AN EFFICIENT TEXT ANALYZER WITH PROSODY GENERATOR-DRIVEN APPROACH FOR MANDARIN TEXT-TO-SPEECH

Shaw-Hwa Hwang and Cheng-Yu Yeh
Department of Electrical Engineering,
National Taipei University of Technology, Taipei, Taiwan, R.O.C.
hsf@ee.ntut.edu.tw, s1669009@ntut.edu.tw

ABSTRACT

A new approach for efficient text analyzer is proposed. The prosody generator driven method is employed to design an efficient text analyzer for Mandarin text-to-speech. Three heuristic and theoretical methods are used to examine the capability of each linguistic feature. Firstly, the contribution of each linguistic feature on prosody generator is examined experimentally. Secondly, the cross-influence of each linguistic feature on the prosody generator is analyzed. Thirdly, the problem of over- and under- classification on the linguistic feature will be inspected.

Finally, these three analytic results are referenced to design an efficient text analyzer. More than 39,103 Chinese characters are employed to examine the performance of our text analyzer. Less than 78ms is need for word tagging under P4-1.4G PC. The correction rate with 97% is achieved. It confirms that the performance of our text analyzer is very good. Moreover, more natural and fluent speech is obtained under the lower computation.

1. INTRODUCTION

Text-to-Speech(TTS) which automatically converts the text into the running speech is an important technology for the application on multimedia and friendly UI. Many attractive applications such as email reader, e-book, news reader,.. etc, are designed based on the TTS technology. In general, the natural and fluent speech is the main goal for TTS.

A general TTS system includes text analysis(TA), prosody generator(PG), synthesis unit generator(SUG), and speech synthesizer(SS) [1]. The TA resolves the text syntactically or semantically and extracts some linguistic features. Usually, the work of TA needs the help from linguist. The PG receives linguistic feature and generates prosodic information. The prosodic information includes the pitch contour, energy contour, and duration. The naturalness of synthesized speech is determined by the prosodic information. The SUG generates the most suitable speech template for synthesized speech. The SS adopts prosodic information and synthesis unit. Then, the algorithm of prosodic modification is implemented on the synthesis unit and the natural speech is generated.

In the past, most linguists pay their effort on the architecture of TA. They do their best to find as more linguistic features as possible. Thus, some high-level linguistic features such as the boundary of phrase, prosodic phrase, sub-sentence, etc,.. are analyzed and extracted. In a general TTS system, good prosodic information will generate the good and natural speech. However, more linguistic features will not guarantee with good prosodic information. But it will need much effort and dramatic computation for the high-level linguistic feature. Moreover, some linguistic feature will interfere and degrade the performance of PG.

In other side, most experts of acoustics and computer science pay their effort on the good statistical model for PG and SUG. However, in order to have the best prosodic information, not only needs a good PG model, but also needs the most suitable linguistic feature. Thus, the suitable linguistic feature which is driven by the performance of PG is the best policy. In the other word, the best linguistic features which are used to generate the best prosodic information must be determined by the performance of the PG. It means that, linguistic and acoustic analysis need to be considered together for the best speech.

In this paper, an efficient TA by PG-driven approach is proposed. Three important topics are analyzed. Firstly, the contribution of each linguistic feature on the PG is examined. Secondly, the cross-influence of each linguistic feature will be analyzed. Lastly, the problem of over-classification on linguistic feature will be examined. Finally, an efficient TA is implemented according to these analysis results. The efficient TA with low computation and high performance on PG will be achieved.

In the following sections, the RNN-based prosody generator will be described in Section 2. Three topics of the analytic methods will be described and defined clearly in Section 3. Section 4 will shows the experimental results and discussions. The conclusion will be given in the last section.

2. THE RNN-BASED PROSODY GENERATOR

The RNN-based prosody generator will be used in this paper. Fig.1 depicts the block diagram of the four-layer RNN [2]. The input linguistic features include the tone (Tone), the consonant initial (Ini), the vowel final (Fin), the part-of-speech (POS), the word's length (Len), the punctuation mark (PM), and the indicator (L) which shows the first, the middle, or the last character in a word. The eight outputs of prosodic parameters include four parameters of pitch contour, energy, pause duration, initial duration, and final duration.

