A NEURAL LEARNING APPROACH FOR DURATION PARAMETER GENERATION IN MANDARIN SPEECH SYNTHESIS

HUANG Yan and HUANG Tai-yi
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, P. O. Box 2728, 100080, Beijing, P. R. China
Tel: 86-10-62566666-1954, Fax: 86-10-62553991
E_mail: hy@prldec3.ia.ac.cn huang@prldec3.ia.ac.cn

ABSTRACT

In this paper, a neural learning approach is investigated, which is designed to generate duration parameter for mandarin speech synthesis. Unlike traditionally used rule-based methods, the novelty of this method lies in that it combines neural learning strategy and prior linguistic knowledge to obtain duration parameter. Rules generalized by linguists are used to encode input vectors of the neural network, and five multi-layer neural networks are built to determine the duration parameter for each tonal syllable. Experiment results show that it is a flexible and effective way to determine duration parameter, and perhaps it provides a helpful way of thought to obtain other prosodic parameters for speech synthesis.

1. INTRODUCTION

The goal of speech synthesis is to enable a machine to transmit orally information to a user in a man machine communication context[1]. This requires the synthesized speech be natural or pleasant, which is always the most difficult subject in text-to-speech synthesis. In general, there are two key points in developing a high-quality speech synthesis system: one is how to find out the prosodic model, which can be used to control the prosodic feature; the other is how to build a flexible speech synthesizer, which permits to modify the prosodic features effectively. This paper focuses on automatic duration parameter generation, which is one factor in prosodic model.

Duration parameter of each syllable in continuous sentence is an important factor in the transmission of linguistic information.[2] In fact, a poor duration model will badly lower the naturalness and hence the quality of synthesized speech, which makes listeners feel quite uncomfortable and even feel unacceptable. Therefore, developing an effective approach for automatic duration parameter generation is essential.

Speech synthesis benefit a lot from linguists who generalized a series of prosodic rules [3]. Many Chinese TTS systems [4] are based on these rules, which are of different levels such as syllable, word, and sentence level. They are effective for duration parameter generation. But there are some drawbacks. Firstly, the duration parameter is affected by many factors, such as the attributes of the current syllable, of the prior syllable and of the posterior syllable, the position of the current syllable in word, the position of the current syllable in sentence etc. Perhaps it is not appropriate to use fixed values for syllables in various situations. Secondly, it is not easy to obtain all possible rules that relate to this problem, especially when multi-language synthesis system and personal characteristic speech synthesis system are taken into account. Custom rules are difficult to describe personal prosodic characteristics. A more flexible and robust way is needed for such tasks. Then statistical method is adopted [5], in which online linguistic database is used for prosody extraction and generation. When online database is replaced by offline training, ANN, another powerful tool, is introduced for prosodic generation. [6][7].

Here an approach for automatic duration generation based on neural learning strategies combined with rules is proposed. It shows to be a promising way to obtain duration parameter, especially for preferred personal characteristic speech synthesis system. It takes advantage both of the prior linguistic knowledge and of the flexibility of ANN. Experiments show that the result is satisfying.

In following chapters the neural learning approach for duration parameter generation for mandarin speech synthesis is introduced, including 2.1 database, 2.2 rule-based network input, 2.3 ANN structure, 2.4 network training. The experiment results and some discussions are showed in part 3. Part 4 is a brief conclusion. Future work is discussed in part 5.

2. A NEURAL LEARNING STRATEGY FOR DURATION PARAMETER GENERATION

2.1 Database

A phonetically oriented speech database for mandarin is used in the experiment. The database consists of 511 sentences, which are collected from newspaper. Most of them are from news articles. All the sentences are recorded by a male speaker at normal speaking rate with sampling rate of 16k.

The beginning and the ending point of each syllable in each sentence are labelled, thus the duration of each syllable in continuous speech is obtained. Table1 shows an example, a sentence extracted from the database and its label information.

Table1.An example sentence and its annotation

Shang4 Hai3 De0 Gong1 Ren2 Shi1 Fu0 Ke4 Fu2 Kun4 Nan0.

Annotation

Silence(14771) *Shang4*(20055) *Hai3*(22320) *Silence*(22967) *De0*(24584) *Silence*(25878) *Gong1*(29437) *Ren2*(30946) *Shi1*(34936) *Fu0*(38063) *Silence*(44102) *Ke4*(47013) *Fu2*(49709) *Silence*(50679) *Kun4*(54238) *Nan0*(57041).

