Cover Page

- *Submitted to* **ICS2002 Workshop on Artificial Intelligence**
- *Title of the paper:* **The Study of Prosodic Modeling for Mandarin Speech**
- *Abstract:*

Prosodic modeling plays an important role in the integration of speech recognition and natural language understanding. The function of prosodic modeling for the integration of speech recognition and natural language understanding 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 build a model describing the relationship between input prosodic features extracted from the testing utterance and output linguistic features of the associated text. In this paper, three prosodic modeling approaches based on the VQ, SOFM, and RNN techniques have been discussed in detail. Experimental results have shown that they all functioned well on detecting prosodic states from acoustic cues. By evaluating on perplexity reduction, we found that the RNN-based approach outperformed the other two approaches.

- *Authors:* Wern-Jun Wang[1] and Sin-Horng Chen[2]

[1] Internet & Multimedia Application Technology Laboratory,
Chunghwa Telecommunication Laboratories, Taiwan, R.O.C.
[2] Department of Communication Engineering,
National Chiao Tung University, Taiwan, R.O.C.

Tel: +886-3-5731822, Fax: +886-3-5710116, Email: schen@cc.nctu.edu.tw
Tel:+886-3-4244536, Fax:+886-3-4244619, Email: wernjun@cht.com.tw

- *Contact author:*
Wern-Jun Wang **王文俊**
網際網路室
中華電信研究所
桃園縣楊梅鎮民族路五段 551 **巷** 12 **號**
TEL: 886 3 4244536
FAX: 886 3 4244619
email: wernjun@cht.com.tw

- *Keywords:* **Speech Recognition, Prosody, Feature Map, Neural Network**

1. Introduction

As for the research of spoken language understanding as shown in Fig. 1, it can be resolved to two major component technologies: speech recognition and natural language understanding. From this figure, we can find that prosodic information can be regarded as a by-product (A_1) of speech recognition and used to help natural language understanding (A_2). Prosodic information is known to carry some information related to syntax and semantics of the text associated with the testing utterance. Besides, prosodic information may also provide additional information (B) to help improving speech recognition. On the other hand, natural language understanding can provide syntax and semantics-related information (C_1). These additional knowledge sources are certainly helpful for improving speech recognition (C_2). This is especially true for speech recognition in domain-specific applications. The above-mentioned integration of speech recognition and natural language understanding is analogous to human perception process. However, even many impressive achievements have been attained in spoken language understanding after a long tradition of studies and experiments, current technology is still far from human-like. Most problems are solved based on engineering approaches without considering human perception process. For example, prosodic information was rarely used in speech recognition (B), not to mention about the incorporation of it into natural language understanding (A_2). In this paper, we will focus on the discussions about the generation of prosodic information (A_1). The incorporation of prosodic information into the speech-to-text conversion task of natural language understanding (A_2) can be found in the study of Wang et al. (2002).

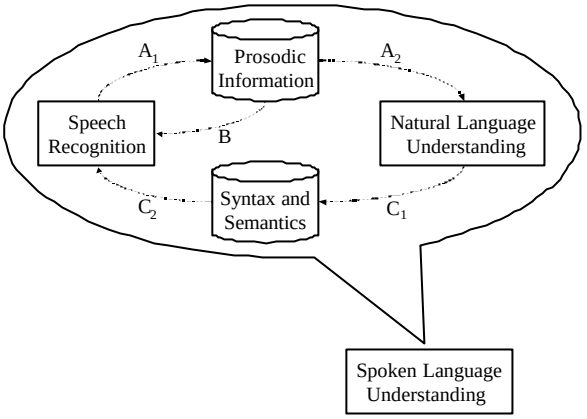

Fig. 1 The description of spoken language understanding.

The function of prosodic modeling for the integration of speech recognition and natural language understanding 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 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 statis tical 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.

