Efficient text analyser with prosody generator-driven approach for Mandarin text-to-speech

C.-Y. Yeh and S.-H. Hwang

Abstract: A new approach for an efficient text analyser is proposed. The prosody generator-driven method is employed to design an efficient text analyser for Mandarin text-to-speech. More simple structure of text analysis, more suitable classification of linguistic features and more efficient contribution of linguistic features to the prosody generator can be achieved. Three heuristic and theoretical methods are used to analyse and examine the capability of each linguistic feature. First, the contribution of each linguistic feature to the prosody generator is examined experimentally. Secondly, the cross-influence of each linguistic feature on the prosody generator is analysed. Thirdly, the problem of over- and under- classification of the linguistic features is inspected. Finally, these three analytic results are referenced to design an efficient text analyser. In total 35 243 Chinese characters are employed to examine the performance of our text analyser. Only 79 ms CPU time on a P4-1.4G PC is needed for word segmentation and POS tagging. Correction rates of 97.5 and 93.2% are achieved for the word segmentation and POS tagging, respectively. This confirms that the performance of our text analyser is very good. Moreover, a Mandarin text-to-speech system is implemented to inspect the performance of the text analysis and the contribution to the prosody generator. More natural and fluent speech is obtained under the lower computation. The MOS of prosody of the synthesised and original speech are 4.2 and 4.8, respectively, which is reasonably good.

1 Introduction

Text-to-speech (TTS), which automatically converts text into running speech, is an important technology for applications in multimedia and friendly UI. Many attractive applications, such as e-mail reader, e-book, news reader etc. are designed based on TTS technology. In general, TTS can generate speech without limit according to the input text. Natural and fluent speech is the most important issue for the development of TTS.

A general TTS [1–6] system includes text analysis (TA) [7–11, 35–37] prosody generator (PG) [12–23], synthesis unit generator (SUG) [24], and speech synthesiser (SS) [25, 26]. The TA resolves the text syntactically or semantically and extracts some linguistic features. Usually, the work of TA needs help from a linguist. The PG receives linguistic features and generates prosodic information. The prosodic information includes the intonation contour, energy envelope and duration pattern. The naturalness of synthesised speech is controlled by the prosodic information. The SUG generates the most suitable speech templates for synthesised speech. Co-articulation rules for each two adjacent templates are also employed to improve the fluency of synthesised 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.

IEE Proceedings online no. 20045095

doi: 10.1049/ip-vis:20045095

Paper received 5th July 2004. Originally published online 20th June 2005

The authors are with the Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan, Republic of China

E-mail: hsf@ntut.edu.tw

In the past, much effort was paid to design a TTS with high quality [1–6]. However, naturalness and fluency are two important issues for the TTS system. Thus, most researchers put their efforts into the prosody generator [12–23] for the TTS system. In the general prosody generator, two problems must be overcome to achieve natural and fluent speech. One is a suitable model for the prosody generator and the other is a suitable linguistic feature for the prosody generator.

In the first problem, in the past, the rule-based and the statistical-based approaches were employed to generalise suitable prosodic information. The rule-based approach [2, 12, 17] used many pronunciation rules inferred by the linguist to improve the speech quality for the TTS system. However, the cross-influence of the pronunciation rules on the prosodic information cannot be easily quantified and inferred as independent rules. Moreover, these pronunciation rules must be inferred from the acoustic expert and the linguist. The statistical approach [13, 15–19, 21–23] uses the probability model or the neural network to infer automatically the pronunciation rules. The natural prosodic information and pronunciation rules are automatically learned from the large database of natural speech. Moreover, the cross-influence of pronunciation rules on the prosodic information can be memorised and simulated in the joint probability of the probability model or the weight of the neural network.

In the second problem, in the past, most linguists put their efforts into the architecture of TA [7–11, 35–37]. They did their best to find as many linguistic features as possible. Thus, some high-level linguistic features, such as the boundary of phrase [34], prosodic phrase, sub-sentence etc. were analysed and extracted. In a general TTS system, good prosodic information will generate good and natural speech. However, more linguistic features will not guarantee good prosodic information. But it will need much effort

and dramatic computation for high-level linguistic features. Moreover, some linguistic features will interfere and degrade the performance of PG.

On the other hand, most experts on acoustics and computer science put their efforts into a good statistical model for PG and SUG. However, in order to have the best prosodic information, not only is a good PG model needed, but also the most suitable linguistic features are needed. Thus, suitable linguistic features driven by the performance of the PG is the best policy. In the other words, the best linguistic features used to generate the best prosodic information must be inspected and determined by the performance of the PG. It means that linguistic and acoustic analyses need to be considered together for the best speech.