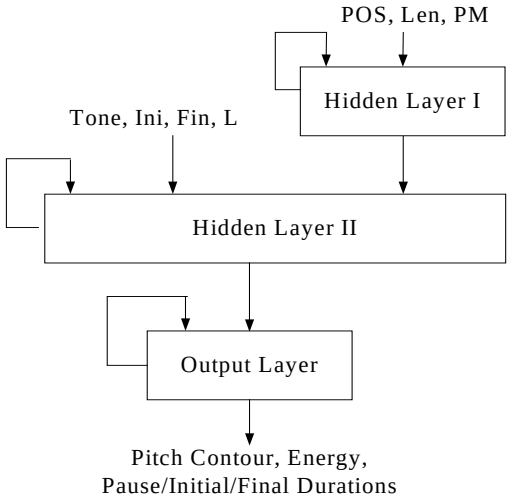

Fig.1. The block diagram of RNN-based prosody generator.

3. SYSTEM DESCRIPTION

In this section, three topics will be discussed in detail.

3.1 The Contribution of Linguistic Feature

The contribution of each linguistic feature(LF) on the PG is inspected by the performance of the RNN-based PG. The score function for each LF's contribution is defined as ***S(LF)*** which is equal to the value of root-mean-square-error(RMSE) between the real and synthesized prosodic information.

$$S(LF_i) = \left[\frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{J} (RP_{n,j} - SP_{n,j})^2 \right]^{1/2}, \quad (1)$$

the $RP_{n,j}$ is the ***j***-th real prosodic information at the ***n***-th syllable and ***N=35,000, J=8***. More than 35,000 syllables are employed to estimate the value of ***S(LF)***. For a fixed linguistic feature LF_i, the more less $S(LF_i)$ is obtained, the more contribution on PG will be presented. The value of $S(LF_i)$ can help us to realize the capability of each linguistic feature. Furthermore, an efficient text analyzer can be implemented according to these results.

For more precise definition, the RMSE of each LF versus the pitch contour can be defined as below:

$$R(X = x_k, Pitch) = \left\{ \frac{1}{N_I} \sum_{i=1}^{N_I} \sum_{j=0}^{3} \{T[p_j(i)] - O[p_j(i)]\}^2 \right\}^{1/2}, \quad (2)$$

where $X = x_k = \{Tone, Ini, Fin, L, Len, POS, PM\}, 1 \le k \le 7$ represents the different LF. The N_I is the number of training data. The $T[p_j(i)]$ and $O[p_j(i)]$ are the target and output values of the i-th coefficient of pitch contour respectively. The RMSE of pause duration, initial duration, final duration, and energy for each LF can also be obtained from the following equations respectively.

$$R(X = x_k, Pause) = \left\{ \frac{1}{N_I} \sum_{i=1}^{N_I} \{T[d_1(i)] - O[d_1(i)]\}^2 \right\}^{1/2}, \quad (3)$$

$$R(X = x_k, Initial) = \left\{ \frac{1}{N_I} \sum_{i=1}^{N_I} \{T[d_2(i)] - O[d_2(i)]\}^2 \right\}^{1/2}, \quad (4)$$

$$R(X = x_k, Final) = \left\{ \frac{1}{N_I} \sum_{i=1}^{N_I} \{T[d_3(i)] - O[d_3(i)]\}^2 \right\}^{1/2}, \quad (5)$$

$$R(X = x_k, Energy) = \left\{ \frac{1}{N_I} \sum_{i=1}^{N_I} \{T[e(i)] - O[e(i)]\}^2 \right\}^{1/2}. \quad (6)$$

Finally, these RMSE values can used to realize the capability of each linguistic feature.

