable. The RNN is a three-layer network with all outputs of the hidden layer being fed back as additional inputs. It is a dynamic system suitable for realizing a complex mapping between the context of the current input prosodic features and the output word-boundary information. The RNN can be trained using a large speech database by the back propagation

through time (BPTT) algorithm. Input features includes prosodic features extracted from each training utterance with all syllable boundaries being given by the preceding acoustic decoding, while output targets are word-boundary information extracted from the word sequence of the associated text tagged by a statistical model-based method. After properly training, the RNN can then be used in the testing phase to generate word-boundary information using the prosodic features extracted from the testing utterance.

In this study, two types of prosodic features are extracted from the vicinity of the current syllable. One is the local features of the current syllable including: (1) the mean and slope of its pitch contour, (2) the means of log-energy and normalized log-energy of its *final* segment, and (3) the normalized duration. Here the normalization is performed with respect to the *final* type of the current syllable. The other is the contextual features of the current syllable including: (1) 2 flags indicating whether the current syllable is, respectively, the beginning and ending syllables of a sentence, (2) 2 values showing the durations of the non-pitch segments between the current pitch contour and its two nearest neighbors, (3) 2 normalized silence durations besides the current syllable, (4) 2 pitch mean differences and (5) 2 log-energy mean differences between the current syllable and its two nearest neighbors. Here the two normalized values in (3) are processed according to the *initial* sub-groups of the current and succeeding syllables. There are in total 15 prosodic features extracted from the vicinity of the current syllable. The RNN takes 7 sets of prosodic features of the current syllable and its 6 nearest neighboring syllables as inputs.

The RNN uses 8 output linguistic features. They include 4 flags indicating whether the current syllable is a monosyllabic word or is the beginning syllable, the intermediate syllable, or the ending syllable of a polysyllabic word; and 4 flags indicating whether there exist a PM in the left side, a PM in the right side, two PMs in both sides, or no PMs in both sides of the current syllable. These 8 output features are denoted as MSW, BPSW, IPSW, EPSW, LPM, RPM, BPM, and NPM, respectively. To prepare output targets for training the RNN, texts associated with all training utterances are tagged into word sequences in advance. An automatic statistical model-based tagging algorithm based on the long-word-first criterion is first applied. A large lexicon containing about 110,000 words is used in the tagging algorithm. Then, several simple word-merging rules are used to modify the resulting word sequences. Lastly, tagging errors are corrected manually. All output targets to train the RNN are extracted from these well-tagged word sequences.

To check the classification ability of the RNN, the 4 flags showing the location of the current syllable in a word and another 4 flags showing the PM status are separately considered. These two flag sets are referred to as Word-tag and PM-tag, respectively. Two FSMs are used to examine whether the responses of the RNN are good enough to make reliable classifications for these two flag sets. The topology of the Word-tag related FSM was shown in *Figure 1*. For each FSM, when one RNN output is higher than the other three outputs by a threshold Th_d and is also higher than a high threshold Th_h, it moves to the associated state if it is a legal one; otherwise it moves to an uncertain state. These two thresholds are determined empirically.

3. THE INTEGRATED MANDARIN SPEECH RECOGNITION SYSTEM

The complete block diagram of an HMM-based continuous Mandarin speech recognition system incorporating with the prosodic model is shown in *Figure 2*. Input speech is firstly processed to extract acoustic features. An HMM-based base-

Figure 1. The topology of word-tag related finite state machine .

Figure 2. A functional block diagram of the continuous Mandarin speech recognition system.

syllable recognizer with the viterbi-parallel-backtrace [6] DP search algorithm is then used to generate a top-N base-syllable lattice. Based on the segmentation information provided by the top-1 base-syllable sequence, prosodic features are extracted for tone recognition and for prosodic modeling. The tone recognizer also uses a three-layer RNN to discriminate 5 tones for the current base-syllable using input features extracted from the fundamental frequency (F0) contour and energy contours of several base-syllables centered around the current base-syllable. Meanwhile, the prosodic modeling RNN detects the word-boundary information for the current base-syllable. Then, all these three recognition results are fed into the lexical decoder. The lexical decoder uses a statistical language model with a lexicon to find the best word sequence from the top-N base-syllable lattice and the top-2 tone lattice.

The word-boundary information provided by the prosodic model is used in the decoding search to help setting some path constraints. Necessary information of the lexical decoding is described a follows. A backward lexical tree was created from a lexicon with 111243 entries. The number of syllables of each entry in this lexicon is from one to five. Each node in the lexical tree is associated with a tonal syllable. Important information registered on every node includes a flag showing whether the associated syllable is a beginning syllable of a word, the unigram probability of the associated word, and parameters related to homonym words. The statistical language model uses a word unigram model and a word-class bigram model. Both models are obtained via analyzing a 3-million-word corpus provided by Academia Sinica of ROC [7]. We adopt some back-off modification to the word-class bigram model. The words with the same beginning base-syllable or the same ending base-syllable are clustered into the same word-class, respectively.

