ELSEVIER

Speech Communication 36 (2002) 247–265

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

RNN-based prosodic modeling for mandarin speech and its application to speech-to-text conversion

Wern-Jun Wang [a,b,*], Yuan-Fu Liao [a], Sin-Horng Chen [a,1]

[a] *Department of Communication Engineering, National Chiao Tung University, Taiwan, ROC*
[b] *Advanced Technology Research Laboratory, Chunghwa Telecommunication Laboratories, Taiwan, ROC*

Received 24 December 1999; received in revised form 5 September 2000; accepted 7 November 2000

Abstract

In this paper, a recurrent neural network (RNN) based prosodic modeling method for Mandarin speech-to-text conversion is proposed. The prosodic modeling is performed in the post-processing stage of acoustic decoding and aims at detecting word-boundary cues to assist in linguistic decoding. It employs a simple three-layer RNN to learn the relationship between input prosodic features, extracted from the input utterance with syllable boundaries pre-determined by the preceding acoustic decoder, and output word-boundary information of the associated text. After the RNN prosodic model is properly trained, it can be used to generate word-boundary cues to help the linguistic decoder solving the problem of word-boundary ambiguity. Two schemes of using these word-boundary cues are proposed. Scheme 1 modifies the baseline scheme of the conventional linguistic decoding search by directly taking the RNN outputs as additional scores and adding them to all word-sequence hypotheses to assist in selecting the best recognized word sequence. Scheme 2 is an extended version of Scheme 1 by further using the RNN outputs to drive a finite state machine (FSM) for setting path constraints to restrict the linguistic decoding search. Character accuracy rates of 73.6%, 74.6% and 74.7% were obtained for the systems using the baseline scheme, Schemes 1 and 2, respectively. Besides, a gain of 17% reduction in the computational complexity of the linguistic decoding search was also obtained for Scheme 2. So the proposed prosodic modeling method is promising for Mandarin speech recognition.

Keywords: Recurrent neural network; Prosodic modeling; Speech-to-text conversion; Acoustic decoding; Linguistic decoding

1. Introduction

Prosody is an inherent supra-segmental feature of human's speech. It carries stress, intonation pattern, and timing structure of continuous speech which, in turn, decide naturalness of the speech (Wightman and Ostendorf, 1994). In the past, prosodic information was rarely used in speech recognition. But it became an interesting research issue in recent years. A general statement of the function of prosodic modeling for speech recognition is to explore the prosodic phrasing of the testing utterance for providing useful information to help linguistic decoding in the next stage. The main concern of the prosodic phrasing issue is to

* Corresponding author. Mailing address: 6F, No. 82 Tz-Chung Street, Chung-Li, Taoyuan 320, Taiwan. Tel.: +886-3-4244536; fax: +886-3-4244147.

E-mail addresses: wernjun@ms.chttl.com.tw (W.-J. Wang), schen@cc.nctu.edu.tw (S.-H. Chen).

[1] Tel.: +886-3-5731822; fax: +886-3-5710116.

0167-6393/02/$ - see front matter
PII: S0167-6393(01)00006-1

build a model describing the relationship between input prosodic features extracted from the testing utterance and output linguistic features of the associated text. Two primary approaches of prosodic modeling are based on the boundary-labeling scheme with or without using phonological knowledge. Most methods in the approach using phonological knowledge employ statistical models, such as decision-tree and hidden Markov model (HMM), to detect prosodic phrase boundaries, word prominence, or word accent type (Bou-Ghazale and Hansen, 1998; Wightman and Ostendorf, 1994; Iwano and Hirose, 1998, 1999). These detected cues are used to help resolving syntactic boundary ambiguity (Niemann et al., 1997; Price et al., 1991), reordering *N*-best acoustically decoded word sequences (Hunt, 1994; Kompe et al., 1995), or improving mora recognition for Japanese speech (Iwano and Hirose, 1999). The statistical model used in the approach can be trained using a large speech database with major and minor breaks of the prosodic phrase structure and/or prominence levels of words being given via properly labeling. A well-known prosody labeling system is the Tones and Break Indices (TOBI) system (Grice et al., 1996; Silverman et al., 1992) which labels prosodic phrase boundaries using a seven-level scale. Two main problems of the approach can be found. One is that the prosodic labeling of training utterances must be done by linguistic experts. This is a cumbersome work. Besides, the consistency in labeling is difficult to maintain over the whole database. The other is that it needs to further explore the relationship between labels of prosodic phrase boundary and the syntactic structure of the associated text for properly using the detected prosodic phrasing information in linguistic decoding. The other approach which does not use phonological knowledge directly uses syntactic features of the associated text as the output targets for modeling the prosodic features of the input utterance (Batliner et al., 1996; Kompe et al., 1997; Price et al., 1991; Hirose and Iwano, 1997, 1998). One problem of the approach is that the syntactic phrase structure is not completely matched with the prosodic phrase structure. Most prosodic phrases contain one to several syntactic phrases. In some cases, a long syntactic phrase can split into several prosodic phrases. This mismatch may degrade the accuracy of the prosodic modeling and hence decreases its usability in linguistic decoding. A hybrid approach which takes the prosodic tendency and syntactic compatibility into consideration was also studied recently (Batliner et al., 1996; Kompe et al., 1997). A new prosody-labeling scheme which includes perceptual-prosodic boundaries and syntactic boundaries was developed in such a hybrid approach.

Prosodic modeling is even more important for Mandarin speech recognition as compared with that for other word-based languages such as English. This is because Chinese is a character-based language. A Chinese text is composed of clauses and sentences which are ended with punctuation marks (PMs) such as comma and period. A clause or sentence is formed by concatenating words which consist of one to several characters. Although word is the smallest meaningful unit in syntax, character is the basic pronunciation unit. Each character is pronounced as a syllable with which a tone associates. Due to the fact that there are no special marks used to delimit word boundaries, inter-word syllable boundaries are, in general, not specially emphasized in pronouncing a Chinese text. This makes the determination of word boundary be a problem to be solved in Mandarin speech-to-text conversion. Conventionally, the problem is solved in linguistic decoding using the acoustic decoding results and some statistical language models such as word-class bigram model. But, due to the fact that human beings rely mainly on prosodic information in their word perception, a proper prosodic modeling can surely provide useful cues to assist in solving the problem. In some cases, cues provided by a prosodic model are more efficient than a statistical language model. This is demonstrated by the following example in which S1 and S2 show two output candidate sentences of a linguistic decoding with and without the help of prosodic information for the same acoustically decoded syllable sequence:

S1. 居庸關|是|兵家|必爭之地。(Chu-Yung Kuan is a military strategic frontier pass.)

S2. 居庸關士兵家必爭之地。 (Chu-Yung Kuan soldier's home military strategic frontier pass.)

Here, the vertical bars in S1 indicate the word boundary cues suggested by a prosodic model (to be discussed in Section 2). The sentence S2, which is an incorrect result, is generated by a conventional linguistic decoder based on a language model involving word unigram and word-class bigram probabilities. The sentence S1, which is a correct one, is obtained by our proposed method which incorporates an RNN prosodic model with the above conventional linguistic decoder. In the past, the studies of prosodic modeling for Mandarin speech recognition were few (Bai et al., 1997; Hsieh et al., 1996; Lyu et al., 1995). In the Golden Mandarin dictation machine (Lyu et al., 1995), the concept of breath group was used to make the recognition operate on a prosodic-segment-by-prosodic-segment mode. Bai et al. (1997) used pitch and energy features to detect possible syllable and word boundaries for pruning unnecessary path searches in keyword recognition. Hsieh et al. (1996) used energy and pitch information to detect syllable boundaries and used syllable-lengthening factor to detect phrase boundaries for reducing the computational complexity of recognition search. Although the recognition speeds improved significantly in those studies, slight losses on the recognition rates were paid.

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 acoustic decoding and aims at detecting word-boundary cues of the input utterance to help the following linguistic decoder solving the problem of word-boundary ambiguity. It uses an RNN to detect the word-boundary information from the input prosodic features extracted from the testing utterance with syllable boundaries pre-determined by the preceding acoustic decoder. The detected word-boundary information is then used in linguistic decoding to assist in determining the best word (or character) sequence.