Moreover, the conditional entropy of linguistic feature and prosodic information can help us to predict the capability of each LF before the training process. In this paper, the normalized conditional entropy of pitch contour with regard to each LF is discussed. It can be calculated by three steps. Firstly, the vector quantization (VQ) algorithm is used to classify the pitch contour pattern of training data into 64 clusters. Secondly, the conditional entropy of pitch contour with regard to each LF can be estimated from the following equation.

$$H(Pitch \mid X) = \sum_{i=1}^{N(X)} p(C_X(i)) \cdot H(Pitch \mid C_X(i)), \quad (7)$$

the $N(X) = \{5, 22, 39, 4, 5, 43, 12\}$ is the class numbers of each LF. The $p(C_X(i))$ is the probability of the i-th class on LF, $C_X(i)$. The $H(Pitch \mid C_X(i))$ is the conditional entropy of pitch contour with regard to i-th class of LF. It can be obtained by

$$H(Pitch \mid C_X(i)) = -\sum_{j=1}^{N_V} p(C_V(j) \mid C_X(i)) \cdot \log_2[p(C_V(j) \mid C_X(i))], \quad (8)$$

where $p(C_V(j) \mid C_X(i))$ is the conditional probability of j-th cluster $C_V(j)$ in VQ algorithm under the i-th class of LF. The N_V is the number of cluster and is defined as 64. Lastly, the normalized conditional entropy with their maximum entropy can be defined as

$$H_{nor}(Pitch \mid X) = \frac{H(Pitch \mid X)}{H_{\max}(Pitch \mid X)}, \quad (9)$$

where the $H_{\max}(Pitch \mid X)$ is the maximum entropy and is defined as

$$H_{\max}(Pitch \mid X) = \sum_{i=1}^{N(X)} p(C_X(i)) \cdot \log_2[NC_X(i)]. \quad (10)$$

The $NC_X(i)$ represents the pattern numbers in i-th class and is expected to be an uniform distribution.

3.2 The Cross-Influence of Each Linguistic Feature

The cross-influence is regarded as the relation between each two or more linguistic features. The analysis result of cross-influence can help us to select the optimal combination of linguistic feature and design an efficient TA. There are four relations of cross-influence which is defined as below.

a. *Cooperation*:

$$\Delta R(AB, Y) > \Delta R(A, Y) + \Delta R(B, Y), \quad (11)$$

where $\Delta R(X, Y) = R(Null, Y) - R(X, Y)$, is the differential RMSE between the 'Null' case($R(Null, Y)$) and the 'X' case. The 'X' is one of LF. The 'Y' is one of the prosodic parameter which is defined as $Y = \{Pitch, Pause, Inifin, Final, Energy\}$. Then, $\Delta R(AB, Y)$ represents the differential RMSE with two LFs('A' and 'B') simultaneously.

b. *Independence*:

$$\Delta R(AB, Y) = \Delta R(A, Y) + \Delta R(B, Y), \quad (12)$$

c. *Overlapped*:

$$Max[\Delta R(A, Y), \Delta R(B, Y)] < \Delta R(AB, Y) < \Delta R(A, Y) + \Delta R(B, Y), \quad (13)$$

d. *Interference*:

$$\Delta R(AB, Y) < Max[\Delta R(A, Y), \Delta R(B, Y)]. \quad (14)$$

3.3 The Classification of Linguistic Feature

The suitable classification on each LF will make the best performance on the prosody generator. In the other side, the redundancy of computation on TA and degradation of natural on PG will be obtained. There are two problems of classification. One is the over-classification. The other is the under-classification. The problems of classification can be inspected by the value of normalized differential RMSE which is normalized by their entropy with respect to each LF. The normalized differential RMSE is defined as

$$NR(X, Y) = \frac{\Delta R(X, Y)}{H(X)}, \quad (15)$$

where $H(X) = \sum_{i=1}^{N(X)} p(C_X(i)) \cdot \log_2[p(C_X(i))]$, represents the entropy of each LF. Two conditions of classification will be discussed in the following item.