From the labelled beginning and ending point of each syllable, Table2 is obtained:

Table2

Shang4	0.330s
Hai3	0.142s
De0	0.101s
Gong1	0.222s
Ren2	0.094s
Shi1	0.249s
Fu0	0.195s
Ke4	0.182s
Fu2	0.169s
Kun4	0.222s
Nan0	0.175s

2.2 Rule-based network input

In rule-based system, duration parameter is obtained according to rules on different levels such as sentence level, word level, and syllable level. As to syllable level two factors, the type of tone and the type of syllable, are considered. Table3 shows different coefficients for different vowels [4], Table4 shows different coefficients for different tones[4].

Table3 [4]

a	1.2
e	1.1
i	0.9
o	1.2
u	1.0
v	0.9
Compound Vowels	1.3
Nasal-ending Vowels	1.4

Table 4[4]

(Fives tones are further divided into 14 types of tones, which are tone0(1,2,3), tone1ˉ(1,2), tone2ˊ(1,2,3), tone3ˇ(1,2), tone4ˋ(1,2,3,4))

Tone0 (1,2,3)	0.7
Tone3ˇ(2)	1.2
Tone4ˋ(2, 3, 4)	0.9
Others	1.0

Because further division information of different tones is not acquirable from the database, the average value for each tone is used here. See Table5:

Table5

Tone0	Tone1	Tone2	Tone3	Tone4
0.7	1.0	1.0	1.1	0.925

The prior knowledge of syllable level generalized by linguistics is adopted here to encode vectors of the network. Different tones have different duration characteristics, so here five networks are built for five different tones(0, ˉ, ˊ, ˇ, ˋ). The five networks will each be trained separately.

It is considered that the duration parameter of the current syllable has close relation with its context in a sentence. So the information of the current syllable combined with context information is input into the network for training. The input of the network is a seven-dimension vector, which consists of the vowel type of the current syllable, of the prior syllable, and of the following syllable, the tone type of the prior syllable, the tone type of the following syllable, the position of the syllable in word, and the position of the syllable in sentence.

The vowel types and the tone types are encoded according to Table3 and Table5. As to the position of the syllable in word, Table6 [4] is referred.

Table6 [4]

	First	Second	Third	Forth
Mono-syllable word	1.00			
Bi-syllable word	0.90	0.95		
Tri-syllable word	0.85	0.80	0.90	
Quad-syllable word	0.85	0.75	0.80	0.90

Here the concept of word is different from the usual concept. It is decided by the duration of silence between the syllables, which can be obtained from the label information.

The position of syllable in sentence is encoded as 0.2+1.0/(2* index), in which index is the physical position of the syllable in sentence. From the equation it can be seen that as the sentence moving forward, the duration shows a trend of decreasing.

All the data are normalized to (0,1].

2.3 ANN structure

Five three-layer neural networks are used here. The number of input node is 7, as described before. The two hidden layers have 10 and 7 nodes respectively. The number of output node is 1. Thus five neural networks by 7*10*5*1 are built for training.

2.4 Training

350 sentences are randomly selected as the training set, the remaining 161 sentences are used as test set.

Five networks are trained respectively, in which back propagation algorithm is used. Output is the duration of each tonal syllable in continuous speech. Experiments show that the networks converge quickly.

3. RESULTS AND DISCUSSIONS

To test the method, 100 sentences are randomly selected from the test set.

Firstly, five networks are tested respectively to obtain the performance of each network. For tone 0 the average error is 10ms(maximum error 15ms, minimum error 2ms). For tone1 the average error is 13ms(maximum error 17ms, minimum error 3ms). For tone2 the average error is 12ms(maximum error 16ms, minimum error 5ms). For tone3 the average error is 15ms(maximum error 21ms, minimum error 5ms). For tone4 the average error is 13ms(maximum error 18ms, minimum error 5ms).

The ultimate goal of this approach is applying it to sentences synthesis. In order to test the overall performance of the method, 10 sentences are chosen from the test set as test sentence. Table7 is the testing result of the following sentence:

You3 san1 bai3 wan4 ou1 gong4 ti3 guo2 jia1 de0 gong1 ren2 yi1 kao4 jun1 gong1 sheng1 chan3 sheng1 huo2

Table7

Syllable	Error	Syllable	Error
You3	*10ms*	*Gong1*	*9ms*
San1	*3ms*	*Ren2*	*6ms*
Bai3	*5ms*	*Yi1*	*6ms*
Wan4	*6ms*	*Kao4*	*8ms*
Ou1	*8ms*	*Jun1*	*5ms*
Gong4	*13ms*	*Gong1*	*5ms*
Ti3	*13ms*	*Sheng1*	*5ms*
Guo2	*12ms*	*Chan3*	*6ms*
Jia1	*6ms*	*Sheng1*	*3ms*
De0	*4ms*	*Huo2*	*3ms*

To evaluate it subjectively, this result is added to the mandarin synthesis system. Informal listening test shows it is satisfying.