The lexical decoding uses a Viterbi search to find the best word sequence with maximal scores. The cumulative score combines likelihood scores of all constituent base-syllables, scores (log of RNN responses) of all constituent tones, unigram probabilites of all constituent words, and bigram probabilities of all word-pairs. Two schemes of using the prosodic model are suggested. One is using the word-boundary information provided by the Word-tag related FSM to set path constraints to eliminate unnecessary lexicon tree accesses. The other is to use the outputs of the prosodic modeling RNN as an additional score of the cumulative score.

4. EXPERIMENTAL RESULTS

Effectiveness of the proposed method was examined by simulations on a single-dependent continuous Mandarin speech recognition task using a large single-speaker database. The database contained 452 sentential utterances and 200 paragraph utterances. All utterances were generated by a male speaker. They were all spoken naturally at speed in the range of 3.5-4.5 syllables per second. The database was divided into two parts. The one containing 491 utterances (or 28060 syllables) was used for training and the other containing 161 utterances (or 7034 syllables) was used for testing.

4.1. Results of prosodic modeling

For training and testing the prosodic model, all speech signals were manually segmented into syllable sequences. The pitch contour was detected by the SIFT algorithm and corrected by hand. 15 prosodic features mentioned previously were extracted for each syllable. All texts were also tagged into word sequences. Eight output linguistic features were then extracted for each character. An RNN taking all prosodic features of seven syllables surrounding the current syllable as inputs was then trained using the BPTT algorithm. The number of nodes in the hidden layers was determined empirically and set to be 30.

Tables 1 and 2 show the classification results for Word-tag and PM-tag by the RNN without invoking FSMs. Best output sequence was searched with a simple word-structure network. Experimental results showed that 71.9% of Word-tag and 95.3% of PM-tag can be correctly detected. Further examination of the RNN classifier was performed by replacing the word-structure network with two FSMs with Th_h=0.6 and Th_d=0.3, which additionally included uncertain states for the cases when the RNN did not respond well for making reliable classifications. Performances were improved to 87.6% and 97.4% with 32.3% and 4.8% outputs staying in uncertain states for Word-tag and PM-tag, respectively.

4.2. Results of continuous speech recognition

The effectiveness of incorporating the RNN prosodic model into the lexical decoding process was then examined. The HMM-based base-syllable recognizer generated a top-10 base-syllable lattice. The base-syllable inclusion rate was 96.5%. The tone recognizer generated a top-2 tone lattice. The tone inclusion rate is 98%. The outputs of the base-syllable

Table 1. The confusion matrix for Word-tag classification. Overall classification rate equals 71.9%.

Result / Desired	BPSW	IPSW	EPSW	MSW
BPSW	1861	317	110	52
IPSW	288	1106	341	11
EPSW	135	369	1774	62
MSW	160	28	107	313

Table 2. The confusion matrix for PM-tag classification. Overall classification rate equals 95.3%.

Result / Desired	NPM	LPM	RPM	BPM
NPM	5487	88	101	0
LPM	65	610	1	0
RPM	65	2	609	0
BPM	1	4	1	0

recognizer, the tone recognizer, and the prosodic model were all fed into the lexical decoder. To compare the result obtained by incorporating the prosodic information with the baseline system, three experiments were conducted in this study. The first experiment was to add the RNN outputs of Word-tag as additional scores to the cumulative scores of the lexical decoding. The second experiment was an extension of the first one by additionally utilizing the outputs of the Word-tag FSM to set path constraints in the decoding search. different search strategies were applied for different the five Word-tag states: MSW, BPSW, IPSW, EPSW and the uncertain state. A full search was used for the uncertain state and restrictive searches were used for other four decisive states. More specifically, the lexical search will be active only when the current syllable is MSW, EPSW or uncertain state and the backward word construction processes will stop if the processing syllable is with BPSW or the syllable before the processing syllable is MSW or EPSW. In this way, the complexity of the search space can be reduced.

The third experiment is a modification of the second one. By detail error analysis of the results of the prosodic modeling, we found that most errors could be categorized into two types. One was that long words were prosodically broken into two or three short character sequences. Nevertheless, in most cases these short words were legal words so as to make it no harm to the lexical decoding. The other type of error was merging several short words to form a long prosodic word. In this case, decoding errors might occur if the lexical search neglects the syllables with IPSW state. To overcome the problem, we

relaxed the path constraints of the IPSW state in the third experiment.

