A First Study on Mandarin Prosodic State Detection

Yuan-Fu. Liao, Wern-Jun Wang, Shu-Ling Lee, Sin-Horng Chen
Institute of Communication Engineering
National Chiao Tung University, HsinChu, Taiwan, ROC
TEL: +886-3-5711431, FAX: +886-3-5710116, Email: schen@cc.nctu.edu.tw

Abstract

In this paper, a method to detect prosodic phrase structure of Mandarin speech is proposed. It first employs an RNN to discriminate each input frame of an utterance among three broad classes of syllable initial, syllable final, and silence. Outputs of the RNN are then used to drive an FSM for segmenting the input utterance into four types of segment. They include three stable-segment - I (initial), F (final), and S (silence), and a transition-segment - T (transition). Appropriate modeling features are thus extracted from the vicinities of F-segments, and used to model the prosodic states for inter-F-segment intervals. Two prosodic-state modeling schemes are studied. One uses VQ to encode the modeling features and directly classify inter-F-segment intervals into 8 prosodic states. The other uses an RNN, trained with relevant linguistic features as output targets, to implicitly represent the prosodic status by the outputs of its hidden layer. Prosodic states can be obtained by vector-quantizing the outputs of the hidden layer of the RNN. Experimental results showed that linguistically meaningful interpretations of these prosodic states can be observed.

1. Introduction

Continuous speech contains the actual words spoken as well as supra-segmental information, such as stress, timing structure, and fundamental frequency (F0) contour patterns. This information is generally referred to as the prosody of the speech, which is affected in turn by the sentence type, the syntactical structure, the semantics, the emotional state of the speaker, etc [Sagisaka 1996]. Traditional speech recognition methods have totally neglected this prosodic information in their recognition process. However, prosodic modeling gets more and more attentions in recent years in the area of speech recognition. Many researches have been devoted to the exploration of relevant prosodic cues from input speech utterance with the purpose of assisting in speech recognition [Compbell 1993, Kompe 1995, Wang 1994, Wightman 1994]. A prosodic model can be generally defined as a mechanism to describe the relationship between the acoustic features extracted from the prosodic parameter contours representing the prosodic phrase structure of speech and the linguistic features extracted from the associated text. Two basic types of prosodic model can be found. One is a model designed to detect the prosodic phrase structure of an utterance by using some features extracted from the prosodic parameter contours. Its main purpose is to provide an additional score to help speech recognition. The other type of prosodic model is designed to predict the embedded prosodic phrase structure from a text by using some linguistic features extracted from the text. Obviously, it is mainly used in text-to-speech to help the generation of prosodic information for synthesizing natural speech [Chen 1996a].

In this paper, we are interested in the first type of prosodic model. A method to detect the prosodic phrase structure of the input speech is proposed. Our final goal is to derive a prosodic model in the pre-processing stage of a speech recognizer for the use in the following recognition process to assist in speech recognition. In this preliminary study, only the prosody modeling is discussed. The primary problem encountered in this study is how to extract appropriate modeling features from the input speech under the constraint that *a priori* information about syllable or word boundaries is not available [Wightman 1994]. The problem is solved by first dividing the input utterance into labeled-segments, and then extract modeling features from stationary voiced segments.

The organization of the paper is stated as follows. Section 1 briefly describes background information and states the problem. Section 2 presents the proposed method of prosodic-state detection. Experimental results are given in Section 3. Conclusions are given in the last section.

2. The proposed method

Fig. 1 shows a block diagram of the proposed method of prosodic-state detection. It consists of five main parts: spectral feature extraction, recurrent neural network (RNN) pre-classifier, finite state machine (FSM) based segmentation [Chen 1996b], modeling feature extraction, and prosodic state classification. Input speech signal is first divided into frames. Some spectral features are then extracted for each frame. An RNN is then employed to discriminate each input frame among three broad classes. They include syllable initial, syllable final, and silence. Outputs of the RNN are then used to drive an FSM to segment the input utterance into four types of segment. They include three stable-segment types of I (initial), F (final), and S (silence), and a transition-segment type of T (transition). Appropriate modeling features are then extracted from the vicinities of F-segments, and used to model the prosodic states for inter-F-segment interval. Two prosodic-state modeling schemes are proposed. One uses VQ to directly classify input features of two contiguous F-segments into 8 prosodic states. The other uses an RNN, trained with relevant linguistic features as output targets, to implicitly represent the prosodic status by the outputs of its hidden layer. Finite number of explicit prosodic states can then be obtained by vector-quantizing the outputs of the hidden layer of the RNN. In the following, these five parts are discussed in more detail.

