Estimation of Prosodic Information for Persian Text-To Speech System Using a Recurrent Neural Network

Ali Farrokhi[1], Shahrokh Ghaemmaghami, [2] and Mansur Sheikhan[1]

1. Azad University, South Tehran Branch 2. Sharif University of Technology, Tehran, Iran
ali_farrokhi@azad.ac.ir, ghaemmag@sharif.edu, msheikhn@azad.ac.ir

Abstract

A simplified four-layer RNN (recurrent neural network) based architecture is introduced to generate prosodic information for improving naturalness in Persian TTS (text-to-speech) systems. The proposed RNN uses the first two layers at word level and the last two layers at syllable level to provide the TTS system with major prosodic parameters, including: pitch contour, energy contour, length of syllables, length and onset time of vowels, and duration of pauses. The experimental results show improvement of accuracy in prediction of prosodic parameters, as compared to similar prosody generation systems of higher complexity.

1. Introduction

Speech prosody is attributed to the hierarchical structure from speech rhythm and intonation phrase to the smallest components of the syllable. It conveys important information about the suprasegmental features such as F0 (fundamental frequency), intensity, and duration, which have shown to be crucial to natural sounding in TTS systems. The dynamism of prosodic features is very language based, however, the relationship between the linguistic features and the prosodic data is not well known in certain languages. Accordingly, recent findings suggest the use of data-driven methods to generate prosodic information employing neural networks or statistical models to achieve naturalness and fluency in automatic speech synthesizers [1]. While relatively high performance prosody generators have been developed for many languages, very limited work has been done on prosody generation in Persian.

In this paper, we propose a new, simplified RNN-based approach to generate prosodic information to achieve natural sounding in Persian TTS systems. In the proposed method, unlike some other approaches [2-6], all major prosodic parameters are generated simultaneously by the RNN. Besides, in comparison with most such complete prosody generators, which are mostly developed for tonal languages [1], the proposed method does not use any sort of complex syntactic or grammar-based structure, such as major and minor phrases [7], or accent values of syllables [2]. Indeed, very simple inputs at word and syllable levels are used as linguistic features and prosody rules are all embedded in the network weights, which are learnt automatically.

2. Architecture of prosody generator

We use a four layer RNN, as shown in figure 1. For training the network, about 1000 Persian sentences are employed and segmented using the automatic segmentation methodwe developed earlier at phoneme, syllable, and word levels [8]. Prosodic data of each syllable, which are pitch contour, energy contour, duration and place of commencement of vowels, length of syllable, and duration of pauses are extracted from the signal. The inputs and outputs of the neural network are summarized in table 1.

Table 1. *Inputs and outputs of the neural network and the number of associated nodes in RNN*

	Input size	Definition	Symbol
In puts (word layer)	5	Length of current word in terms of syllable counts	L-W-0
	5	Length of next word in terms of syllable counts	L-W-1
	10	Number of words in sentence	NUM-WORD
	10	Position of current word in sentence	POSITION-WORD
	4	Sentence type	TYPE_SEN
	12	POS of current word	POS (Wi)
	12	POS of following word	POS (W_{i+1})
Inputs (syllable layer)	2	Punctuation marks after current syllable(: / ,)	INTERNAL PM
	6	Type of 1st consonant in current syllable	I-0
	6	Type of 1st consonant in next syllable	I-1
	6	Type of vowel in current syllable	V-0
	6	Type of vowel in next syllable	V-1
	6	Type of 2nd consonant in current syllable	SC
	6	Type of 3rd consonant in current syllable	TC
	4	Syllable's position (first, middle, last, monosyllable)	POS-SYLAB
Out puts(Syllable)	4	Log-Pitch freq. curve-Legendre coefficients	PITCH
	4	Log-energy curve-Legendre coefficients	ENERGY
	1	Pause duration before current syllable	PAUSE
	1	Length of syllable	LEN-SYL
	1	Length of vowel	LEN-VOWEL
	1	Place of vowel onset in current syllable	START-VOWEL