In the following sections, three prosodic modeling approaches, such as vector quantization (VQ), self-organizing feature map (SOFM), and recurrent neural network (RNN), are proposed. According to the comparisons between the performances attained by these three approaches, the RNN-based prosodic modeling approach outperformed the other two approaches. Comparing with previous studies (Batliner et al., 1996; Grice et al., 1996; Kompe et al., 1997; Silverman et al., 1992), the distinct property of the proposed methods 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. Firstly, it uses VQ, SOFM or RNN to automatically learn to do prosodic phrasing of Mandarin utterance and implicitly stores the mapping within the internal representation. No explicit prosodic labeling of the speech signals is needed. Secondly, it is easy to incorporate the prosodic model into the linguistic decoder by directly taking the outputs of prosodic model as additional scores (RNN), or by using the outputs of prosodic model to drive a finite state machine (FSM) for setting path constraints to restrict the linguistic decoding search (VQ, SOFM, RNN) (Wang et al. 2002). Both of them can cope with the performance degradation on word boundary detection caused by pronouncing a word chunk within a prosodic phrase. Thirdly, 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.

2. The Descriptions of Three Prosodic Modeling Approaches

Many prosody-related researches provide a mechanism for mapping sequence of observations as a vector of acoustic features to prosodic label. Features used in these studies are selected based on their close 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 previous 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 the following study, we choose syllable segment as the basic unit to extract features for prosodic modeling because syllable is the basic pronunciation unit of Mandarin speech. Four prosodic features are extracted for each syllable segment. They include three features representing the F0 values of the starting and ending points and the mean of the F0

contour, and one feature representing log-energy mean of the *final* part of the current syllable segment. Furthermore, to effectively model the variation between speech segments, the feature vector of each syllable segment is constructed by combining its prosodic features and those of the following syllable. Besides, one additional feature, which represents the duration of the silence between these two syllable segments, is used. In this way, there are in total 9 prosodic features used for detecting the prosodic state of an inter-syllable segment interval. Note that special treatment must be given to the last syllable of each utterance because there are no prosodic features of the following syllable segment associated with it. The alternative we take is the average mean of the corresponding prosodic feature of the starting syllable segment of all utterances in the training database. The reasons of using these 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 silence between syllable segments is useful in identifying the ending point of a prosodic phrase because the pause usually can be found at the last syllable of a prosodic phrase.

2.1 Vector Quantization

As the feature vector has been defined, the most straightforward approach of prosodic modeling is to employ the vector quantization (VQ) technique to classify prosodic feature vectors into different classes. In the training phase, all prosodic feature vectors in the training database are collected and used to train a codebook by the LBG algorithm (Linde et al., 1980). The number of codewords which implies the number of prosodic states used is determined empirically and set to be 8 in this study. In the testing phase, we can assign a prosodic state to each input feature vector via VQ encoding. The detailed analyses of the VQ approach in prosodic state detection will be presented in Section 3.

2.2 Self-organizing Feature Map

In the past, SOFM has been used in many phoneme recognition applications (Morgan and Scofield, 1991). A special characteristic of SOFM is that it is an unsupervised clustering technique with no need of the a priori knowledge about the training samples. This can lower the load of prosodic modeling by skipping the prosodic labeling process of the training data. As shown in Fig. 2, the SOFM we used is a one-dimensional 8-node network with 9 input prosodic features. It is similar to a two-layer neural network with full connections between the nodes of output layer and input layer. The training procedure is to sequentially present the input feature vector $f(t)$ to the network. The node of output layer, which minimizes the following distance, is selected as the winning node (prosodic state), i.e.

$$n_{win} = \arg\min_i \left| f(t) - w_i(t) \right| \tag{1}$$

where $w_i(t)$ is the weighting vector connecting to the *i*-th output node. The feature vector $f(t)$ and the weighting vector $w_i(t)$ are normalized by their respective norms before the calculation of Eq. (1). It is noted that the decision criterion can be based on either the differences or the inner products of the input vector and those weighting vectors. The winner node is then selected as the center of a modification region in the network. The weighting vectors of the winner node, n_{win}, and each of its spatial neighbors in a region $R(t)$ are adjusted according to the following rule:

$$w_i(t+1) = w_i(t) + h(t)\,(f(t) - w_i(t)) \tag{2}$$

where $h(t)$ stands for the learning rate. The weighting vectors of the nodes which fall outside the region $R(t)$ are not modified. The modification range $R(t)$ shrinks from 2 in the beginning stage to 1 in the middle stage, and to 0 in the final stage of the

training process. The different modification ranges are also displayed in Fig. 2 via using different gray levels to show the neighboring nodes of the winning node.