2.1. Spectral feature extraction

A pre-processing was used to extract spectral features for RNN-based pre-classification. In the pre-processing, a short-time spectral analysis by 256-point FFT was performed for each of 200-sample frame padding with 56 zero-samples. The frame shift is 100 samples. The spectrum of each frame was compressed non-linearly into 20 frequency channels (in mel-scale) using a bank of 20 triangular windows. The energy spectrum was then log-compressed and cosine-transformed to calculate 12 mel cepstral coefficients. Besides, 12 delta mel cepstral coefficients and one delta log-energy were also calculated using a 7-frame window. So, there are in total 25 spectral features used in the pre-classification.

2.2. The RNN pre-classifier

The function of the RNN pre-classifier is to discriminate each input frame among three broad classes of syllable initial, syllable final, and silence. Fig. 2 shows the architecture of the RNN. It is a two-layer network with all outputs of the hidden layer being fed-back to itself as additional inputs [Elman 1990]. The RNN has a distinct property of using its hidden nodes to represent the contextual status of the previous inputs. It is therefore suitable for discriminating dynamic speech patterns [Elman 1991, Robinson 1994]. The RNN can be trained by the back-propagation (BP) algorithm with output targets being set according to the segmentation positions given by an HMM recognizer [Lee 1991].

2.3. FSM-based segmentation

The function of the FSM is to segment the input utterance into I-, F-, S-, and T-segments. The FSM is designed to conform to the phonetic structure of Mandarin base-syllables. Fig. 3 shows the state transition diagram of the FSM. To drive the FSM, all the three outputs of the RNN are compared with two threshold values, TH_L and TH_H. While one output is higher than TH_H and the other two are all lower than TH_L, the FSM moves into the corresponding stable state if it is a legal one. Otherwise, the FSM stays at T state. In this study, TH_L and TH_H were set to be 0.2 and 0.8, respectively. After obtaining the state sequence encoded by the FSM, we divide the input utterance into I-, F-, S- and T-segments.

2.4. Modeling feature extraction

We then extract 4 features relevant to prosody modeling for each F-type segment. They include three features representing the two ending points and the mean of the pitch frequency contour overlapping with the current F-segment and one feature representing log-energy mean of the F-segment. There are in total 9 modeling features used for detecting the prosodic state of an inter-F-segment interval. They include the 8 features of the two neighboring F-segments and one additional feature which represents the duration of the S-segment located between the two F-segments.

2.5. Prosodic state classification

Two prosodic state classification schemes are proposed in this study. One uses VQ to encode the input modeling feature vector and directly classify it as the prosodic state associated with the encoded codeword. The other uses a two-layer RNN, trained with appropriate linguistic features extracted from the associated text as output targets, to implicitly represent the prosodic status of the current inter-F-segment interval by using the outputs of its hidden layer. The RNN has the same structure shown in Fig. 2. The inputs of the RNN consist of the same 9 modeling features used in the first scheme. And 6 output linguistic features are used in this study. Two indicators showing, respectively, whether the current inter-F-segment interval is an inter-word boundary and an intra-word boundary; Two indicators showing whether the current inter-F-segment interval 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-F-segment interval; One indicator showing whether the two neighboring F-segments belong to the same syllable. The prosodic state is finally obtained by vector quantizing the outputs of its hidden layer. Here, the codebook size is also set to 8.