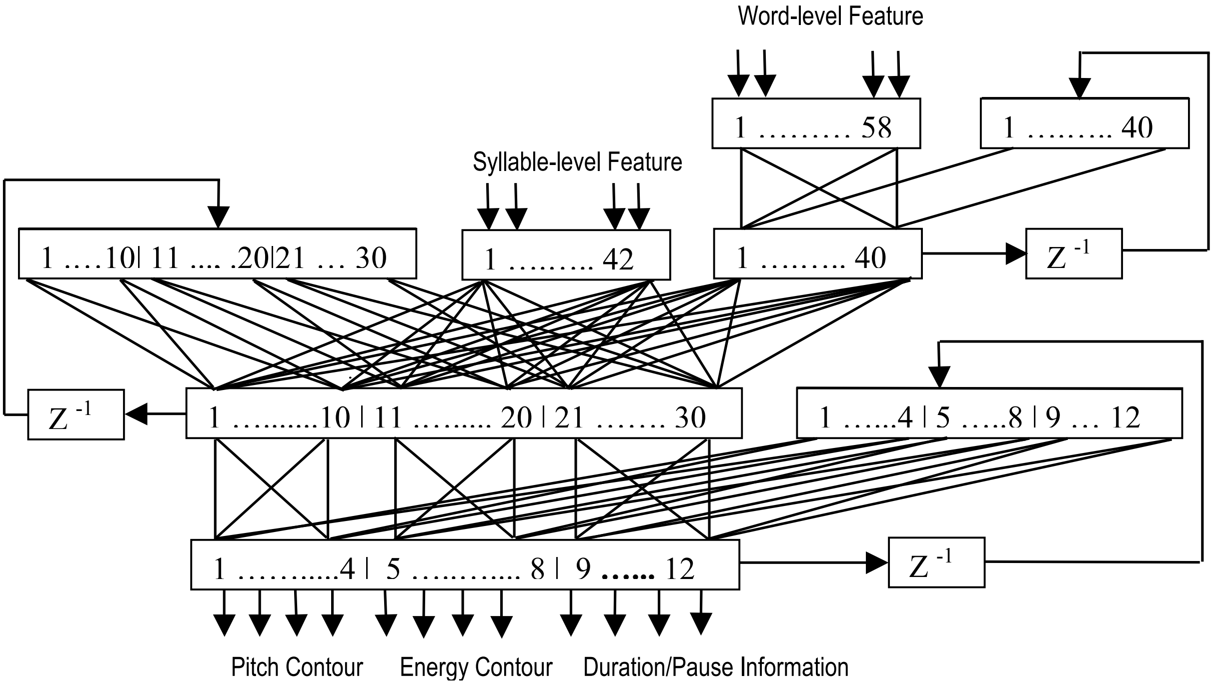

Figure 1**.** *Structure of the neural network generating prosodic data.*

In the past few years, many studies have been published on deriving prosodic model of spoken language for TTS [9-14]. Ostendorf and Veilleux [12] used a hierarchical stochastic mode to automatically predict prosodic phrasal boundaries in text, achieving promising results in determining where major and minor prosodic breaks occur in input text. Sanders and Taylor [13] identified phrasal breaks in text using a statistical model that described the relationship between phrase break and part of speech (POS) trigrams. Although these two methods are potentially suitable for use in TTS synthesis, further studies on assigning proper prosodic parameter patterns to the detected prosodic phrases are still needed [1].

As shown in Table 2, the POS set consists of verb, noun, adverb, adjective, infinitive, number, pronoun, indefinite, interjection, preposition, and conjunction.

Table 2***:*** *POS Tags*

No.	POS (W)	No.	POS (W)
1	Verb	7	Indefinite
2	Adjective	8	Interjection
3	Noun	9	Preposition
4	Infinitive	10	Conjunction
5	Adverb	11	Sign of direct object
6	Number	12	Pronoun

Table 3***:*** *Classification of consonants in RNN*

C1	p, t, k	C4	s, sh, ch, f
C2	b, d, g	C5	z, j, zh, ch, v
C3	m, n, l, r, y	C6	h, x, e'

Different categories of sentence type (such as predicative, interrogative, imperative and exclamatory) are shown with TYPE_SEN in Table.1. The input word-layer and the first hidden layer operate with a word-synchronized clock to represent current word phonological states within the prosodic structure of text to be synthesized. The second hidden layer uses syllable-level inputs, along with outputs from the preceding layer, to generate desired prosodic parameters. Denoting vowels by V and consonants by C, Persian syllables can be found in one of the forms of V, VC, CV, CVC, and CVCC [15]. To reduce the number of input nodes of the second layer, containing syllable data, Persian consonants are set into six groups, as shown in table3. Each of six Persian vowels, including "a" (as in "ran"), "e" (as in "bed"), "o" (as in "more"), "ee" (as in "sheep"), "A" (as in "car"), and "u" (as in "dub"), are treated separately.

The syllables are grouped based on their location in the input word: primary, middle, last, or monosyllabic (POS-SYLAB in table 1). Moreover, four Legendre coefficients are employed to represent the curve of log F0 (for voiced) and the curve of log energy of each syllable to decrease the number of output nodes of the network [1].

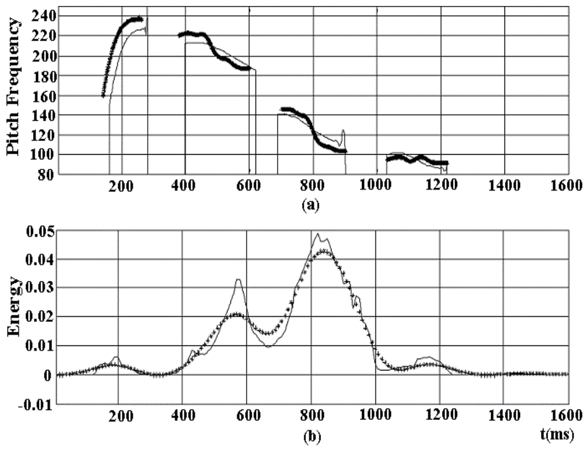

Figure 2. *Performance of prosody generator (solid: original, +: predicted), a) pitch contour, b) energy contour.*