In this paper, an efficient TA using a PG-driven approach [27] is proposed. An RNN-based prosody generator [16, 18, 21] is employed to analyse and inspect this approach. Three important topics are analysed. First, the contribution of each linguistic feature on the PG is examined. Secondly, the cross-influence of each linguistic feature is analysed. Lastly, the problem of over-classification of linguistic features is examined. Finally, an efficient TA is implemented according to these analysis results. A Mandarin TTS with RNN-based PG is implemented to examine the performance. An efficient TA with low computation and high performance on PG is achieved. The synthesised speech is more natural and fluent. The difficulty of high-level analysis on input text is avoided. The computation requirement for the text analysis is reduced dramatically.

2 Text-to-speech system

In the general text-to-speech (TTS) system, there are four basic subsystems. Figure 1 shows the basic block diagram of a general TTS system. It consists of the text analysis, the prosody generator, the synthesis unit generator and the speech synthesiser. The text analysis analyses text syntactically and/or semantically to extract some linguistic features. The prosody generator adopts linguistic features and generates prosodic information, such as pitch contour, energy contour and duration pattern. To make the synthesised speech more natural and intelligent is its main goal. The synthesis unit generator generates the synthesis unit according to the phonetic symbol. To make the synthesised speech more clear is its main goal. In general, the co-articulation effect elimination will also be employed to smooth the spectrum and energy between the two synthesis units and make synthesised speech sound more fluent. The speech synthesiser makes the prosodic modification on the synthesis unit and generates the synthesised speech. The ability of prosodic modification of the synthesis unit is important for the general speech synthesiser.

In this paper, efficient text analysis for the prosody generator-driven approach is considered. Thus, the basic structure of text analysis and prosody generator in our implementation will be described in the following Subsections.

2.1 Text analysis

In the structure of text analysis, a dictionary that contains more than 80 000 Chinese words and a score function with first-order Markov chain are employed to implement the word segmentation and POS (part of speech) tagging. The score function $S(w_i, t_i)$ is defined as below:

$$S(w_i, t_i) = L^2(w_1) + \alpha_1 F(w_1) + \alpha_2 \log(P(t_1)) + \sum_{i=2}^{N} [L^2(w_i) + \beta_1 F(w_i) + \beta_2 \log(P(t_{i-1}|t_i))] \quad (1)$$

The w_i and t_i is the ith word and POS, respectively. $L(w_i)$ is the number of the Chinese character in the word w_i. $P(t_{i-1}|t_i)$ is the conditional probability of POS t_{i-1} and t_i. $F(w_i)$ is the occurrence probability of word w_i. $\alpha_1, \alpha_2, \beta_1$ and β_2 are constants. The linguistic features include the tone type (Tone), the consonant initial (Ini), the vowel final (Fin), the part-of-speech (POS), the word length (Len), the punctuation mark (PM), and the indicator (L), which shows the first, middle or last character in a word used to analyse the contribution to the prosodic information.

2.2 Prosody generator

The RNN-based prosody generator is used in this paper. Figure 2 shows the block diagram of the three-layer RNN [21]. The input linguistic features include the tone (Tone), the consonant initial (Ini), the vowel final (Fin), the part-of-speech (POS), the word 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. The RNN operates with two clocks: one is the word sequence and the other is the Chinese character sequence. 'Hidden layer I' adopts the word-level linguistic