3. Simulations

Effectiveness of the proposed method was examined by simulations. A continuous-speech Mandarin database provided by the Telecommunication Laboratories was used. The database contains 452 sentential utterances and 200 paragraphic utterances. Texts of these 452 sentential utterances are well-designed, phonetically-balanced short sentences with lengths less than 18 characters. Texts of these 200 paragraphic utterances are news selected from a large news corpus to cover a variety of subjects including business, medicine, social event, sport, literature, etc. All utterances were generated by a single 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: a training set and an open test set. These two sets consist of 28060 and 7034 syllables, respectively.

We first examine the performance of the RNN pre-classifier and the FSM-based segmentation. Figs. 4(a-e) shows a typical example. It can be seen from these figures that all syllable finals have been properly detected. Each syllable final has an F-segment associating with it. By carefully examining all segmentation results, we find that the error rate for syllable-final segmentation is about 7%. Two major types of error had been found. One is the missing of an F-segment, i.e., a syllable final has no F-segments assigned to it. This usually occurred for syllables with Tone 5. Another type of error is caused by the missing of an I-segment to make two neighboring finals being connected together to form a long aggregate F-segment. This usually occurred at short voiced initials including nasal, liquid, and null initials. It is noted that these errors are relatively unimportant to our prosodic-state detection because our purpose is to model the global characteristics of the prosody-phrase structure. No significant cues of the prosodic phrase structure are lost due to these errors. So both the RNN and the FSM functioned quite well to meet our requirement [Chen 1996b].

We then examine the performance of the first prosodic state classification scheme using VQ to encode the modeling features. Fig. 5 shows the codewords associated with the 8 prosodic states. By examining the pitch parameters of these codewords, we find that a meaningful mapping between these prosodic states and the prosodic phrases structure does exist. Specifically, the first two states correspond to the beginning part of a prosodic phrase. States 6 and 8 correspond to the ending part. States 4, 5 and 7 correspond to the intermediate part. States 3 and 6 correspond, respectively, to minor and major breaks. Table 1 lists some statistics, including the distributions of beginning and ending of sentential and paragraphic utterances, and the distribution of the existence of a PM in the current inter-F-segment interval. It can be seen from this table that most utterances begin with State 1 and end with State 6. And most PMs are associated with State 6. Table 2 lists the state transition probabilities. It can be found from the table that States 2 and 4 follow State 1 to locate in the beginning part of a prosodic phrase. State 5 follows States 2 and 4. State 7 follows State 4 and 5. States 6 and 8 follow State 7. Fig. 4(h) shows a typical example of the encoded state sequence of an input utterance. And Fig. 6 shows a finite state automata (FSA) obtained by drawing only significant transition probabilities. Based on above discussions, the FSA is a meaningful model to describe the prosodic phrases structure of Mandarin speech [Wang 1994, Chen 1996a, Elman 1990, Elman 1991].

Lastly, we examine the performance of the second RNN-based prosodic-state-detection scheme. The same statistics and state transition probabilities were calculated and listed in Tables 3 and 4. It can be seen from Table 3 that almost all utterances start with State 7 and end with State 2. And most PMs are associated with State 2. As compared with the results shown in Tables 1 and 2, the RNN-based method performed better on prosodic-state detection than the VQ-based method. Similar FSA can also be obtained by counting only significant state transition probabilities (see Fig. 7). Fig. 4(h) shows a typical example of the encoded state sequence of an input utterance. Finally, we examine the outputs of the RNN. Fig. 4(g) shows the outputs of inter-word and intra-word indicators. It can be seen from the figure that significant inter-word response always occurs at a word boundary. This property might be useful to assist in speech recognition.

4. Conclusions

A new prosody modeling method has been discussed in this paper. Experimental results have shown that a meaningful mapping between the resulting detected prosodic states and the prosodic phrase structure can be found. So its performance is quite well. Due to its effectiveness, further studies to incorporate it into a conventional speech recognizer is worthy doing in the future.

Acknowledgment

This work was supported by the National Science Council, Taiwan, ROC under the contract NSC-86-2213-E009-097.