3. Experiment

We conducted a simulation of the proposed system using MATLAB, where the training function was GDX [16], activation function of output nodes was linear, and tang-sigmoid function was selected for the activation function of the remaining nodes. All output quantities were evaluated linearly between zero and one and the error function was RMSE (root mean square error). The data set for training the network was obtained from the segmentation of about 1000 Persian sentences, uttered by a native Persian male speaker, containing about 10000 syllables at an average speed of 4 syllable/sec. The sentences were selected from a set of everyday conversational speech of different types: declarative, interrogative, exclamatory, and imperative, with affirmative and negative modes, sampled at 10 KHz with 16-bit resolution.

To evaluate performance of the system, 50 sentences (780 syllables), outside the training set, were used. The resulting RMSE of the trained Network for the test set were as follows (see table 1): LEN-SYL 40.8 ms, LEN-VOWEL 36 ms, PAUSE 31 ms, F0 18.7 Hz, and ENERGY 2.1 dB, which showed slightly higher performance, as compared to the prosody generators reported in [1] and [6]. One of the major reasons for such superiority is the use of syllable's energy contour in the proposed network, rather than ignoring the syllable energy in [6] or taking a simple scalar as the syllable energy used in [1]. Besides, in comparison with the network introduced in [1], the purposed network has lesser feedback connections that reduce the system complexity without sacrificing the performance. Figure 2 shows synthesized pitch and energy contours for a test Persian sentence "/kojA Amadee?/" (meaning "where did you come?" in English).

4. Conclusion

A new neural network-based prosodic information generator for Persian TTS systems has been introduced in this paper. It employs a compact four-layer RNN that simultaneously generates prosodic information including pitch and energy contours, syllable and vowel lengths, place of commencement of vowels, and inter-syllabic pause duration. Experimental results show that the proposed system outperforms some prosody generation systems of even more complexity, in terms of error in prediction of prosodic parameters.

5. References

[1] S. Chen, 1998 S. Hwang, and Y. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-_speech'' *IEE Transactions on speech and audio processing,* vol.6, No.3, 226-238, May.

[2] C. Tbler, 1992,"F0 generation with a database of natural F0 patterns and with a neural network,'' *in Talking Machines: Theories Models and Applications. Amsterdam, The Netherlands: Elsevier*.

[3] M. S. Scordilis and J. N. Gowdy, 1989, "Neural network based generation of fundamental frequency contours,'' *in Proc. ICASSP*, 219-222.

[4] M. Riedi, 1995, " A neural-network-based model of segmental duration for speech synthesis,'' in *Proc. EUROSPEECH*, 599-602.

[5] Y. Hifny and M. ashwan, 2002 "Duration modeling for Arabic text to speech synthesis,'' *7th International conference on spoken language processing ICSLP*, 1773–1776.

[6] O. Jokisch, H. Ding, H. Kruschke, and G. Streacha, 2002, "Learning syllable duration and intonation of Mandarin Chinese,''*7th International conference on spoken language processing ICSLP*, 1777-1780.

[7] Y. Sagisaka, 1990, "On the prediction of global F0 shape for Japanese text-to-speech.'' i*n Proc. ICASSP, 325-328.*

[8] Farrokhi, S. Ghaemmaghami, M, Tebyani, and M. Sheikhan, 2002, "Automatic segmentation of speech signal to extract prosodic information," *Proc. 10th Iranian Conf. Electr. Eng. (Computer Sessions),* 424-431, Tabriz, Iran.

[9] H. Mixdorff and H. Fujiski, 1995, " A scheme for a model-based synthesis by rule of F0 contours of German utterance,'' in *Proc. EUROSPEECH*, 1823-1826.

[10] E. Lopez-Gonzalo and L. A. Hernandez-Gomez, 1995, "Automatic data-driven prosodic modeling for text to speech,'' in *Proc.EUROSPEECH*, 585-588.

[11] Y. Yamashita and R. Mizoguchi, 1995, "Modeling the contextual effect on prosody in dialog,,'' in *Proc. EUROSPEECH*, 1329-1332 .

[12] M. Ostendorf and N. Veilleux, 1994, "A hierarchical stochastic model for automatic prediction of prosodic boundary location, '' *Comutat. Linguist.,*vol. 20, 27-54.

[13] E. Sanders and P. Taylor, 1995, "Using statistical model to predict phrase boundaries for speech synthesis, '' in *Proc. EUROSPEECH*, 1811-1814.

[14] H. Fujisaki and S. Ohno, 1993, " Prosodic modeling in Swedish speech synthesis,'' *Speech Commun.,*, vol. 13, 63-73.

[15] Y. Samareh, 1995, "Phonetics of Persian Language,'' *University Press Center*, University of Tehran, Iran.

[16] D. Hahn, "Essential, 2002, MATLAB for scientists and engineers,'' *Oxford*, *Butterworth*.