The results of baseline system and the three experiments are listed in Table 3. The character accuracy was calculated by considering the insertion, deletion and substitution errors. The complexity counted the number of access to the lexical tree. It can be found from Table 3 that the character accuracy was improved from 73.6% to 74.6% in Experiment 1 by incorporating the Word-tag information into the lexical decoding process of the baseline system. In Experiment 2, the complexity was reduced by 52.4% but the character accuracy decreased to 69.7%. On the contrary, a complexity reduction of 30.5% was obtained in Experiment 3 with a slight increase on the character accuracy.

Table 3. The performance comparisons for different experiments.

Method	Character Accuracy	Complexity Reduction
Baseline	73.6%	X
Experiment 1	74.6%	X
Experiment 2	69.7%	52.4%
Experiment 3	73.9%	30.5%

To more precisely examine the effect of the Word-tag FSM on the performance of Experiment 3, we calculated the reliability score of the four decisive output states for difference sets of thresholds. The reliability score was defined as the conditional probability of correct classified Word-tag given the detected Word-tag. Table 4 lists the reliability scores for three cases including no threshold and two sets of thresholds. It can be found from the table that more restrictive thresholds will result in more reliable Word-tag classification. The side effect of using more restrictive thresholds lies in the increase of the number of syllables staying in the uncertain state. It can also be found from Table 4 that IPSW has the lowest reliability scores in all three cases. This finding gives a strong support to our approach in used in Experiment 3. To further examine the tradeoff between the character accuracy and the complexity, another test of Experiment 3 was lastly performed for the third case of using more restrictive thresholds. A character accuracy of 74.7% was obtained with a complexity reduction of 17.0%. Undeniably, the result is a compromise between accuracy and complexity. It is worth noting that using a more sophisticated language model to solve the homonym problems can further increase the performance. The character accuracy achieved in the current system is equivalent to 82.5% if all homonym errors, which were not solvable by the current word-class bigram and word unigram models, were excluded

Table 4. The comparison of the reliability of the classifications by the Word-tag FSM.

Reliability	No Threshold	Th_h =0.6 Th_d =0.3	Th_h =0.8 Th_d =0.6
BPSW	76.1%	84.0%	91.2%
IPSW	60.8%	69.4%	78.2%
EPSW	76.8%	87.0%	94.8%
MSW	71.5%	79.2%	87.6%

5. CONCLUSION

A new prosodic modeling method for Mandarin speech recognition has been discussed in this paper. Its effectiveness on both Word-tag and PM-tag classifications has been confirmed by simulations on a speaker-dependent speech recognition task. The Word-tag information provided by the prosodic model has also been successfully incorporated into the conventional HMM-based continuous Mandarin speech recognition system to integrate the acoustic and linguistic knowledge. Word boundary hypotheses conjectured via the local and global prosodic features do help in pruning unpromising word boundary and increasing the character accuracy.

6. REFERENCES

[1] C. W. Wightman and M. Ostendorf, "Automatic Labeling of Prosodic Patterns," *IEEE Trans. Speech and Audio Proc.*, Vol.2, No.4, pp.469-480, Oct. 1994.

[2] S. E. Bou-Ghazale and J. H. L. Hansen, "HMM-Based Stressed Speech Modeling with Application to Improved Synthesis and Recognition of Isolated Speech under Stress," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 3, pp.201-216, May 1998.

[3] K. Silverman, M. Beckman, etc., "TOBI: A standard for labeling English prosody," in *Proc. Int. Conf. On Spoken Language Processing (ICSLP)*, Vol. 2, pp. 867-870, Banff, 1996.

[4] A. Batliner, R. Kompe, etc., "Syntactic-Prosodic Labeling of Large Spontaneous Speech Data-Base," in *Proc. Int. Conf. On Spoken Language Processing (ICSLP)*, Vol. 3, pp. 1720-1723, 1996.

[5] H. Y. Hsieh, R. Y. Lyu and L. S. Lee, "Use of Prosodic Information to Integrate Acoustic and Linguistic Knowledge in Continuous Mandarin Speech Recognition with Very Large Vocabulary," in *Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1995.

[6] E. F. Huang and H. C. Wang, "An efficient algorithm for syllable hypothesization in continuous Mandarin speech recognition," *IEEE Trans on Speech and Audio Processing*, July, 1994.

[7] Chinese Knowledge Information Processing Group, "The contents and descriptions of Sinica Corpus," Technical Report no. 95-02, Academia Sinica, September, 1995.