Reference

Chen S. H., Hwang S. H., and Wang Y. R. (**1996a**), "An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech", Accepted by *IEEE Trans. on Speech and Audio Processing.*

Chen S. H., Liao Y. F., Chiang S. M., and Chang S. (**1996b**), ``An RNN-Based Pre-classification Method for Fast Continuous Mandarin Speech Recognition'', Accepted by *IEEE Trans. on Speech and Audio Processing.*

Compbell N. (**1993**), ``Automatic Detection of Prosodic Boundaries in Speech,'', *Speech Communication*, Vol.13, pp.343-354.

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

Elman J. L. (**1991**), ``Distributed Representations, Simple Recurrent Networks, and Grammatical Structure'', *Machine Learning*, Vol.7, pp.195-224.

Kompe R., Kiebling A., Niemann H., Noth E., Schukat-Talamazzini E.G., Zottmann, A. and Batliner A. (**1995**), ``Prosodic Scoring of Word Hypotheses Graphs,'' in *Proc. EUROSPEECH,* pp.1333-1336.

Lee S. J., Kim K. C., Yoon H. and Cho J. W. (**1991**), ``Application of fully recurrent neural networks for speech recognition", *Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 77-80.

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, March.

Sagisaka Y., Campbell N. and Higuchi N. edited (**1996**), ``Computing Prosody - Computational Models for Processing Spontaneous Speech'', *Springer-Verlag*, New York, Inc..

Wang Y. R. and Chen S. H. (**1994**), ``Tone recognition of continuous Mandarin speech assisted with prosodic information", *J. Accoust. Soc. Am.*, **96** (5), Pt. 1, pp. 2637-2645, Nov..

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, Oct..

State Prob. Prosodic cue	1	2	3	4	5	6	7	8
Beginning of utterance	0.64	0.17	0.13	0.03	0.03	0.00	0.00	0.00
Ending of utterance	0.03	0.01	0.06	0.01	0.01	0.85	0.00	0.03
PM	0.04	0.02	0.12	0.02	0.08	0.63	0.01	0.08

Table 1: The statistics of the prosodic states detected by the VQ-based scheme.

Next state Prob. Previous state	1	2	3	4	5	6	7	8
1	0.38	0.37	0.02	0.21	0.01	0.00	0.01	0.00
2	0.02	0.31	0.13	0.24	0.22	0.00	0.08	0.00
3	0.31	0.37	0.01	0.30	0.00	0.00	0.01	0.00
4	0.00	0.00	0.14	0.00	0.40	0.10	0.21	0.16
5	0.00	0.08	0.17	0.10	0.38	0.06	0.20	0.01
6	0.47	0.27	0.11	0.13	0.02	0.00	0.00	0.00
7	0.00	0.00	0.05	0.00	0.04	0.35	0.08	0.48
8	0.00	0.03	0.13	0.05	0.27	0.18	0.19	0.16

Table 2: Prosodic state transition probabilities of the VQ-based scheme.

State Prob. Prosodic cue	1	2	3	4	5	6	7	8
Beginning of utterance	0.00	0.00	0.03	0.00	0.01	0.00	0.96	0.00
Ending of utterance	0.08	0.92	0.00	0.00	0.00	0.00	0.00	0.00
PM	0.17	0.73	0.04	0.02	0.01	0.01	0.00	0.01

Table 3: The statistics of the prosodic states detected by the RNN-based scheme.

Next state Prob. Previous state	1	2	3	4	5	6	7	8
1	0.00	0.01	0.22	0.18	0.08	0.04	0.32	0.14
2	0.00	0.00	0.06	0.02	0.01	0.00	0.76	0.12
3	0.03	0.15	0.08	0.05	0.29	0.11	0.04	0.39
4	0.04	0.01	0.10	0.14	0.05	0.24	0.00	0.28
5	0.14	0.13	0.21	0.07	0.36	0.21	0.00	0.12
6	0.38	0.17	0.09	0.32	0.01	0.31	0.00	0.05
7	0.00	0.00	0.27	0.01	0.55	0.09	0.00	0.06
8	0.03	0.20	0.19	0.29	0.00	0.27	0.00	0.02

Table 4: Prosodic state transition probabilities of the RNN-based scheme.