a. *Over-classification:*

$$H(X_{c1}) < H(X_{c2}), \; NR(X_{C1}, Y) > NR(X_{C2}, Y). \quad (16)$$

The C_1 is small than C_2. The $H(X_{c1})$ and $H(X_{c2})$ represent the entropy of different classification with the LF 'X' respectively.

b. *Under-classification:*

$$H(X_{c1}) < H(X_{c2}), \; NR(X_{C1}, Y) < NR(X_{C2}, Y). \quad (17)$$

4. EXPERIMENTAL RESULT

In this paper, more than 35,000 syllables waveform and its relative Chinese characters are employed to train and examine our approach. A complete TA is employed to extract the LF as many as possible firstly. The RNN-based PG is employed to examine the contribution of each LF secondly. Seven types of LF and five types of prosodic information are analyzed, respectively and simultaneously. Three important topics are analyzed and discussed via the experimental result.

In the first topic, the RMSE of RNN-based pitch generator for each LF is estimated and depicted in Fig.2. The 'Tone' will have the greatest contribution than the others. It means that the 'Tone' is the best LF for pitch generator. Moreover, the 'Fin' and 'L' will almost have no any contribution on pitch generator. The 'Null' means that no any LF is employed to generate the pitch. Its RMSE is equal to the standard deviation of pitch. Table 1 lists the normalized conditional entropy which is estimated by using Eq. (7)-(10). The large value of the entropy means that the pitch is almost uniform distributed for each type of LF. It will decrease the performance of PG. Table 1 is estimated by theoretical analysis and Fig.2 is obtained by experimental result. They have the same result and conclusion.

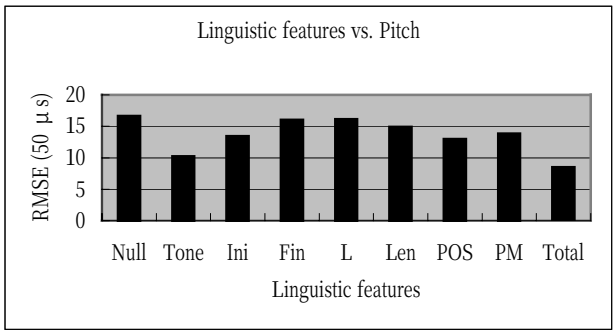

Fig. 2. The RMSE of each LF vs. Pitch

Table 1. The normalized conditional entropy of pitch with regard to each LF.

	Tone	Ini	Fin	L	Len	POS
H(Pitch | X)	0.390542	0.525332	0.556035	0.433985	0.430717	0.523520

In the second topic, Table 2 shows the differential RMSE of each prosody with 'Tone', 'Ini', 'Fin', 'IF', and 'TIF', respectively. The 'IF' means that the initial and final types of syllable are taken as LF, simultaneously. The 'TIF' means that the tone, the initial, and the final types of syllable are taken as LF, simultaneously. Table 3

shows the type of cross-influence of each LF. The 'Cooperation' case is the best solution and the 'Interference' case must be avoided. Table 3 points out the best direction for the implementation of TA system.

Table 2. The differential RMSE of each LF

	Pitch (50 μs)	Pause Duration (10ms)	Initial Duration (10ms)	Final Duration (10ms)	Energy (dB)
Tone	6.379313	0.096487	0.168615	0.809698	0.297440
Ini	3.168502	0.171919	2.678541	0.967988	0.625193
Fin	0.590008	0.019043	0.371558	0.501969	0.414191
IF	4.183168	0.148693	2.680984	1.432132	0.964136
TIF	6.771739	0.192073	2.744602	1.613471	1.383982

Table 3. The type of cross-influence for each LF

	IF	TIF
Pitch	Cooperation	Overlapped
Pause Duration	Interference	Overlapped
Initial Duration	Overlapped	Overlapped
Final Duration	Overlapped	Overlapped
Energy	Overlapped	Cooperation