Two distinct properties of the proposed method can be found as compared with previous studies (Batliner et al., 1996; Grice et al., 1996; Kompe et al., 1997; Silverman et al., 1992). One is that it adopts word boundary information as the output targets to be modeled instead of the conventional multi-level prosodic marks, such as the TOBI system (Grice et al., 1996; Silverman et al., 1992), or the prosodic-syntactic features (Batliner et al., 1996; Kompe et al., 1997). This leads to the following three advantages although the performance of word boundary detection may be degraded for some cases when two or more words are combined into a word chunk and pronounced within a prosodic phrase. First, it uses an RNN to automatically learn to do prosodic phrasing of Mandarin utterance and implicitly stores the mapping within the internal RNN representation. No explicit prosodic labeling of the speech signals is needed. Second, it is easy to incorporate the prosodic model into the linguistic decoder by a soft-decision scheme which directly takes the RNN outputs as additional scores, or by a partial-soft-and-partial-hard-decision scheme which uses the RNN outputs to drive an FSM for setting path constraints to restrict the linguistic decoding search (to be discussed in Section 3). Both of them can cope with the performance degradation on word boundary detection caused by pronouncing a word chunk within a prosodic phrase. Third, it is relatively easy to prepare a large training database without the help of linguistic experts. Only a simple word tokenization system is needed to analyze the texts associated with the training utterances for finding the output targets to be modeled. Neither complicated syntactic analyses nor cumbersome prosodic-mark labeling are needed. Another property of the proposed method lies in its use of neural network technology to solve the prosodic phrasing problem.

The organization of the paper is stated as follows. Section 2 presents the proposed prosodic modeling method. Section 3 describes the functional blocks of the Mandarin speech-to-text conversion system. The way of incorporating the prosodic model into linguistic decoding is discussed in detail. Effectiveness of the prosodic modeling method and its usefulness in helping linguistic decoding were evaluated by simulation experiments and are discussed in Section 4. Some conclusions are given in Section 5.

2. The proposed prosodic modeling method

Before discussing the proposed prosodic modeling method, we briefly introduce the characteristics

<table>
<tr><td colspan="4">Tonal Syllable (1345)</td></tr>
<tr><td colspan="3">Base Syllable (411)</td><td rowspan="3">Tone (5)</td></tr>
<tr><td rowspan="2">INITIAL (21)</td><td colspan="2">FINAL (39)</td></tr>
<tr><td>Medial (3)</td><td>Nucleus (9)</td><td>Ending (5)</td></tr>
</table>

Fig. 1. The phonological hierarchy of Mandarin syllables. Here, the number within a parenthesis indicates the total number of the specified unit in Mandarin Chinese.

of Mandarin Chinese. Mandarin Chinese is a tonal and syllabic language. There exist more than 80,000 words, each composed of one to several characters. There are more than 10,000 commonly used characters, each pronounced as a mono-syllable with one of five tones. The total number of phonologically allowed mono-syllables is only 1345. All mono-syllables have a very regular, hierarchical phonetic structure as shown in Fig. 1 (Cheng, 1973). A mono-syllable is composed of a base-syllable and a tone. There are in total 411 base-syllables. A base-syllable can be further decomposed into two parts: an optional *initial* (*onset* (Yin, 1989)) and a *final* (*rime* (Yin, 1989)). The *initial* part contains a single consonant if it exists. The *final* part consists of an optional *medial* (semivowel (Wu, 1998)), a *vowel nucleus*, and an optional *nasal ending* (*coda* (Yin, 1989)). These 411 base-syllables are formed by all legal combinations of 21 *initials* and 39 *finals*. These 39 *finals* are, in turn, formed by the combinations of 3 *medials*, 9 *vowel nuclei* and 5 *nasal endings*. There are only five lexical tones, namely, Tone 1 (or high-level tone), Tone 2 (or high-rising tone), Tone 3 (or low-dipping tone), Tone 4 (or high-falling tone), and Tone 5 (or neutral tone). Conventionally, a complete continuous Mandarin speech recognition system is generally composed of two components: acoustic decoding for mono-syllable identification and linguistic decoding for word (or character) string recognition. Owing to the regular hierarchical phonetic structure of mono-syllables, acoustic decoding is traditionally further decomposed into two sub-components: base-syllable recognition and tone recognition. In this study, we add a new RNN-based prosodic model to the conventional continuous Mandarin speech recognition system via inserting it in between acoustic decoding and linguistic decoding.

Fig. 2 shows a block diagram of the proposed RNN-based prosodic modeling method. It operates in two phases: a training phase and a testing phase. In the training phase, each training utterance is first processed in acoustic decoding to obtain the best syllable-boundary segmentation matching with the associated text. Some prosodic features are then extracted based on the best syllable-boundary segmentation. Meanwhile, the text associated with the input utterance is tokenized using a statistical model-based method to extract some word-boundary information. An RNN prosodic model is then trained to learn the relationship between the input prosodic features of the training utterance and the output word-boundary information of the associated text. In the testing phase, the input utterance is first processed in acoustic decoding to generate a top-N base-syllable lattice (to be discussed in detail in Section 3). Some prosodic features are then extracted based on the syllable-boundary segmentation of the top-1 base-syllable sequence. The well-trained RNN prosodic model is then employed to generate output word-boundary cues for the base-syllable lattice by using these input prosodic features. An FSM is then used to discriminate reliable outputs of word-boundary cues from unreliable ones. These detected word-boundary cues are lastly used to help linguistic decoding in the next stage. Fig. 3 shows the architecture of the simple RNN. It is a three-layer network with all outputs of the hidden layer being fed back to the input layer as additional inputs (Elman, 1990). An RNN of this type has been shown in some previous studies (Chen et al., 1998; Elman, 1990; Robinson, 1994) to possess a good ability of learning the complex relationship of the input feature vector sequence and the output targets via implicitly storing the

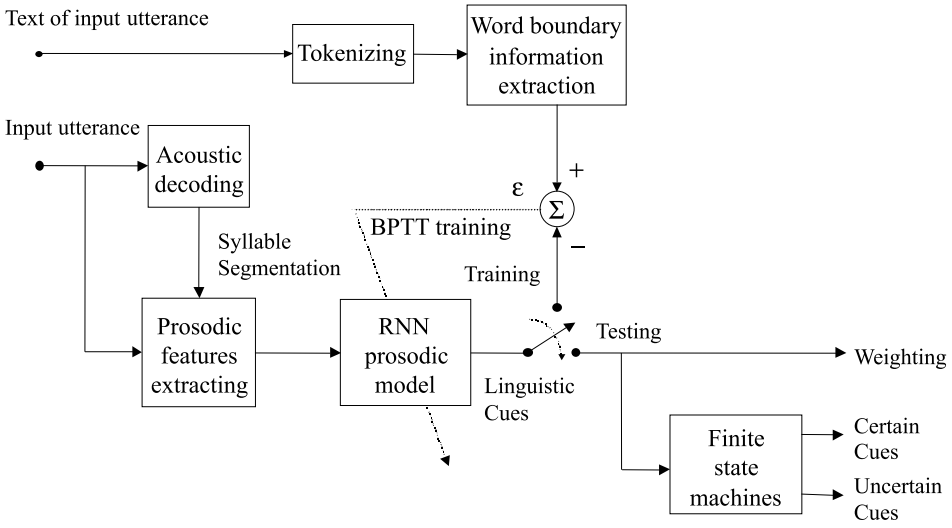


Fig. 2. A block diagram of the RNN-based prosody modeling method.

contextual information of the input sequence in its hidden layer. So it is suitable for the problem of realizing a complex mapping between the input prosodic features and the output linguistic features. The RNN can be trained by the back propagation through time (BPTT) algorithm (Haykin, 1994).

Inputs of the RNN prosodic model include some acoustic features extracted from several syllable segments, of the top-1 base-syllable sequence, surrounding the current syllable segment. Features used in this study are selected based on their close

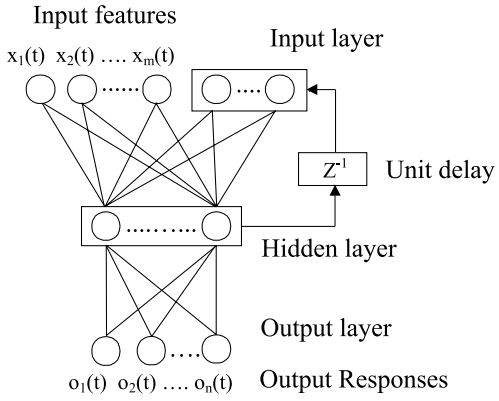


Fig. 3. The structure of the prosodic-modeling RNN.

relation with the prosody of speech signal. It is known that pitch, energy and timing information are prosody-related features and, hence, widely used in some previously prosodic-modeling studies (Campbell, 1993; Hirose and Iwano, 1997, 1998; Kompe et al., 1995; Wightman and Ostendorf, 1994). Since prosody is a supra-segmental feature of speech signal, prosody-related features to be considered must be for speech segments much larger than frame. In this study, we choose syllable segment as the basic unit to extract features for prosodic modeling because syllable is the basic pronunciation unit. For each syllable segment of the top-1 base-syllable sequence, two prosodic feature sets are extracted. One contains some local features of the current syllable segment, while the other contains some contextual features extracted from the syllable segment and its two nearest neighbors. Fig. 4 shows a schematic diagram of the feature extraction. Local features in the first set include: (1) the mean M_t and slope S_t of the pitch contour of the current syllable segment, (2) the log-energy mean E_t and the normalized log-energy mean NE_t of the *final* part of the syllable segment and (3) the normalized duration ND_t of the syllable segment. Here t is the index of the syllable segment and the two normalization operations are performed with respect to the *final* type $F(t)$ of the tth