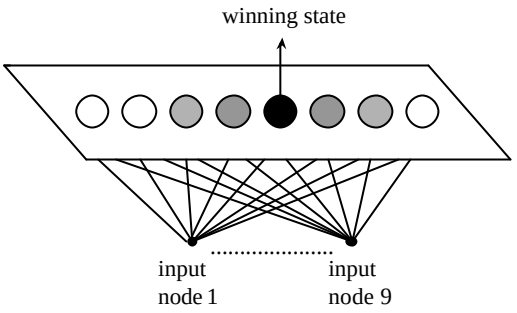

Fig. 2 The structure of self-organizing feature map.

2.3 Recurrent Neural Network

Different from the above two unsupervised clustering methods, VQ and SOFM, the RNN-based approach is arranged by using acoustic features carrying prosodic information as input and setting some appropriate linguistic features extracted from the associated text as output targets in the training phase. Fig. 3 shows the architecture of the prosodic-modeling 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 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. In our study, the inputs of the RNN consist of the same 9 prosodic features used in the VQ and SOFM approaches. Five output linguistic features extracted from the context of the current inter-syllable boundary. They include one indicator showing whether the current inter-syllable boundary is an inter-word boundary or an intra-word boundary; two indicators showing whether the current inter-syllable boundary is a left boundary and a right boundary of a long word with length greater than or equal to 3 syllables; one indicator showing whether there exists a punctuation mark (PM) in the current inter-syllable boundary; and one indicator showing whether the current syllable is of a neutral tone. The RNN can be trained by the back propagation through time (BPTT) algorithm (Haykin, 1994). The prosodic state is obtained by vector quantizing the outputs of its hidden layer.

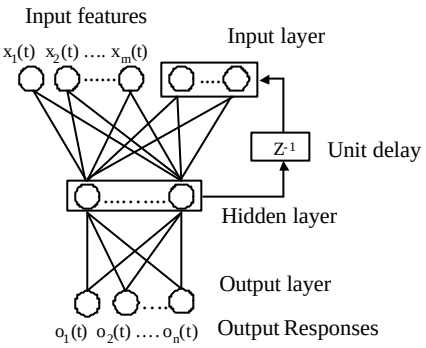

Fig. 3 The structure of the prosodic-modeling RNN.

3. Performance Evaluation of the Three Approaches

Effectiveness of the three proposed approaches was examined by simulations on a speaker-dependent continuous Mandarin speech database. It is a speaker-dependent continuous speech database provided by Chunghwa Telecommunication Laboratories.

The database contained 652 continuous utterances including 452 sentential utterances and 200 paragraphic utterances. All speech signals were recorded with a sampling rate of 20 kHz. 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 extracted from a large news corpus to cover a variety of subjects were 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.

All speech utterances were segmented into syllable sequences based on an HMM base-syllable recognizer. The F0 contour of each utterance was detected by the simplified inverse filter tracking (SIFT) algorithm (Markel and Gray, 1976) with manual correction. To prepare output targets for training the RNN, texts associated with all training utterances were tokenized into word sequences in advance by an automatic statistical model-based algorithm with a long-word-first criterion and several simple word-merging rules (Su, 1994). A lexicon containing 111,243 words was used in the tokenizing process.

We first examined the performance of the prosodic state classification using the VQ approach. Fig. 4(a) shows the VQ codewords associated with the 8 prosodic states. There are two segments for each codeword representing, respectively, the current syllable *final* segment and the following syllable *final* segment. Each segment has three values standing for the F0 values of the starting point, the mean, and the ending point of its F0 contour. By examining these 8 codewords, we find that a meaningful prosodic phrases structure can be interpreted by using these 8 prosodic states. Specifically, the first two states correspond to the beginning part of a prosodic phrase; State 5 corresponds to the ending part; and States 3, 6, and 7 correspond to the intermediate part. Besides, States 2 and 4 correspond to minor break and State 5 corresponds to major break. The same analyses can be applied to interpret functions of

the F0-related weighting coefficients of the SOFM approach shown in Fig. 4(b). From this figure, we can find that States 1 and 6 correspond to the beginning part of a prosodic phrase; State 5 corresponds to the ending part; and States 3, 4, and 7 correspond to the intermediate part. Besides, States 0 and 2 correspond to minor break and State 5 corresponds to major break.