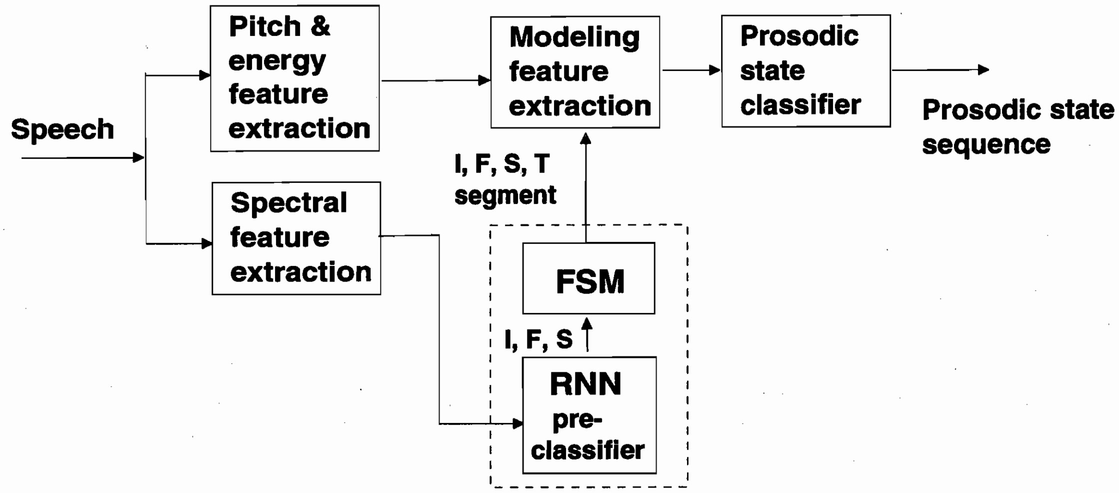

Speech
Pitch & energy feature extraction
Spectral feature extraction
Modeling feature extraction
I, F, S, T segment
FSM
I, F, S
RNN pre-classifier
Prosodic state classifier
Prosodic state sequence

Figure 1: A block diagram of the proposed prosodic-state detection method.

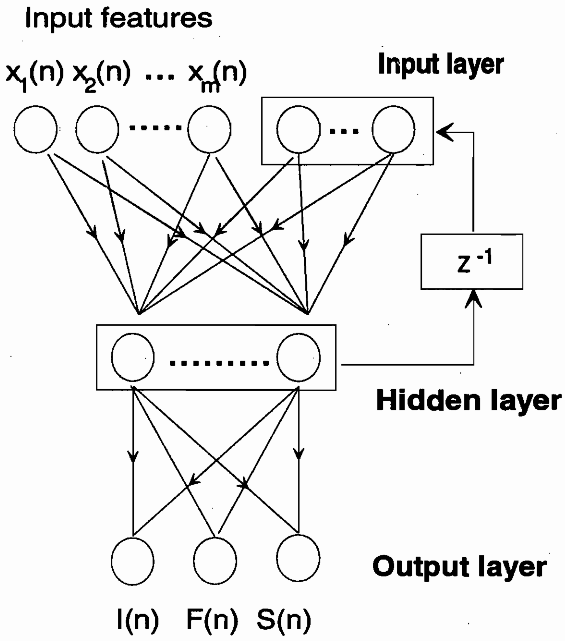


Figure 2: The architecture of the RNN.