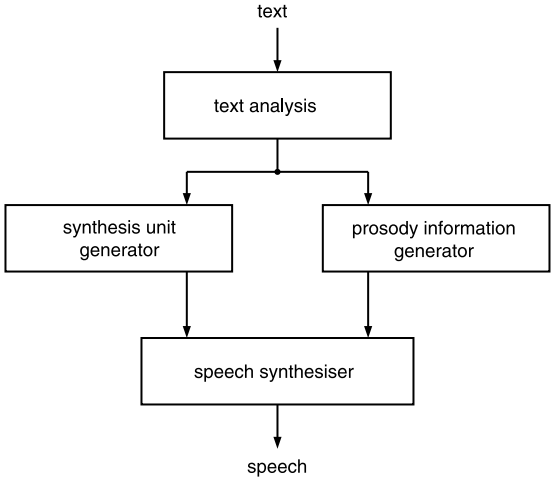

Fig. 1 *Block diagram of general TTS system*

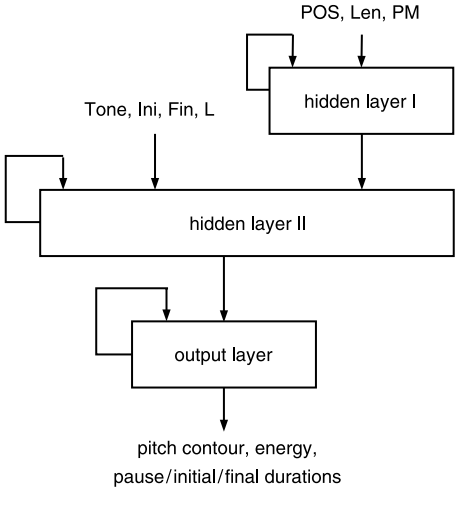

Fig. 2 *Block diagram of RNN-based prosody generator*

features, including 'POS', 'Len', and 'PM' and simulates the global declination effect of a whole sentence. 'Hidden layer II' adopts the syllable-level linguistic features, including 'Tone', 'Ini', 'Fin', and 'L' and generates the prosodic information associated with each syllable. The 'output layer' is a mechanism of linear combination. It can generate the value of prosodic information directly. The error back propagation (EBP) algorithm is used to adapt the weights of RNN.

3 System description

In this Section, three topics will be discussed in detail: the contribution of each linguistic feature on the prosody generator; the analysis of the cross-influence on the prosody generator between each linguistic feature; and the analysis of classification of each linguistic feature.

3.1 Contribution of linguistic features

The contribution of each linguistic feature (LF) to the PG is determined by the performance of the RNN-based PG. The score function for each LF's contribution to the prosodic parameter PP is defined as $R_{pp}(LF)$, which is equal to the value of the root-mean-square-error (RMSE) between the real and synthesised prosodic parameters.

$$R_{pp}(LF) = \left\{ \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{J} [P_j(n) - \hat{P}_j(n)]^2 \right\}^{0.5} \quad (2)$$

The $P_j(n)$ is the original prosodic parameter with the jth dimension at the nth syllable. The $\hat{P}_j(n)$ is the synthesised prosodic parameter of the jth dimension at the nth syllable. The values of N are 28 293 and 6 950 for the inside and outside tests, respectively. The value of J is the dimension of the prosodic parameter PP. LF, defined below, is the set of linguistic features:

$$LF \subseteq \{Tone, Ini, Fin, L, PM, Len, POS\}$$

PP is the set of prosodic parameters, which is defined as:

$$PP \subseteq \{Pitch, Pause\_Dur., Initial\_Dur., Final\_Dur., Energy\}$$

For a fixed linguistic feature LF, the R greater value of that is obtained, the more contributions on PG will be presented. The value of R can help us to realise the capability of each linguistic feature. Furthermore, an efficient text analyser can be implemented according to these results.

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

$$R_{Pitch}(LF) = \left\{ \frac{1}{N} \sum_{n=1}^{N} \sum_{j=0}^{3} [pth_j(n) - \hat{pth}_j(n)]^2 \right\}^{0.5} \quad (3)$$

$pth_j(n)$ is the real pitch coefficient at the jth dimension at the nth syllable. $\hat{pth}_j(n)$ is the synthesised pitch coefficient. These pitch coefficients are extracted from the pitch contour by the orthogonal expansion algorithm [3]. These relative equations are defined as:

$$pth_j(n) = \frac{1}{L} \sum_{l=0}^{L-1} \Phi_j\left(\frac{l}{L}\right) \cdot pitch(n, l) \quad (4)$$

$$\hat{pitch}(n, l) = \sum_{j=0}^{3} \Phi_j\left(\frac{l}{L}\right) \cdot pth_j(n) \quad (5)$$

$pitch(n, l)$ is the real fundamental frequency at the lth frame and the nth syllable. The $\hat{pitch}(n, l)$ is the synthesised fundamental frequency. $\Phi_j(l/L)$ is the orthonormal coefficient derived by the Gram-Schmit algorithm [3]. The value of L is the frame number of the nth syllable. The RMSE of pause duration, initial duration, final duration and energy for each LF can also be obtained from the following equations:

$$R_{Pause\_Dur.}(LF) = \left\{ \frac{1}{N} \sum_{n=1}^{N} [pause(n) - \hat{pause}(n)]^2 \right\}^{0.5} \quad (6)$$

$$R_{Initial\_Dur.}(LF) = \left\{ \frac{1}{N} \sum_{n=1}^{N} [initial(n) - \hat{initial}(n)]^2 \right\}^{0.5} \quad (7)$$

$$R_{Final\_Dur.}(LF) = \left\{ \frac{1}{N} \sum_{n=1}^{N} [final(n) - \hat{final}(n)]^2 \right\}^{0.5} \quad (8)$$

$$R_{Energy}(LF) = \left\{ \frac{1}{N} \sum_{n=1}^{N} [energy(n) - \hat{energy}(n)]^2 \right\}^{0.5} \quad (9)$$