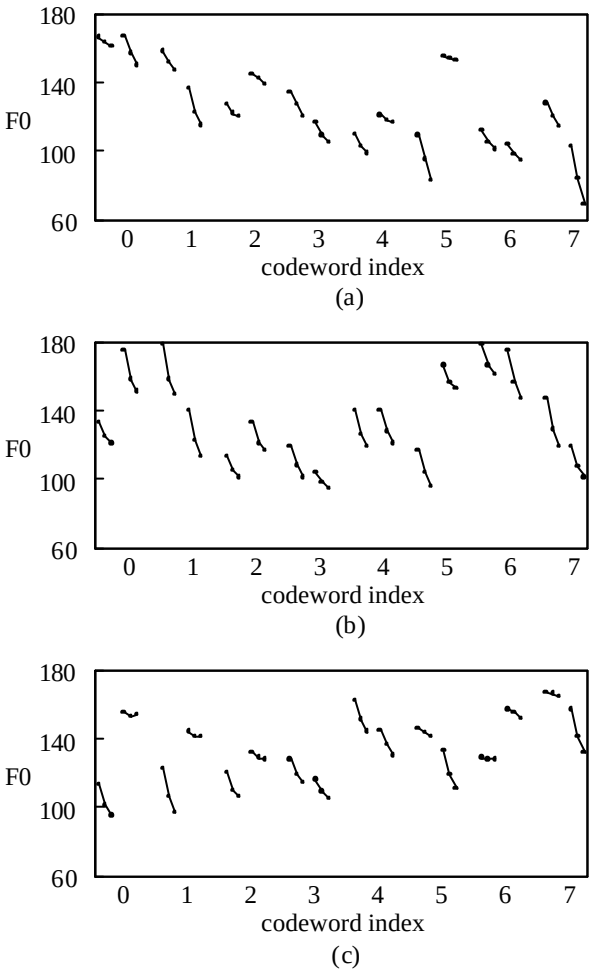

Fig. 4 The plots of F0-related parameters of 8 prosodic states based on (a) VQ, (b) SOFM, and (c) RNN approaches.

Different from the above discussions for the VQ and SOFM approaches, it is difficult to find the physical meaning directly from the hidden layer outputs in the RNN-based approach. We therefore adopt a backward analysis process to find the related F0 parameters of 8 prosodic states. All training syllables which were assigned to the same prosodic state were clustered and used to calculate the average value of starting, mean and ending F0 value of the F0 contour of the current and following syllable *final* segments. The F0-related parameters of these 8 prosodic states are displayed in Fig. 4(c). From this figure, we find that State 7 corresponds to the beginning part of a prosodic phrase; State 0 corresponds to the ending part; and States 3, 4, and 5 correspond to the intermediate part. Besides, States 1 and 2 correspond to minor break and State 0 corresponds to major break. It is noted that almost all the simplified 3-point F0 contours shown in Fig. 4 demonstrate a declination phenomenon. The high-rising and low-dipping phenomena of F0 contour, respectively, for Tone 2 and Tone 3 were not found in this figure. The reason is that the features of all training syllables with the same prosodic state were clustered to generate the features of the corresponding codeword. These syllables with the same prosodic state were associated with different lexical tones. In addition, the F0 contours of all syllables in a sentential utterance were also seriously adjusted to meet the declined intonation pattern of sentence. The amalgamated F0 contour patterns were thus obtained as those shown in Fig. 4.

The prosodic state sequences of some typical sentences labeled by the VQ, SOFM, and RNN approaches are shown in Fig. 5. The number within a parenthesis stands for the encoded prosodic state associated with the preceding character. From this figure, we can find that the encoded prosodic states of most characters were properly assigned because their functions match quite well with our priori linguistic knowledge. However, some prosodic states with the function of minor break have been assigned to intermediate characters of polysyllabic words, such as **行**, **閃**, **找**, **麻**, **期**, and **粉**. Obviously, they are false alarms. An interesting phenomenon of repeating prosodic states of 4, 6, and 7 can be found in the example for the SOFM approach. This is relatively infrequent to occur for both the VQ and RNN approaches.