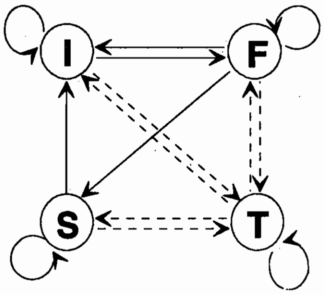
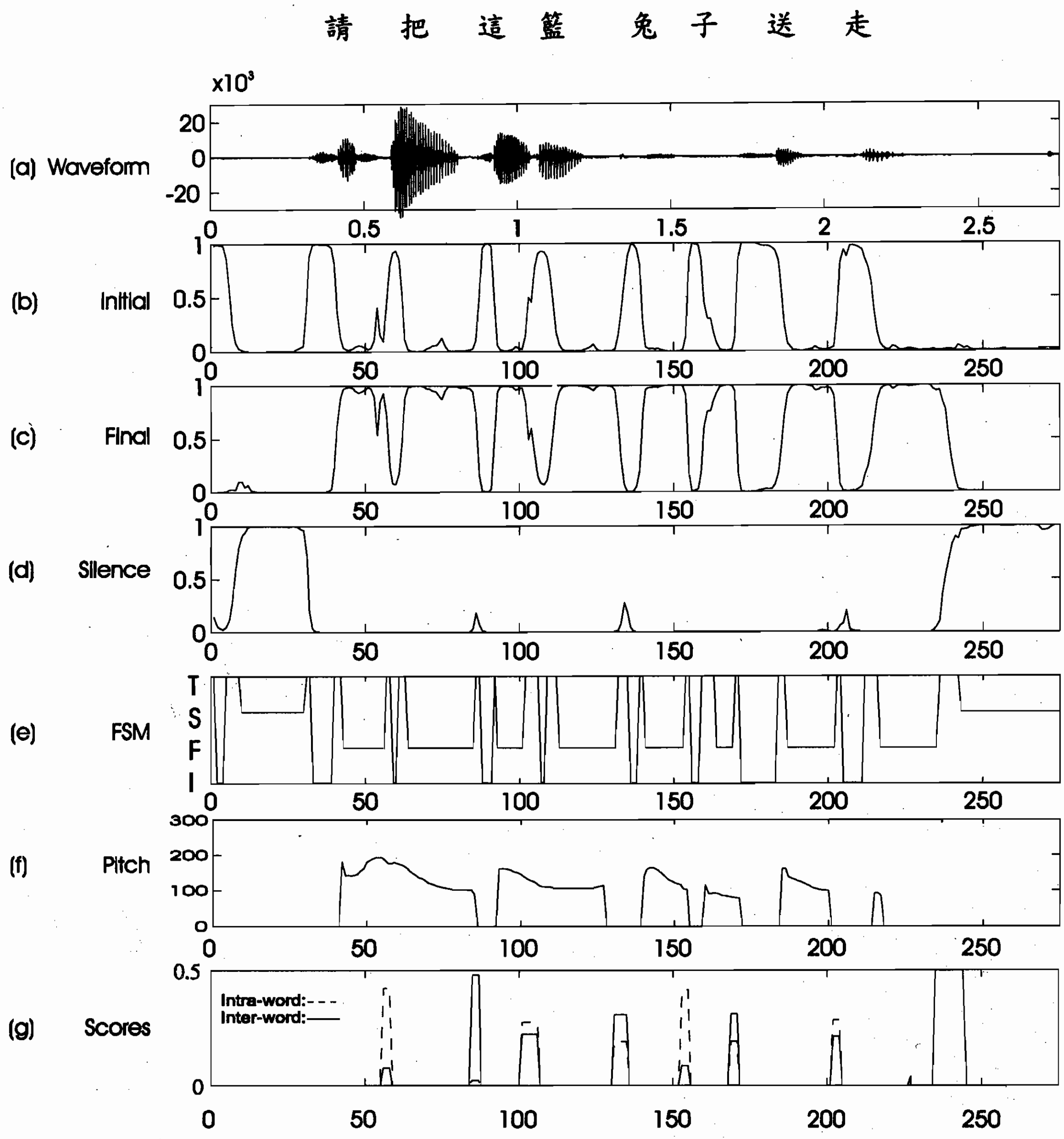


Figure 3: The state transition diagram of the FSM.



(h) Prosodic state sequence :

VQ : 4 5 4 5 7 8 7 6

RNN : 7 3 8 4 8 4 8 2

Figure 4: A typical example of the proposed mrthod : (a) Waveform of the input speech; (b) Initial-, (c) final- and (d) silence-outputs of the RNN pre-classifier; (e) Segmentation results of the FSM; (f) Pitch contour; (g) The inter- and intra-word scores generated by the RNN-based prosodic-state classifier; (h) Prosodic state sequences generated by the VQ-based and RNN-based schemes.