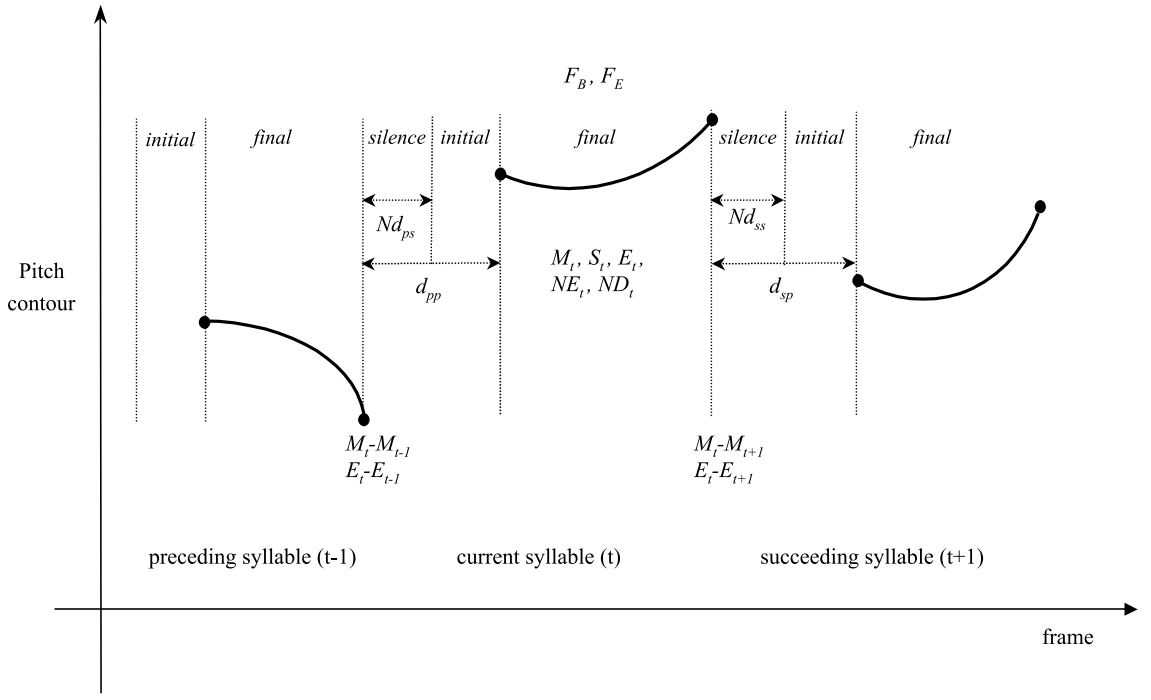


Fig. 4. A schematic diagram showing the extraction of the input prosodic features for the RNN-based prosodic modeling.

syllable segment. The normalization operation is discussed as follows. Let the mean and standard deviation of the log-energies of *finals* with type $F(t)$ be $\mu^E_{F(t)}$ and $\sigma^E_{F(t)}$, respectively. Then NE_t is obtained by first subtracting $\mu^E_{F(t)}$ from the log-energy mean E_t and then being divided by $\sigma^E_{F(t)}$. The same normalization process is performed on the duration of the syllable segment to calculate ND_t. We note that these two normalization operations are to compensate the high variabilities of *final* on both log-energy level and syllable duration. The reasons of using these local features in the prosodic modeling study are briefly discussed as follows. Pitch mean and log-energy mean of syllable segment are useful in discriminating different states of a prosodic phrase because both the pitch level and the log-energy level in the beginning part of a prosodic phrase are usually much higher than those in the ending part. Duration of syllable segment is useful in identifying the ending point of a prosodic phrase because the lengthening effect always occurs at the last syllable of a prosodic phrase.

The second set contains contextual features extracted from the current syllable segment and its two nearest neighbors. They include: (1) two flags indicating whether the current syllable segment is, respectively, the beginning and ending syllable segments, F_B and F_E, of a sentence, (2) two values showing the durations, d_{pp} and d_{sp}, of the non-pitch segments between the current pitch contour and its two nearest neighbors, (3) two normalized inter-syllable pause durations, Nd_{ps} and Nd_{ss}, besides the current syllable, (4) two pitch mean differences, $M_t - M_{t-1}$ and $M_t - M_{t+1}$, and (5) two log-energy mean differences, $E_t - E_{t-1}$ and $E_t - E_{t+1}$, between the current syllable and its two nearest neighbors. Here the two normalized values, Nd_{ps} and Nd_{ss}, are performed with respect to the *initial* types, $I(t)$ and $I(t-1)$, of the current and succeeding syllables, respectively. Specifically, let the mean and standard deviation of the pause durations preceding *initials* of type $I(t)$ be $\mu^{pd}_{I(t)}$ and $\sigma^{pd}_{I(t)}$. Then, Nd_{ps} is obtained by subtracting $\mu^{pd}_{I(t)}$ from the pause duration d_{ps} preceding the syllable segment and then being divided by $\sigma^{pd}_{I(t)}$. The same

normalization process is applied to the pause duration d_{ss} following the syllable segment to obtain Nd_{ss}. These two normalization operations are used to compensate the affection of the succeeding *initial* on pause duration. It is noted that the pitch mean and log-energy mean are set to zero for a missing preceding or succeeding syllable segment. Besides, d_{pp} and d_{sp} are set to zero for the first and last syllable segments of a sentence, respectively. The reasons of using these features in the prosodic modeling study are discussed as follows. The first two parameters in (1) are to set the boundary conditions. The use of inter-syllable pause duration is to consider its relatively large value for both major and minor breaks. The uses of pitch mean difference and energy mean difference are to consider the relatively large jumps of the pitch level and the energy level at prosodic phrase boundaries. Notice that similar features have been used in some previous studies. For instances, fundamental frequency mean difference of mora was used in the detection of the accent type of prosodic word (Iwano and Hirose, 1998, 1999; Iwano, 1999). There are in total 15 prosodic features extracted for each syllable segment. The number of input features for each syllable segment with its two nearest neighbors therefore equals to 105. The issue of the suitable amount of input features for this problem can be raised for discussion. However, based on the dimensionality reduction function of multi-layer neural network (Morgan and Scofield, 1991), the extracted features in the hidden layer of a multi-layer neural network correspond to a low-dimensional projection from the input feature space to the input pattern space. Non-essential information for output classification can be removed automatically.

The evolution of the calculation of prosodic features in a sample utterance is shown in Fig. 5. The waveform and the segmentation information are shown in the upper part of this figure. The utterance contains 11 syllables. The solid lines, dashed lines and dotted lines represent the starting points of syllables, the junctions of *initial* and *final* and the ending points of syllables, respectively. The middle part of this figure shows the $F0$ contour of the utterance. The star symbols represent the $F0$ means of these 11 syllables. The energy

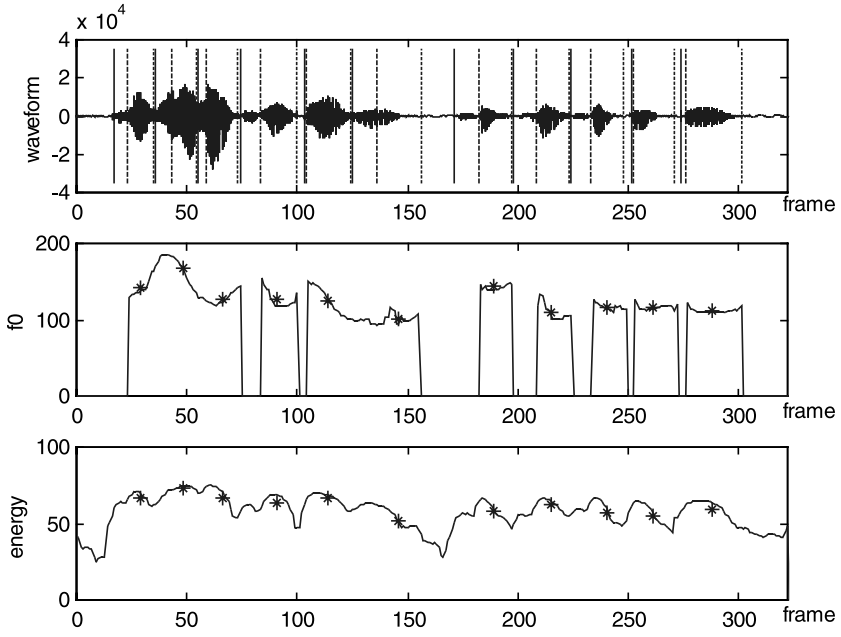


Fig. 5. The waveform, segmentation information, $F0$ contour and energy contour of the sample utterance "/ch2/ /yeh4/ /lan2/ /chiu2/ /dui4/ /yuan2/ /shen1/ /tsai2/ /hsiang1/ /dang1/ /kao1/" (The professional basketball players are very tall.).

contour of the utterance is shown in the lower part of this figure. The star symbols represent the energy means of these 11 syllables. All the prosodic features can be calculated based on the segmentation information, the $F0$ mean and the energy mean shown in this figure. It can be seen from this figure that the features of $F0$ mean and energy mean can represent well the global variations of utterance. Moreover, in order to considering the affection from a wider context in the prosodic modeling, the prosodic feature sets of the current syllable segment and its six nearest neighbors are assembled and taken as the inputs of the RNN prosodic model. In the feature-assembling process, two things need to be specially taken care. One is the elimination of some duplicate contextual prosodic features collected across several neighboring syllable segments. This can increase the efficiency of the prosodic modeling. The other is the setting of boundary conditions for some syllable segments in the beginning and ending parts of every utterance. A simple scheme to set all non-existing prosodic features to zero is adopted in this study. Investigating the responses of the RNN prosodic model to the boundary conditions confirmed the suitability of the simple boundary-condition-setting scheme.