工(0)業(1)時(1)代(2)分(1)秒(2)必(3)爭(2)行(0)動(0)必(0)須(0)快(1)如(3)閃(4)電(5)
別(2)自(1)找(4)麻(4)煩(6)地(4)把(3)水(2)噴(3)到(2)篩(3)好(6)的(4)麵(6)粉(6)上(5)
另(0)為(0)防(1)止(2)新(0)春(1)期(2)間(2)民(2)眾(1)酒(4)後(2)開(0)車(1)肇(3)事(5)
(a)

工(6)業(6)時(1)代(0)分(1)秒(4)必(4)爭(0)行(6)動(6)必(6)須(6)快(1)如(7)閃(7)電(5)
別(6)自(1)找(4)麻(4)煩(7)地(2)把(4)水(4)噴(4)到(4)篩(4)好(3)的(2)麵(3)粉(2)上(5)
另(6)為(6)防(1)止(0)新(6)春(1)期(4)間(4)民(0)眾(1)酒(4)後(4)開(6)車(6)肇(1)事(5)
(b)

工(7)業(4)時(3)代(1)分(7)秒(2)必(3)爭(0)行(6)動(4)必(3)須(2)快(5)如(3)閃(2)電(0)
別(7)自(4)找(2)麻(3)煩(3)地(2)把(5)水(2)噴(5)到(2)篩(5)好(3)的(2)麵(3)粉(2)上(0)
另(7)為(4)防(5)止(2)新(5)春(4)期(3)間(1)民(6)眾(4)酒(2)後(2)開(5)車(5)肇(3)事(0)
(c)

Fig. 5 The prosodic state sequences of three typical sentences labeled by the (a) VQ, (b) SOFM, and (c) RNN approaches.

Another measure we used to compare the performances of these three approaches is the reduction in perplexity. It measures the difference between the uncertainties with and without applying a prosodic model. First, the uncertainty of a target linguistic feature observation, *O*, without applying any prosodic model is defined by

$$2^{H_{org}(O)} \tag{3}$$

where

$$H_{org}(O) = -\sum_{i=1}^{N} P(o_i) \log P(o_i), \tag{4}$$

o_i is the *i*-th outcome of the observation *O* and *N* is the total number of outcomes in *O*. The new measure of the uncertainty obtained after applying a prosodic model is then defined by

$$2^{H_{new}(O)} \tag{5}$$

where

$$H_{new}(O) = \sum_{j=1}^{M} P(s_j)\, H(O \mid s_j) \tag{6}$$

is the weighted sum of conditional entropies related to prosodic state, s_j stands for the *j*–th prosodic state, and *M* is the total number of prosodic states. The conditional entropy $H(O \mid s_j)$ used in Eq. (6) is defined by

$$H(O \mid s_j) = -\sum_{i=1}^{N} P(o_i \mid s_j) \log P(o_i \mid s_j) \tag{7}$$

The analyses of perplexity reduction for five linguistic features, including the tone of syllable, the right boundary of long word, the left boundary of long word, the existence of punctuation mark, and the inter-word boundary, are shown in Table 1. It is noted that the number of outcomes (*N* in Eqs. (5) and (7)) for tone is 5 while those for all other 4 features are 2. We can find from Table 1 that, among these 5 features, better improvements can be attained for the predictions of inter-word boundary and PM. The prediction results are better for the right boundary of long word than those for the left boundary of long word. Moreover, it can be found that the performance of the RNN-based approach is better than both of the VQ and SOFM approaches.

Table 1. Experimental results of the perplexity measurements for five linguistic features with and without using prosodic models.