In the third topic, Table 4 lists the situation of classification on the final type of syllable and the POS type of word. The 'Fin39' and 'Fin17' represent the different classification of 39 and 17 classes on the final type respectively. The 'POS43' and 'POS13' represent the 43 and 13 classes on the POS type. In Table 4, the final type with 'Fin39' is over-classified for the initial duration, final duration, and energy generators. But, the 'Fin17' is under-classified for the pitch and pause duration generators. Moreover, the POS type with 'POS43' is over-classified for pitch, initial duration, and final duration generators. According to the results on Table 4, the suitable classification on each LF for prosody generator can be achieved. Furthermore, Fig. 3 shows the RMSE with the original classification on LF(Total) and the simple classification on final, POS, and PM types(Total'). The performance with the Total' approach has no obvious degradation. But the computation of the Total' approach is reduced dramatically in the TA.

Table 4. The situation of classification on the final type of syllable and the POS type of word

$NR(X,Y)$		Pitch	Pause Duration	Initial Duration	Final Duration	Energy
Final	Fin39	0.122852	0.003965	0.077366 **(Over)**	0.104520 **(Over)**	0.086243 **(Over)**
	Fin17	0.102503 **(Under)**	0.003741 **(Under)**	0.082753	0.118892	0.088066
POS	POS43	0.699324 **(Over)**	0.198106	0.031626 **(Over)**	0.120975 **(Over)**	0.301990
	POS13	0.969081	0.184069 **(Under)**	0.045709	0.184358	0.160280 **(Under)**

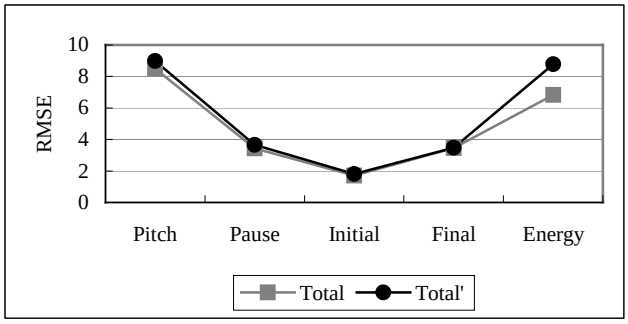

Fig. 3. The RMSE with original classification on LF vs. simple classification on LF.

According to above analysis results, an efficient TA with the best performance can be easily achieved. Totally 39,103 Chinese characters are employed to test the performance of our TA. Only 78ms CPU time under the PC(Pentium-IV, 1.4GHz) is achieved. Moreover, the correction rate of word tagging with 97% is achieved. It confirms that the performance of our text analyzer is very good.

5. CONCLUSION

A new approach for the implementation of an efficient TA for Mandarin TTS is proposed. Three heuristic and theoretical analysis methods are employed to examine the capability of each LF. The problem of contribution, cross-influence, and over-classification of each LF can be easily inspected. Finally, an efficient TA can be easily achieved.

REFERENCES

[1] S. H. Cheng, S. H. Hwang, and Y. R. Wang, "A Mandarin Text-to-Speech System," Proc. ICSLP, Vol. 3, pp. 1421-1424, 1996.

[2] S. H. Cheng, S. H. Hwang, and Y. R. Wang, "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, 1998.

[3] F.C. Chou, C.Y. Tseng, K.J. Chen, and L.S. Lee, "A Chinese text-to-speech system based on part-of-speech analysis, prosodic modeling and non-uniform units," Proc. ICASSP, Vol. 2, pp. 923-926, 1997.

[4] Zhiwei Ying, and Xiaohua Shi, "An RNN-based algorithm to detect prosodic phrase for Chinese TTS," Proc. ICASSP, Vol. 2, pp. 809-812, 2001.

[5] F.C. Chou, C.Y. Tseng, and L.S. Lee, "Automatic Generation of Prosodic Structure for High Quality Mandarin Speech Synthesis," Proc. ICSLP, Vol. 3, pp. 1624-1627, 1996.