The outputs of the prosodic-modeling RNN include four linguistic features. These four flags indicating whether the current syllable is a mono-syllabic word or is the beginning syllable, the intermediate syllable, or the ending syllable of a polysyllabic word. These four output features are denoted as MW (a mono-syllabic word), BPW (the beginning syllable of a polysyllabic word), IPW (an intermediate syllable of a polysyllabic word) and EPW (the ending syllable of a polysyllabic word), respectively. To prepare output targets for training the RNN, texts associated with all training utterances are tokenized into word sequences in advance by an automatic statistical model-based algorithm with a long-word-first criterion (Su, 1994). The algorithm was tested to achieve a word tokenization accuracy rate around 97% (Su, 1994). Here, a lexicon containing 111,243 words is used in the tokenizing process. Besides, several simple word-merging rules are used to improve the word tokenization. They are bracketing rules which construct some types of compound words, missing in the current lexicon, including character-duplicated compound words (e.g., 快快樂樂 (happy)), determiner-measure compound words (e.g., 一種 (one type)), short verb–preposition compound words (e.g., 坐在 (sit at)), noun-localizer compound words (e.g., 森林中 (inside forest)), short adverb–verb compound words (e.g., 可預測 (can predict)), negation-verb compound words (e.g., 不守時 (not keep time)), short adjective–noun compound words (e.g., 小石頭 (small stone)), etc. Some tokenization errors are corrected manually. All output targets to train the RNN are extracted from these tokenized word sequences. It is worth noting that, owing to the following two reasons, we do not use high-level syntactical features, such as syntactic phrase boundaries, in this study. One is that it is generally not easy to do syntactic analysis for unlimited texts of natural Chinese language. The other is that the syntactic structure of a Chinese text is not isomorphic to the prosodic phrase structure of the corresponding Mandarin speech.

To check the classification ability of the RNN, the four flags showing the location of the current syllable in a word are considered. This flag set is referred to as Word-tag. An FSM is used to examine whether the responses of the RNN are good enough to make reliable classifications for this flag set. The purpose of the FSM is to provide a partial-hard-and-partial-soft classification scheme for safely invoking the classification results into the speech recognition process. We can therefore take the advantage of pre-classifying the input speech signal to design a sophisticated search procedure for the recognition process when the pre-classification is reliable and to simultaneously avoid unamendable speech recognition errors caused by pre-classification errors. The topology of the Word-tag FSM is shown in Fig. 6. The symbols "M", "B", "I", "E" and "U" shown in the Word-tag FSM denote "MW", "BPW", "IPW", "EPW" and uncertain states, respectively. The operation of the FSM is stated as follows. 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 stable state if it is a legal one; otherwise it moves to an uncertain (U) state. These two thresholds are de-

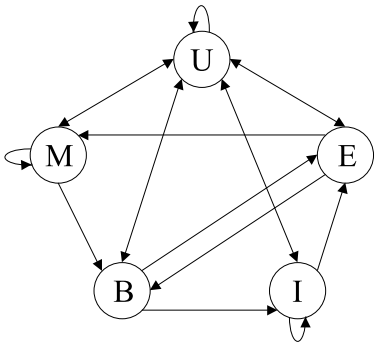


Fig. 6. The topologies of the Word-tag FSM.

termined empirically by considering the tradeoff between the classification accuracy rate and the number of undetermined responses.

3. The mandarin speech-to-text conversion system invoking with the RNN prosodic model

A complete block diagram of the proposed speech-to-text system invoking with the RNN prosodic model is shown in Fig. 7. The input speech is first preprocessed to extract some acoustic features for base-syllable recognition. Recognition features extracted include 12 Mel-frequency cepstral coefficients (MFCCs), 12 delta MFCCs, and a delta log-energy. Then, HMM-based base-syllable recognition is done to generate a top-N base-syllable lattice. The HMM recognizer uses 100 three-state right-*final*-dependent (RFD) *initial* models and 39 five-state context-independent (CI) *final* models to form 411 eight-state base-syllable models and is trained using the maximal likelihood criterion. For silence, a single-state model is used. The observation features in each HMM state is modeled by a mixture Gaussian distribution. The number of mixture components in each state of these models is variable with a range from 1 to 20. The number of mixture components used in each HMM state depends on the number of training data. The top-N base-syllable lattice is generated by the Viterbi-parallel-backtrace method (Huang and Wang, 1994). The method consists of two steps: a forward one-stage Viterbi search and a parallel-backtracking (PB) procedure. In the first step, the top-1 base-syllable string matching with the input utterance is obtained by a one-stage Viterbi search. The corresponding base-syllable boundaries are detected by decoding the top-1 base-syllable string through a simple backtracking procedure. Then, in the second step, the PB procedure is applied to expand the top-1 base-syllable string to the top-N base-syllable lattice. The PB procedure keeps all base-syllable boundaries and expands all top-1 base-syllables in parallel to produce the base-syllable candidates of the top-N base-syllable lattice. For each top-1 base-syllable, it first sets a search range by letting the end boundary be fixed to the ending point of the top-1 base-syllable segment and relaxing the beginning boundary to a small range of the beginning point of the top-1 base-syllable. It then matches the speech segment in the search range with HMM models of all base-syllables other than the top-1 base-syllable to find the remaining top-N base-syllable candidates by time-reversed Viterbi searches.

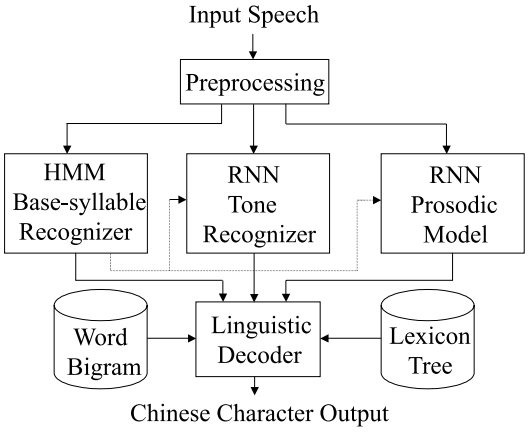


Fig. 7. A functional block diagram of the proposed Mandarin speech-to-text conversion system.

With all base-syllable boundaries pre-determined by the HMM base-syllable recognizer, tone recognition and prosodic modeling are then performed. Since the tonality has lexical meaning as discussed in Section 2, tone recognition is important in Mandarin speech recognition. An RNN tone recognizer (Wang and Chen, 1994) is used in this study. It also employs an RNN with the same

Table 1
The node number distribution of the lexical tree

Layer no.	First layer	Second layer	Third layer	Fourth layer	Fifth layer	Total
Node number	1217	69811	37169	14238	808	123243

structure shown in Fig. 3 to discriminate the five lexical tones using recognition features extracted from the speech part surrounding the current syllable. The feature extraction is discussed as follows. Given with all base-syllable boundaries, two sets of local features and contextual features are extracted for every syllable. The local features are extracted from the current syllable segment and include: (1) four orthogonal transformed coefficients (Chen et al., 1998) representing the mean and the shape of the pitch contour, (2) the log-energy mean, (3) the duration of the syllabic pitch contour, (4) the syllable *initial* type, (5) the syllable *medial* type, (6) the syllable *final* type and (7) the syllable *nasal-ending* type. The contextual features include: (1) eight orthogonal transformed coefficients representing the pitch contours of the two nearest neighboring syllables, (2) F_B and F_E, (3) d_{pp} and d_{sp} and (4) Nd_{ps} and Nd_{ss}. There are in total 53 features used in the tone recognition. Note that some features are simultaneously used for RNN prosodic model and RNN tone recognizer. By using the above tonal features, the RNN tone recognizer generates the best M tone candidates for each base-syllable segment.