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 normalised conditional entropy of pitch contour with regard to each LF is discussed. It can be calculated from three steps. First, the vector quantisation (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:

$$H(Pitch|X) = \sum_{i=1}^{NL(X)} p(C_X(i)) H(Pitch|C_X(i)) \quad (10)$$

$NL(X) \subseteq \{5, 22, 39, 4, 5, 43, 12\}$ are the class numbers of each LF. $p(C_X(i))$ is the probability of the ith class on LF,$C_X(i)$. $H(Pitch|C_X(i))$ is the conditional entropy of pitch contour with regard to the ith class of LF. It can be obtained from

$$H(Pitch|C_X(i)) = -\sum_{j=1}^{N_V} p(C_V(j)|C_X(i)) \times \log_2[p(C_V(j)|C_X(i))] \quad (11)$$

where $p(C_V(j)|C_X(i))$ is the conditional probability of the jth cluster $C_V(j)$ in the VQ algorithm under the ith class of LF. N_V is the number of the cluster and is defined as 64. Lastly, the normalised conditional entropy with their maximum entropy can be defined as

$$H_{nor}(Pitch|X) = \frac{H(Pitch|X)}{H_{\max}(Pitch|X)} \quad (12)$$

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

$$H_{\max}(Pitch|X) = \sum_{i=1}^{NL(X)} p(C_X(i)) \log_2[NC_X(i)] \quad (13)$$

$NC_X(i)$ represents the pattern numbers in the ith class and is expected to have a uniform distribution.

3.2 Cross-influence of each linguistic feature

The cross-influence is regarded as the relation between each set of two or more linguistic features and its contribution to the prosody generator. The analysis result of the cross-influence can help us to select the optimal combination of linguistic features and design an efficient TA. There are four relations of cross-influence, defined as:

a Co-operation:

$$\Delta R_{pp}(A+B) > \Delta R_{pp}(A) + \Delta R_{pp}(B) \tag{14}$$

where $\Delta R_{pp}(X) = R_{pp}(Null) - R_{pp}(X)$ is the differential RMSE between the 'null' case and the 'X' case. 'X' is one or a set of LF. Then, $\Delta R_{pp}(A+B)$ represents the differential RMSE with two LFs ('A' and 'B') simultaneously.

b Independence:

$$\Delta R_{pp}(A+B) = \Delta R_{pp}(A) + \Delta R_{pp}(B) \tag{15}$$

c Overlapped:

$$Max[\Delta R_{pp}(A), \Delta R_{pp}(B)] < \Delta R_{pp}(A+B) < \Delta R_{pp}(A) + \Delta R_{pp}(B) \tag{16}$$

d Interference:

$$\Delta R_{pp}(A+B) < Max[\Delta R_{pp}(A), \Delta R_{pp}(B)] \tag{17}$$

3.3 Classification of linguistic features

Suitable classification of each LF will give the best performance for the prosody generator. On the other hand, the redundancy of computation on TA and degradation of naturalness on PG will be obtained. There are two problems of classification. One is over-classification. The other is under-classification. The problems of classification can be seen from the value of normalised differential RMSE, which is normalised by its entropy with respect to each LF. The normalised differential RMSE is defined as:

$$NR_{pp}(X) = \frac{\Delta R_{pp}(X)}{H(X)} \tag{18}$$

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

a Over-classification:

$$H(X_{C1}) < H(X_{C2}),\ NR_{pp}(X_{C1}) > NR_{pp}(X_{C2}) \tag{19}$$

C_1 is smaller than C_2. $H(X_{C1})$ and $H(X_{C2})$ represent the entropy of different classifications with the LF 'X', respectively.

b Under-classification:

$$H(X_{C1}) < H(X_{C2}),\ NR_{pp}(X_{C1}) < NR_{pp}(X_{C2}) \tag{20}$$

4 Experimental results

In this paper, there are 35 243 syllabic waveforms and their relative Chinese characters, which are divided into inside set with 28 293 characters and outside set with 6950 characters, and are employed to train and examine our approach. A complete TA is first employed to extract the as many LFs as possible. Equation (1) is used as the cost function for the word segmentation and POS tagged. Next the RNN-based PG is employed to examine the contribution of each LF. Seven types of LF and five types of prosodic information are analysed, respectively and simultaneously. Three important topics are analysed and discussed via the experimental results.