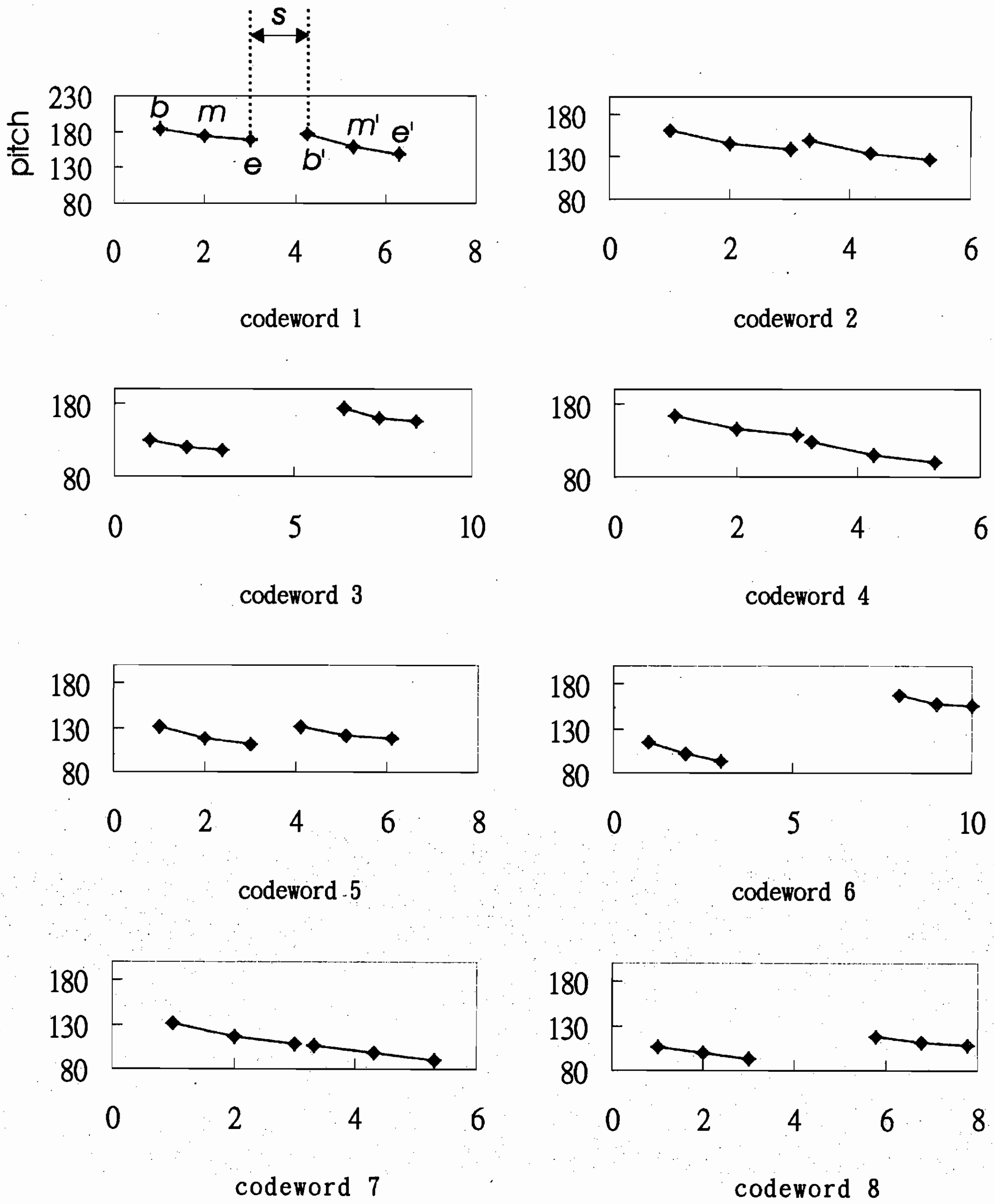


Figure 5: VQ codewords associated with the 8 prosodic states, where *b,b'* are the beginning points of F0 contours, *e, e'* are the ending points of F0 contours, *m, m'* are means of F0 contours, and *s* is the average duration of S-segment.