Then the outputs of the base-syllable recognizer, the tone recognizer, and the prosodic models are all fed into the linguistic decoder to generate the recognized word (character) string. The linguistic decoder first combines the top-N base-syllable lattice generated by the base-syllable recognizer, with the best M tone candidates generated by the tone recognizer, to form an $M \times N$ tonal-syllable lattice. It then performs a decoding search on the syllable lattice to find the best word string. The decoding search employs a Viterbi search algorithm invoking with a word-construction process and a statistical language model. The word-construction process uses a lexicon containing 111,243 entries. Each entry consists of one to five syllables. A backward lexical tree is built from the lexicon for word construction. Each node of the lexical tree represents a tonal syllable. The node number in each layer of the lexical tree is listed in Table 1. Fig. 8 shows a small portion of the lexical tree. Here nodes circled with solid and dashed lines represent leaf and intermediate nodes, respectively. There may exist a list of homonym words on a leaf node. The detailed information registered in each node includes the address of next node in the same layer, the address of child node in the succeeding layer, the tonal-syllable code, a flag showing whether the current syllable is the beginning syllable of a word, the word unigram probability, and a list of all homonym words. The statistical language model used in the linguistic decoding search accommodates both word-unigram and word-class-bigram probabilities. Word classes used in the calculation of word-class-bigram probabilities are generated by a simple word classification scheme which considers a special property of Chinese language. The property says that many polysyllabic words sharing the same beginning or ending character have the same

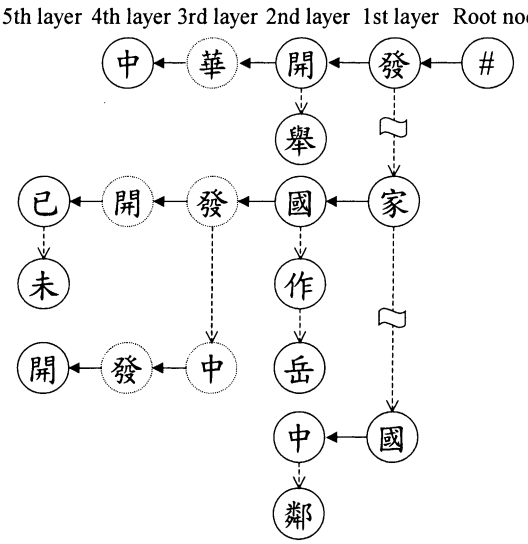


Fig. 8. A small part of the backward lexical tree.

function in syntax or semantics (Yang et al., 1994). For examples, 火車 (train), 汽車 (car), 腳踏車 (bicycle), etc., all end with the character "車" which is a generic name of vehicle; 副主席 (vice chairman), 副總統 (vice president), 副校長 (vice principal), etc., all begin with the character "副". Two sets of word classes are generated to comply with the special property. One is formed by clustering all words with the same beginning base-syllable into a class. It is referred to as *CR*. The other is formed by clustering all words with the same ending base-syllable into a class and is referred to as *CL*. Both *CR* and *CL* contain 411 word classes. Each word-class-bigram probability, therefore, represents the frequency of a word-class-pair with the left and right word classes belonging to *CL* and *CR*, respectively. A well-tagged corpus which contains about three million words (Chinese Knowledge Information Processing Group, 1995) is used to train the statistical language model. Although the number of all possible combinations of right and left word classes is 411×411, there exist only 73761 combinations in the corpus. All missing word-class-bigram probabilities are simply assigned a low probability value.

Invoking with the word-construction process and the statistical language model, the Viterbi search calculates discriminant scores for all word sequence hypotheses and finds the best word sequence with maximal discriminant score as the recognized result. Three searching schemes are used in this study. They include the conventional (baseline) scheme and two schemes assisted with the proposed RNN prosodic model. In the baseline scheme without invoking the prosodic model, the Viterbi search starts a searching process at each syllable segment of the syllable lattice. The searching process continues backward along the syllable lattice and ends at the location of the fourth syllable segment ahead because the maximum word length is five for the lexicon used in the study. In the backward searching process, all possible words with lengths from one to five syllables are constructed using the backward lexical tree. The score of each word-sequence candidate W is formed by combining the likelihood scores of all constituent base-syllables provided by the HMM base-syllable recognizer, the scores of all constituent tones provided by the RNN tone recognizer, the unigram probabilities of all constituent words, and the bigram probabilities of all constituent word-class-pairs, i.e.,

$$\begin{aligned} L(W) = & \sum_{k=1}^{K} w_{\mathrm{HMM}} \ln(p_{\mathrm{HMM}}(b_k)) \\ & + \sum_{k=1}^{K} w_{\mathrm{tone}} \ln(o(t_k)) + \sum_{l=1}^{L} w_{\mathrm{UG}} \ln(p_{\mathrm{UG}}(w_l)) \\ & + \sum_{l=2}^{L} w_{\mathrm{BG}} \ln(p_{\mathrm{BG}}(CR(w_l)|CL(w_l - 1))), \end{aligned} \tag{1}$$

where $W = \{w_1 w_2, \ldots, w_L\} = \{s_1 s_2, \ldots, s_K\}$ consists of L words or equivalently K syllables, w_l is the lth word, $s_k = (b_k, t_k)$ is the kth syllable formed by base-syllable b_k and tone t_k, $o(t_k)$ is the score of tone t_k, $p_{\mathrm{HMM}}(b_k)$ is the likelihood score of base-syllable b_k, $p_{\mathrm{UG}}(w_l)$ is the unigram probability of word w_l, $p_{\mathrm{BG}}(CR(w_l)|CL(w_{l-1}))$ is the word-class-bigram probability of the word-pair (w_{l-1}, w_l), and $w_{\mathrm{HMM}}, w_{\mathrm{tone}}, w_{\mathrm{UG}}$ and w_{BG} denote the weights for the four different scores. A discriminating optimization experiment (Chiang et al., 1996) has been conducted in finding the suitable weights to integrate the four different scores. Two schemes of incorporating the RNN prosodic model into the linguistic decoding are proposed. Scheme 1 modifies the baseline scheme by directly taking the outputs of the prosodic-modeling RNN as additional scores and changing the recognition score by

$$L'(W) = L(W) + \sum_{k=1}^{K} w_{\mathrm{PM}} l_{\mathrm{PM}}(s_k), \tag{2}$$

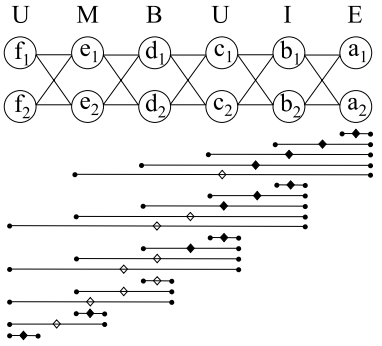
where

$$l_{\mathrm{PM}}(s_k) = \begin{cases} \ln(o_{\mathrm{MW}}(k)) & \text{if } s_k \text{ is a mono-syllabic word,} \\ \ln(o_{\mathrm{BPW}}(k)) & \text{if } s_k \text{ is the beginning syllable of a polysyllabic word,} \\ \ln(o_{\mathrm{IPW}}(k)) & \text{if } s_k \text{ is an intermediate syllable of a polysyllabic word,} \\ \ln(o_{\mathrm{EPW}}(k)) & \text{if } s_k \text{ is the ending syllable of a polysyllabic word,} \end{cases} \tag{3}$$

$o_{MW}(k), o_{BPW}(k), o_{IPW}(k)$ and $o_{EPW}(k)$ are, respectively, the MW, BPW, IPW and EPW outputs of the prosodic modeling RNN at syllable s_k and w_{PM} denotes the corresponding weight for this score. The suitable weights of $w_{HMM}, w_{tone}, w_{UG}, w_{BG}$ and w_{PM} are obtained via using a joint optimization experiment (Chiang et al., 1996). Scheme 2 is an extended version of Scheme 1. In addition to using the new recognition score defined in Eq. (2), Scheme 2 also uses the word-boundary information provided by the Word-tag FSM to set path constraints to further restrict the Viterbi search. Different constraints are set for the five states of M (MW), B (BPW), I (IPW), E (EPW) and U (uncertain). Generally, more restrictive searches are used for the three decisive states of M, B and E while full searches are used for I and U states. Specifically, when a syllable segment of the input syllable lattice is classified as an M state, a B state, or an E state, we restrict it to be a mono-syllabic word, the beginning syllable of a polysyllabic word, or the ending syllable of a polysyllabic word in the Viterbi search. Otherwise, for an I state or a U state, we do not set any restrictions to the searching process. In realization, Scheme 2 modifies the Viterbi search of the baseline scheme by letting a backward searching process be activated only when the current syllable segment is at M, E, I or U state and stopped at a syllable segment with B or M state or before a syllable segment with M or E state. It is noted here that I state is treated in the same way as U state because many MWs, BPWs and EPWs are erroneously classified as IPWs (to be discussed in detail in Section 4). Fig. 9 compares the backward searching processes of the baseline scheme and Scheme 2 for a simplified top-2 syllable lattice labeled with the outputs of the Word-tag FSM. In the figure, all partial paths need to be considered for constructing candidate words via accessing the word-lexical tree are displayed. It can be found from Fig. 9 that Scheme 2 has much less partial paths to be considered than the baseline scheme. So Scheme 2 is more efficient in computational complexity.