The different performance of different networks for five tones are also analyzed. The network for tone3 has the highest average error. Tone3 has the prominent coarticulation phenomenon. It should be turned into tone2 when followed by tone3. When the different syllables from the database are classified according to five tones, such phenomenon was not taken into account. This may account for the higher average error of tone3. To verify this, another experiment is done. When the classification is made, the tone3 is classified into the set of tone2 if the following syllable is also tone3. With such changing being considered, the network for tone3 is trained again. Test result shows that the analysis is right. The error for tone3 falls to 10ms. Therefore, it is necessary to consider several coarticulation phenomena if the training database is not very large.

The sentences of the training set are increased to 500 sentences. The whole process is done again without considering any coarticulation phenomena, the results of the five networks are all improved with varied degrees. For tone 0 the average error is 8ms(maximum error 10ms, minimum error 2ms). For tone1 the average error is 10ms(maximum error 11ms, minimum error 3ms). For tone2 the average error is 9ms(maximum error 10ms, minimum error 5ms). For tone3 the average error is 10ms(maximum error 9ms, minimum error 5ms). For tone4 the average error is 10ms(maximum error 10ms, minimum error 5ms).

Tone3 has the biggest improvement with average error from 15ms to 10ms. The neural network learned the coarticulation phenomenon when a larger database is attainable.

4. CONCLUSION

The ANN-and-rule based method is an effective way to determine duration parameter for speech synthesis. It takes advantage of rule knowledge, and at the same time benefits from the flexibility of ANN.

The prior linguistic knowledge is used for network input encoding, which is a basis of this method. It is taken as the general properties of all speakers and all styles. Then the flexibility of ANN is utilized to catch the different characteristics of different people and different styles. So the training largely depends on the database used. This also makes the synthesis of different styles of speech possible. Such idea of two-step processing of prosody may be a way to solve other similar problems in prosody generation for synthesis.

The scale of database is a factor for the training. When it is not big enough, it is necessary to consider some detailed coarticulation phenomena. When it is big enough, such coarticulation phenomena may be learned by the neural network itself.

It is held that for speech synthesis of different style, ANN-and-rule based method will be a feasible and promising approach.

5. FUTURE WORK

Among prosodic elements, duration parameter is only one part. There are others, such as pitch, amplitude, etc.

Conventional way for pitch, a very important element in synthesis, is also based on rules. The idea of ANN combined with rules is also applicable here. Training a neural network for pitch contour is the next work, which may be more complex than for duration parameter generation. But they have very similar properties.

6. REFERENCE

[1]L.R.Rabiner, “Application of Voice Processing To Telecommunications,” *Proc.IEEE*, vol. 82,pp199-228, Feb 1994.

[2] Deborah L.B., Randy L.D. and Leslie B.C. “Effects of syllable duration on the perception of the Mandarin Tone2/Tone3 distinction: evidence of auditory enhancement ”, Journal of Phonetics(1990) 18, 37.

[3] Wu Zongji, Lin Maocan, “Outline On Experiment Phonetics (in Chinese)”, Higher Education Press, 1987.

[4] Chu Min. “Research in High legible and High Natural Chinese Text to Speech System ”Ph.D. Thesis, Institute of Acoustics, Chinese Academy of Sciences, 1993

[5] Chilin Shih and Benjamin Ao. “Duration Study for the Bell Laboratories Mandarin Text-to-Speech System.”, Progress in Speech Synthesis(Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors), New York: Springer, 1996

[6] Sin-Horng Chen, Shaw Hwa Hwang, and Chun-Yu Tsai, “A First Study of Neural Net Based Generation of Prosodic and Spectral Information for Mandarin Text-to-Speech ”, ICASSP92, Vol. 2, pp. 45-48.

[7] Sin-Horng Chen, Shaw Hwa Hwang and Yih-Ru Wang. “An RNN-based Prosody Information Synthesizer for Mandarin Text-to-Speech. IEEE transactions on Speech and Audio Processing”, May 1998, Vol.6, Number 03, p226