Data	Method	Target Observatio				
		Tone	Long-r	Long-l	PM	Inter
Training	Original	4.36	1.53	1.54	1.35	1.96
	VQ	3.68	1.45	1.50	1.16	1.79
	SOFM	3.66	1.43	1.49	1.12	1.76
	RNN	3.62	1.40	1.47	1.07	1.69
Test	Original	4.38	1.51	1.52	1.30	1.96
	VQ	3.68	1.45	1.49	1.17	1.84
	SOFM	3.65	1.43	1.48	1.13	1.80
	RNN	3.62	1.40	1.46	1.08	1.73

4. Conclusions

In this paper, three prosodic modeling approaches based on the VQ, SOFM, and RNN techniques have been discussed in detail. Experimental results have shown that they all functioned well on detecting prosodic states from acoustic cues. By evaluating on perplexity reduction, we found that the RNN-based approach outperformed the other two approaches.

5. References

Batliner, A., Kompe, R., Kießling, A., Niemann, H. and Nöth, E. (1996), "Syntactic-Prosodic Labeling of Large Spontaneous Speech Data-Base," in *Proc. Int. Conf. On Spoken Language Processing (ICSLP)*, Vol. 3, pp. 1720-1723.

Bou-Ghazale, S. E. and 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*, Vol. 6, No. 3, pp.201-216.

Campbell, W. N. (1993), "Automatic Detection of Prosodic Boundaries in Speech," *Speech Communication*, Vol. 13, pp.343-354.

Chen, S. H., Hwang, S. H., and Wang, Y. R. (1998), "An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 3, pp.226-239.

Chen, S. H., Liao, Y. F., Chiang, S. M. and Chang, S. (1998), "An RNN-based Pre-Classification Method for Fast Continuous Mandarin Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 1, pp.86-90.

Elman, J. (1990), "Finding Structure in Time," *Cognitive Science*, Vol. 14, pp. 179-211.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J. and Batliner, A. (1996), "Consistency in Transcription and Labeling of German Intonation with GToBI," in *Proc. Int. Conf. On Spoken Language Processing (ICSLP)*, pp. 1716-1719.

Haykin, S. (1994), *Neural networks – A comprehensive foundation*, Macmillan College Publishing Company.

Hirose, K. and Iwano, K. (1997), "A Method of Representing Fundamental Frequency Contours of Japanese Using Statistical Models of Moraic Transition," in *Proc. European Conf. On Speech Communication and Technology (EUROSPEECH)*, Vol. 1, pp. 311-314.

Hirose, K. and 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.

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. and 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 Processing (ICSLP)*, Vol. 3, pp. 599-602.

Iwano, K. and 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.

Iwano, K. (1999), "Prosodic Word Boundary Detection Using Mora Transition Modeling of Fundamental Frequency Contours – Speaker Independent Experiments -," in *Proc. European Conf. On Speech Communication and Technology (EUROSPEECH)*, Vol. 1, pp. 231-234.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Schukat-Talamazzini, E. G., Zottmann, A. and Batliner, A. (1995), "Prosodic Scoring of Word Hypotheses Graphs," in *Proc. European Conf. On Speech Communication and Technology (EUROSPEECH)*, Vol. 2, pp.1333-1336.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Batliner, A., Schachtl, S., Ruland, T. and 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.

Linde, Y., Buzo, A., and Gray, R. M. (1980), "An algorithm for vector quantizer design," *IEEE Trans. On Commun*., vol. COM-28, no. 1, pp. 84-95.

Markel, J. D. and Gray, Jr. A. H. (1976), *Linear Prediction of Speech*, Springer-Verlag.

Morgan, D. P. and Scofield, C. L. (1991), *Neural Networks and Speech Processing*, Kluwer Academic Publishers.

Niemann, H., Nöth, E., Kießling, A., Kompe, R. and 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. and Fong, C. (1991), "The Use of Prosody in Syntactic Disambiguation," *J. Acoust. Soc. Am.*, Vol.90, No. 6, pp. 2956-2970.

Robinson, A. J. (1994), "An application of recurrent nets to phone probability estimation," *IEEE Trans. on Neural Networks*, Vol. 5, No. 2, pp. 298-305.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and 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 in Taiwan, R. O. C.

Wang, W. J., Liao, Y. F. and Chen, S. H. (2002), "RNN-based Prosodic Modeling of Mandarin Speech and Its Application to Speech-to-Text Conversion," *Speech Communication*, Vol.36, No.3-4, pp.247-265.

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