4. Experimental results

Effectiveness of the proposed prosodic modeling method was examined by simulation experiments on a speaker-dependent, continuous Mandarin speech recognition task using a large single-speaker database. The database contained 452 sentential utterances and 200 paragraphic utterances. Texts of these 452 sentential utterances were well-designed, phonetically balanced short sentences with lengths less than 18 characters. Texts of these 200 paragraphic utterances were



The path needs to be checked for both baseline and scheme 2.

The path needs to be checked only for baseline method.

Fig. 9. The comparison of the backward searching processes of the baseline scheme and Scheme 2 for a simplified top-2 syllable lattice labeled with the outputs of the Word-tag FSM. Here, all partial paths need to be considered for constructing candidate words via accessing the word-lexical tree are shown.

news selected from a large news corpus to cover a variety of subjects including business (12.5%), medicine (12.0%), social event (12.0%), sports (10.5%), literature (9.0%), computers (8.0%), food and nutrition (8.0%), etc. All utterances were generated by a male speaker. They were all spoken naturally at a speed 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 the prosodic modeling

To test the proposed prosodic modeling method, we first trained the HMM base-syllable recognizer and used it to segment all speech utterances into syllable sequences. Pitch is then detected by the simplified inverse filter tracking (SIFT) algorithm (Markel and Gray, 1976). The SIFT algorithm first uses an LP filter to reduce the bandwidth of the input signal to 0–1 kHz. The LP-filtered signal is then down-sampled. Then, a fourth-order inverse filter designed using the autocorrelation method for linear prediction analysis is applied to flatten the signal spectrum. Pitch period is then detected from the inverse-filtered signal by the autocorrelation method. A second-order interpolation is then applied to increase the resolution of the detected pitch period. Besides, a simple error correction is applied to eliminate some errors with abrupt pitch jumps. Lastly, all remaining errors are manually corrected. It is noted that the manual correction effort is much less than the manual prosodic-labeling effort because the former task is much easier. After we detected pitch, features for RNN-based prosodic modeling were then extracted from all segmented training utterances. Meanwhile, texts of all training utterances are tokenized into word sequences. Word-boundary features for RNN output targets were then extracted. The details of these feature extraction processes have been described in Section 2. The RNN prosodic model was then trained by the BPTT algorithm. The number of nodes in the hidden layer of the RNN is determined empirically and set to be 30 in this study.

Table 2
The confusion matrix of Word-tag classification for the RNN prosodic model. The overall classification rate is 71.9%

	Result			
Desired	BPW	IPW	EPW	MW
BPW	1861	317	110	52
IPW	288	1106	341	11
EPW	135	369	1774	62
MW	160	28	107	313

Table 2 shows the classification results of the RNN-based prosodic modeling for the output set of Word-tag without invoking the FSM. Accuracy rate of 71.9% was achieved for Word-tag set. Then, the performance of the RNN-based prosodic modeling invoking with Word-tag FSM was examined. The Word-tag FSM used thresholds of $\mathrm{Th}_h = 0.6$ and $\mathrm{Th}_d = 0.3$. In this FSM, an uncertain state was added for the cases when the RNN did not respond well for making reliable classifications. Experimental results are shown in Table 3. Accuracy rate increased to 87.6% with 32.3% outputs staying in uncertain states for Word-tag set. Table 4 shows the classification performances of Word-tag FSM for two sets of thresholds. It can be found from the table that the classification accuracy rate increased with a paid of putting more syllables into U state as a more restrictive threshold setting was used. Fig. 10 shows a typical example of the Word-tag responses of the prosodic-modeling RNN and the corresponding FSM to an input sentential utterance. It can be found from the figure that the FSM functioned well to reliably determine all B states and some I and E states. Lastly, it is noted that, although the above prosodic modeling test was conducted in a

Table 3
The confusion matrix of Word-tag classification for the RNN prosodic model invoking with an FSM using $\mathrm{Th}_h = 0.6$, $\mathrm{Th}_d = 0.3$. The classification rate is 87.6%

	Result				
Desired	BPW	IPW	EPW	MW	Uncertain
BPW	1547	147	27	14	605
IPW	141	720	118	1	766
EPW	56	163	1420	37	664
MW	97	7	67	198	239

Table 4
Performance comparison of Word-tag classifications for the RNN prosodic model without FSM and with FSM using two different sets of thresholds

Word-tag	No threshold	$Th_h = 0.6,\ Th_d = 0.3$	$Th_h = 0.8,\ Th_d = 0.6$
BPW	76.1%	84.0%	91.2%
IPW	60.8%	69.4%	78.2%
EPW	76.8%	87.0%	94.8%
MW	71.5%	79.2%	87.6%

speaker-dependent mode, it can be adapted to a speaker-independent mode via properly normalizing the input pitch and energy features by the pitch level and loudness of each individual speaker. This is worth further studying in the future.

4.2. An error analysis of the prosodic modeling

In this study, the linguistic features, used as output targets to train the RNN prosodic model, are primarily extracted based on the word tokenization results obtained by a simple statistical model-based method. But, owing to the out-of-vocabulary (OOV) problem, the tokenization cannot be very accurate. Besides, the prosodic phrasing of an utterance can be influenced by many factors other than the linguistic features of the associated text, such as word emphasis, the need for breathing, speaking rate and emotional status, etc. Those mismatches are harmful to the accuracy of the RNN prosodic modeling. A detail error analysis is therefore needed in order to compensate those effects for making the classification results more useful in linguistic decoding. In the following, the classification errors shown in Table 3 for Word-tag set were investigated.

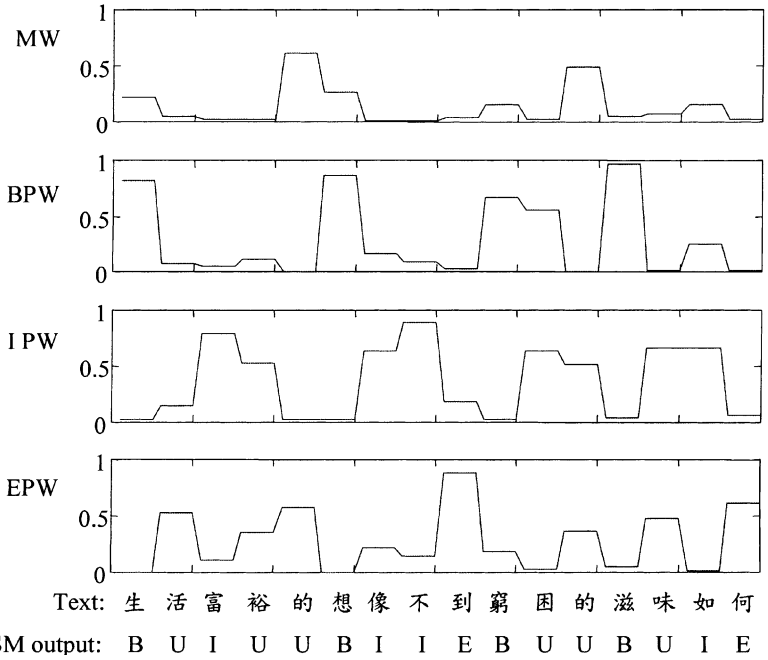


Fig. 10. An example of the Word-tag responses and the corresponding FSM outputs for a sentential utterance.

MW errors. From Table 3, an MW was easy to be misclassified as a BPW or an EPW. This was because it was combined with the following or preceding words to form a compound word. About 70% errors of misclassifying MWs as BPWs occurred when these MWs were short adverb, preposition, conjunction or transitive verb (see example sentences M-a and M-b, where the processing Chinese character and the corresponding word in English translation are marked with an underline). This also easily occurred when MWs were special function words, such as "是" (is), "有" (has) and "的" (of) (see M-c and M-d). For the case of misclassifying MWs as EPWs, we found that 17%, 17%, 30% and 10% of MWs were, respectively, preposition, auxiliary, pronoun and verb words (see M-e and M-f). This also easily occurred when MWs were "是" and "的". For some cases, an MW was combined simultaneously with the following and preceding words to form a word chunk when it was a preposition or conjunction (see M-g).

M-a: 髒東西都擦掉了。(Dirty things are all wiped off.);

M-b: 有點像在暗房裡工作一樣。(Just like working in a dark room.);

M-c: 應不致於有任何變化。(It ought to be no change at all.);

M-d: 電子琴大賽是一年一度的國際性音樂比賽。(The chord organ match is an annual international music contest.);

M-e: 開車去德州西部。(Driving to west Texas.);

M-f: 空氣不好使我頭昏腦脹。(Dirty air makes me feel dizzy.);

M-g: 論體力及臂力。(Talking about physical agility and arm strength.).