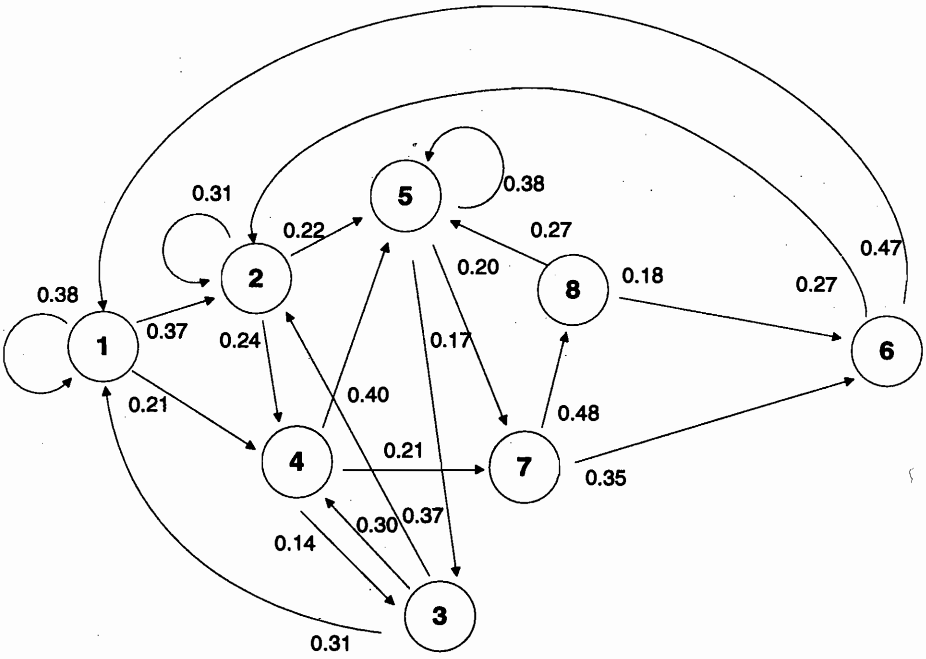


Figure 6: A FSA obtained by the VQ-based prosodic modeling scheme.

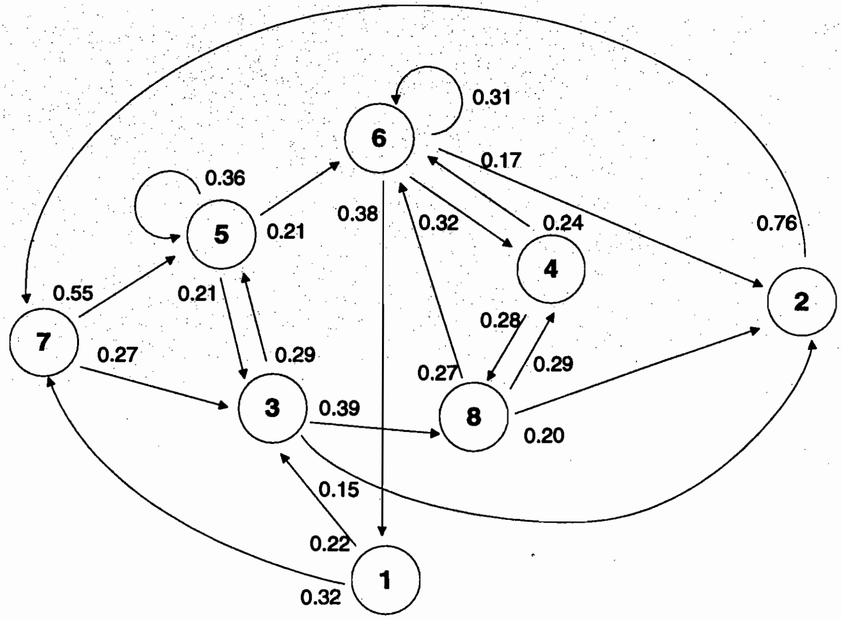


Figure 7: A FSA obtained by RNN-based prosody modeling scheme.