BPW errors. Most classification errors of BPWs occurred when they were combined with the preceding words and misclassified as IPWs. This was especially easy to occur when both the current and preceding words were all short (see example sentences B-a and B-b). For the cases of misclassifing BPWs as EPWs, we found that 42% of them were syllables with Tone 2 or Tone 3 which were pronounced lightly (see B-c and B-d). The processing Chinese characters and the corresponding words in English translation are marked with an underline in the following examples. It is noted that, owing to the difficulties in finding the one-to-one correspondence between the Chinese character and English word, we choose the suitable English word by checking whether it contains the meaning of the underlined Chinese character.

B-a: 學生排隊歡迎得勝的球員。(The students line up to welcome the triumphant players.);

B-b: 台灣氣候潮濕。(The climate in Taiwan is humid.);

B-c: 社會科勞工行政股股長。(The chief of the labor administration section of the social department.);

B-d: 全國總決賽。(The final of national contest.).

IPW errors. Most IPW errors occurred when long words were prosodically broken into pairs of short words and misclassified as BPWs and EPWs (see I-a and I-b). Some other errors were owing to the concatenations with low-energy syllables with Tone 2 or Tone 3 (see I-c and I-d). Some interesting cases occurred when unfamiliar words, such as long translated names of foreigners and infrequently used characters, were pronounced. The speaker seemed to hesitate to utter such characters so as to insert breaks around them (see I-e and I-f).

I-a: 台北市花卉產銷公司。(The company of flower production and distribution in Taipei city.);

I-b: 觀景休閒區及湖濱渡假區。(The sightseeing leisure area and lakeside vacation area.);

I-c: 大陸摩托運動協會。(The association of motor sports in Mainland.);

I-d: 一千六百萬美元。(Sixteen million US dollars.);

I-e: 他仜仜睨睨的四處張望。(He looks around fearfully.);

I-f: 尤里庇底斯的作品將搬上舞台。(The works of Euripides will be performed on the stage.).

EPW errors. Most errors were owing to the connections with short succeeding words to become IPWs (see E-a and E-b). Besides, over 95% errors of misclassifying EPWs as MWs occurred at the ending syllables of compound words formed by combining words with special function words like "的" (of), "個" (a unit) and "在" (at) (see E-c and E-d). Some other errors resulted from the interferences by the preceding low-energy syllables with Tone 2 or Tone 3 (see E-e and E-f).

E-a: 高中職以上學歷。(The degree of senior high school and above.);

E-b: 經營中古機車行。(To manage a used motorcycle shop.);
E-c: 未來的就業機會。(The future job opportunity.);
E-d: 固定在元宵節舉辦的中華民藝華會。 (The Chinese folk fair is held regularly during the Lantern Festivals.);
E-e: 無法與高大型球員相抗衡。(Cannot contend with big and tall players.);
E-f: 娶長妻的男子。(The man married with an elder woman.).

Generally speaking, most word boundary detection errors are due to the tendency of grouping function words and some MWs with adjacent words. This phenomenon was found by Campbell (1993) in a study of detecting prosodic boundaries in British English. It showed that a function word or an MW can be merged either with the preceding word in accordance with rhythmic principles, or with the following word in accordance with syntactic principles to form a word chunk to be pronounced within a prosodic phrase. Aside from the same phenomenon, we also found in the current study that the merging process may also occur at both sides simultaneously for Mandarin speech in case that the two adjacent words are all short. Some other word boundary detection errors are owing to the false alarms occurred at the neighborhoods of characters with Tone 2 or Tone 3. This mainly results from the ambiguity in distinguishing an abrupt $F0$ jump owing to the $F0$ resetting occurred at a prosodic boundary or to the characteristics of these two lexical tones.

4.3. Results of the speech-to-text conversion

The same database used in the prosodic modeling was used in this study to test the speech-to-text conversion system invoking with the proposed RNN prosodic model. A sub-syllable-based HMM recognizer was constructed from the training set by the maximum likelihood training algorithm. Each testing utterance was first preprocessed by the HMM base-syllable recognizer to generate a top-N base-syllable lattice. The top-1 base-syllable recognition rate was 81.4% with substitution, insertion and deletion error rates being, respectively, 16.5%, 1.1% and 1.0%. The base-syllable inclusion rate was 96.5% for the top-10 base-syllable lattice. Then, both features for prosodic modeling and for tone recognition were extracted based on the given syllable-boundary segmentation of the top-1 base-syllable recognition. The top-2 tone inclusion rate for the case of using hand-corrected pitch contours was 98.0%. After performing prosodic modeling and tone recognition, we combined their results with the base-syllable recognition results in the linguistic decoder to generate the best recognized character (word) string.

The two schemes, Schemes 1 and 2, of invoking the RNN prosodic model into the linguistic decoding were then examined. Experimental results are displayed in Table 5. The character accuracy rate was calculated according to the following formula:

$$\text{character accuracy rate} = 1 - \frac{\text{insertion} + \text{deletion} + \text{substitution}}{\text{total character number}}. \quad (4)$$

The search complexity counted the number of accesses to the lexical tree for constructing candidate words to be disambiguated in the linguistic decoding search. It can be found from Table 5 that the character accuracy rate achieved by the baseline scheme without invoking the prosodic model was 73.6%. Detailed analyses revealed that the

Table 5
The performance comparison for speech-to-text conversion using three different linguistic decoding schemes

Method	Use of Word-tag RNN outputs	Word-tag FSM	Character accuracy	Complexity reduction
Baseline	X	X	73.6%	X
Scheme 1	Yes	X	74.6%	X
Scheme 2-1	Yes	$\text{Th}_h = 0.6, \text{Th}_d = 0.3$	73.9%	30.5%
Scheme 2-2	Yes	$\text{Th}_h = 0.8, \text{Th}_d = 0.6$	74.7%	17%

substitution, insertion and deletion error rates were 24.8%, 0.5% and 1.1%, respectively. The character accuracy rate was improved from 73.6% to 74.6% as we invoked the RNN prosodic model in the linguistic decoding by Scheme 1. The improvement resulted from the decrease of the substitution error rate by about 1% with almost no change for both the insertion and deletion error rates. Accuracy rates of 73.9% and 74.7% with 30.5% and 17% of search complexity reductions were achieved for the two cases of Scheme 2. This shows that the threshold setting in Word-tag FSM of Scheme 2 can make a tradeoff between the improvement on recognition performance and the reduction in computational complexity. If we compare these experimental results with those obtained in two related studies of using prosodic information to assist in speech recognition (Hsieh et al., 1996; Kompe et al., 1997), the proposed method is better because it improves not only the computational complexity but also the recognition performance. On the contrary, both previous methods improved the computational complexities in paid of minor losses on the recognition performance. If we compare the proposed method with another method of using prosodic modeling to improve mora recognition for Japanese speech (Iwano and Hirose, 1999), it is slightly inefficient on performance improvement. Based on above discussions, we can conclude that the proposed method of using prosodic information in Mandarin speech recognition is a promising one.

4.4. An error analysis of the speech-to-text conversion

For better understanding the effect of the RNN-based prosodic modeling on the linguistic decoding, we made a detailed error analysis for Scheme 2-2. Table 6 lists six main types of recognition errors. The first three error types, E1–E3, resulted from three types of acoustic processing errors including: (1) base-syllable recognition error – correct base-syllable is not included in top-10 base-syllable candidates; (2) tone recognition error – correct tone is not included in top-2 tone candidates; and (3) syllable-boundary segmentation error – owing to insertion and deletion errors in the top-1 base-syllable recognition. The HMM base-syllable recognizer was responsible for E1- and E3-type errors. Improving the performance of the HMM base-syllable recognizer is surely very helpful to correct them. The E2-type errors were caused by the RNN tone recognizer and may be corrected by including more tone candidates. But this will increase the computational complexity of the linguistic decoding. The two types of errors, E4 and E5, were owing to the incompleteness of word-merging rules applied to construct compound words. The use of a more sophisticated word-construction algorithm is surely helpful to correct them. The error type of E6 is due to the out-of-vocabulary problem. Increase the size of the lexicon may be useful to improve it. Since correct words could not be constructed to form a correct word sequence hypothesis when errors of these six types occurred, the word-boundary information provided by the RNN prosodic model is vain for correcting those errors in the linguistic decoding. If we do not count those uncorrectable errors, the improvement on the character accuracy rate by the proposed method is significant.

Table 6
An error analysis of the linguistic decoding

Error types	Percentage
E1: Base-syllable missing in the base-syllable lattice	11.1%
E2: Tone recognition error	21.6%
E3: Speech segmentation error	9.3%
E4: Compound word error owing to no morphological rules applied	7.0%
E5: Quantity word error owing to no morphological rules applied	4.8%
E6: OOV of the lexicon	8.0%
E7: Others	38.2%

Lastly, many errors were owing to homonymic ambiguity. In Mandarin speech-to-text conversion, homonymic ambiguity is a serious problem for recognizing mono-syllabic words because each syllable maps, in average, to over 10 characters. A more sophisticated language model is helpful to solve the problem. If we neglect all homonym errors, the character accuracy rate increases to 82.5%.

5. Conclusions

A new RNN-based prosodic modeling method for Mandarin speech recognition has been discussed in this paper. It uses an RNN to detect word-boundary information from the input prosodic features with base-syllable boundary being pre-determined by an HMM-based acoustic decoder. Two schemes of using the word boundary information to assist the linguistic decoder in solving word-boundary ambiguity as well as pruning unlikely paths were proposed. Experimental results on a speaker-dependent speech-to-text conversion test have confirmed that the RNN prosodic model is effective on detecting useful word boundary information from the input testing utterance for assisting in the linguistic decoding. Accuracy rate of 71.9% has been obtained for Word-tag detection. The character accuracy rate of speech-to-text conversion has increased from 73.6% to 74.7% with an additional gain of 17% reduction in the computational complexity of the linguistic decoding search. So it is a promising prosodic modeling method for Mandarin speech recognition.

Some further studies to improve the proposed RNN-based prosodic modeling method are worthwhile doing in the future. One is to improve the efficiency of the RNN prosodic model by compensating the effects of other affecting factors, such as tone and phonemic constituents of syllable, on the input prosodic features. Specifically, both the pitch and energy levels of a syllable are seriously affected by its tone. The energy level of a syllable is also seriously affected by its *final* type. If we can isolate those affecting factors from the prosodic modeling, we may achieve a better prosodic phrasing result. Another worthwhile study is to improve the effectiveness of the RNN prosodic model by providing more accurate prosodic phrasing information of training utterances for help preparing output targets. There are many cases that two or more words are combined together to form a word chunk and pronounced within a prosodic phrase in Mandarin speech. Experimental results have confirmed that this effect resulted in degradation on the performance of word boundary detection. If we can extend the current study to additionally include the word-chunk boundary information to help setting output targets, we may obtain a more precise RNN prosodic model. The other worthy further study is to find new ways of incorporating the RNN prosodic model into the speech recognition. In the current study, the RNN prosodic model is combined with the linguistic decoder to provide additional scores for discriminating word boundary from non-word-boundary (Scheme 1) and to additionally set path constraints for restricting the linguistic decoding search (Scheme 2). Other ways of using prosodic modeling information in either linguistic decoding or acoustic decoding is worth further exploring.

Acknowledgements

The database was provided by Chunghwa Telecommunication Laboratories and the basic lexicon was supported by Academia Sinica of Taiwan.

References

Bai, B.R., Tseng, C.Y., Lee, L.S., 1997. A multi-phase approach for fast spotting of large vocabulary Chinese keywords from Mandarin speech using prosodic information. In: Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 903–906.

Batliner, A., Kompe, R., Kießling, A., Niemann, H., Nöth, E., 1996. Syntactic-prosodic labeling of large spontaneous speech data-base. In: Proc. Int. Conf. On Spoken Language Process. (ICSLP), Vol. 3, pp. 1720–1723.

Bou-Ghazale, S.E., Hansen, J.H.L., 1998. HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress. IEEE Trans. on Speech and Audio Processing 6 (3), 201–216.

Campbell, W.N., 1993. Automatic detection of prosodic boundaries in speech. Speech Communication 13, 343–354.

Chen, S.H., Hwang, S.H., Wang, Y.R., 1998. An RNN-based prosodic information synthesizer for Mandarin text-to-speech. IEEE Trans. on Speech and Audio Process. 6 (3), 226–239.

Cheng, C.C., 1973. A Synchronic Phonology of Mandarin Chinese. Mouton, The Hague.

Chiang, T.H., Lin, T.H., Su, K.Y., 1996. On jointly learning the parameters in a character-synchronous integrated speech and language model. IEEE Trans. Speech and Audio Proc. 4 (3), 167–189.

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

Elman, J., 1990. Finding structure in time. Cognitive Science 14, 179–211.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., Batliner, A., 1996. Consistency in transcription and labeling of German intonation with GToBI. In: Proc. Int. Conf. On Spoken Language Process. (ICSLP), pp. 1716–1719.

Haykin, S., 1994. Neural Networks – A Comprehensive Foundation. Macmillan College Publishing Company, New York.

Hirose, K., Iwano, K., 1997. A method of representing fundamental frequency contours of Japanese using statistical models of moraic transition. In: Proc. Eur. Conf. on Speech Commun. Technol. (EUROSPEECH), Vol. 1, pp. 311–314.

Hirose, K., Iwano, K., 1998. Accent type recognition and syntactic boundary detection of Japanese using statistical modeling of moraic transitions of fundamental frequency contours. In: Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP), Vol. 1, pp. 25–28.

Hsieh, H.Y., Lyu, R.Y., Lee, L.S., 1996. Use of prosodic information to integrate acoustic and linguistic knowledge in continuous mandarin speech recognition with very large vocabulary. In: Proc. Int. Conf. On Spoken Language Process. (ICSLP), Vol. 1, pp. 809–812.

Huang, E.F., Wang, H.C., 1994. An efficient algorithm for syllable hypothesization in continuous Mandarin speech recognition. IEEE Trans on Speech and Audio Process 2 (3), 446–449.

Hunt, A., 1994. A generalized model for utilizing prosodic information in continuous speech recognition. In: Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP), Vol. II, pp. 169–172.

Iwano, K., 1999. Prosodic word boundary detection using mora transition modeling of fundamental frequency contours speaker-independent experiments. In: Proc. Eur. Conf. On Speech Commun. Technol. (EUROSPEECH), Vol. 1, pp. 231–234.

Iwano, K., Hirose, K., 1998. Representing prosodic words using statistical models of moraic transition of fundamental frequency contours of Japanese. In: Proc. Int. Conf. On Spoken Language Process. (ICSLP), Vol. 3, pp. 599–602.

Iwano, K., Hirose, K., 1999. Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Vol. 1, pp. 133–136.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Schukat-Talamazzini, E.G., Zottmann, A., Batliner, A., 1995. Prosodic scoring of word hypotheses graphs. In: Proc. Eur. Conf. On Speech Commun. Technol. (EUROSPEECH), Vol. 2, pp. 1333–1336.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Batliner, A., Schachtl, S., Ruland, T., Block, H.U., 1997. Improving parsing of spontaneous speech with the help of prosodic boundaries. In: Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 811–814.

Lyu, R.Y., Chien, L.F., Hwang, S.H., Hsieh, H.Y., Yang, R.C., Bai, B.R., Weng, J.C., Yang, Y.J., Lin, S.W., Chen, K.J., Tseng, C.Y., Lee, L.S., 1995. Golden Mandarin (III) a user-adaptive prosodic-segment-based mandarin dictation machine for Chinese language with very large vocabulary. In: Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 57–60.

Markel, J.D., Gray Jr., A.H., 1976. Linear Prediction of Speech. Springer, Berlin.

Morgan, D.P., Scofield, C.L., 1991. Neural Networks and Speech Processing. Kluwer, Dordrecht.

Niemann, H., Nöth, E., Kießling, A., Kompe, R., Batliner, A., 1997. Prosodic processing and its use in verbmobil. In: Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process (ICASSP), pp. 75–78.

Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., Fong, C., 1991. The Use of prosody in syntactic disambiguation. J. Acoust. Soc. Am. 90 (6), 2956–2970.

Robinson, A.J., 1994. An application of recurrent nets to phone probability estimation. IEEE Trans. on Neural Networks 5 (2), 298–305.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. TOBI: a standard for labeling English prosody. In: Proc. Int. Conf. On Spoken Language Processing (ICSLP), Vol. 2, pp. 867–870.

Su, Y.S., 1994. A study on automatic segmentation and tagging of Chinese sentence. Master Thesis, National Chiao Tung University, Taiwan, ROC.

Wang, Y.R., Chen, S.H., 1994. Tone recognition of continuous mandarin speech assisted with prosodic information. J. Acoust. Soc. Am. 96 (5), 2637–2645.

Wightman, C.W., Ostendorf, M., 1994. Automatic labeling of prosodic patterns. IEEE Trans. Speech and Audio Proc. 2 (4), 469–480.

Wu, Z., 1998. The formalization of segmental coarticulatory variants in Chinese synthesis system. In: Proc. Conf. on Phonetics of the Languages in China, pp. 125–128.

Yang, Y.J., Lin, S.C., Chien, L.F., Chen, K.J., Lee, L.S., 1994. An intelligent and efficient word-class-based Chinese language model for Mandarin speech recognition with very large vocabulary. In: Proc. Int. Conf. On Spoken Language Process. (ICSLP), Vol. 3, pp. 1371–1374.

Yin, Y.M., 1989. Phonological Aspects of Word Formation in Mandarin Chinese. University Microfilms International.