In the first topic, the RMSE of the RNN-based pitch generator for each LF is estimated and given in Table 1. In the case 'Total', the values of RMSE with 8.486 and 10.798 are obtained for inside and outside tests, respectively. The RMSE of pitch contour with inside test is also shown in Fig. 3. The 'Tone' will have a greater contribution than the others. It means that the 'Tone' is the best LF for pitch generator. Moreover, the 'Fin' and 'L' will have almost no contribution to pitch generator. The 'Null' means that no LF is employed to generate the pitch. Its RMSE is equal to the standard deviation of pitch. Table 2 gives the normalised conditional entropy, which is estimated by using (10)–(13). The large value of the entropy means that the pitch is almost uniformly distributed for each type of LF. Table 2 is estimated by theoretical analysis and Fig. 3 is obtained by experimental results. The 'Tone' is the best LF for the pitch generator. Moreover, the 'Fin' has almost no contribution to the pitch generator. The major results of Fig. 3 and Table 2 seem consistent. However, the minor results of the other LFs have some little inconsistency between Fig. 3 and Table 2. If the sequence of training data of RNN is distributed all over, the RNN can reach the best performance. Meanwhile, Fig. 3 and Table 2 will be consistent. Otherwise, the result by theoretical analysis will have some differences from the experimental result. These results of Fig. 3 and Table 2 have some inconsistency. It points out that our training data needs more suitable arrangement.

For the second topic, Table 3 shows the value of differential RMSE of prosodic information for each linguistic feature set. 'Ini + Fin' means that the initial and final types of syllable are taken as LF, simultaneously. 'Tone + Ini + Fin' means that the tone, the initial and the final types of syllable are taken as LF, simultaneously. Table 4 shows the type of cross-influence on the prosodic information for some LFs. The 'co-operation' case is the best solution and the 'interference' case must be avoided. Table 4 points out the best direction for the implementation of the TA system. In Table 4, the pitch and final duration are located in 'overlapped' or 'co-operation' for each set of

Table 1: RMSE of pitch contour with each linguistic feature for inside and outside tests

RMSE of Pitch contour (50 μs)	Inside test	Outside test
Null	16.625362	17.168571
Tone	10.246050	11.127300
Ini	13.456861	14.742570
Fin	16.035355	17.005941
L	16.150660	16.880539
Len	14.898714	15.727992
POS	12.975919	14.729033
PM	13.866839	14.636217
Total	8.485972	10.798090

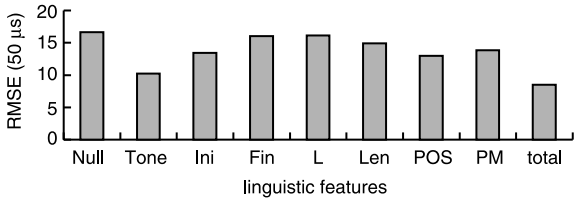

Fig. 3 *RMSE of synthesised pitch for each linguistic feature*

Table 2: Normalised conditional entropy of pitch for each linguistic feature

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

Table 3: Differential RMSE of prosodic information for each linguistic feature

	Pitch (50 μs)	Pause duration (10 ms)	Initial duration (10 ms)	Final duration (10 ms)	Energy (dB)
Tone	6.379312	0.096487	0.168615	0.809698	0.29744
Ini	3.168501	0.171919	2.678541	0.967988	0.625193
Fin	0.590007	0.019043	0.371558	0.501969	0.414191
L	0.474702	0.076356	0.027458	0.347682	0.128108
Len	1.726648	0.110429	0.074208	0.42742	0.359136
POS	3.649443	1.033824	0.165039	0.631312	1.575942
PM	2.758523	0.094571	0.065614	0.461359	1.523791
Tone + Ini	6.599831	0.208539	2.677076	1.238881	0.776281
Tone + Fin	6.587993	0.10044	0.646803	1.263239	0.971365
Tone + L	6.486928	0.223872	0.169743	0.97146	0.376178
Ini + Fin	4.183168	0.148693	2.680984	1.432132	0.964136
Ini + L	3.342702	0.32573	2.675802	1.096333	0.569848
Fin + L	1.343596	0.086047	0.446567	0.880998	0.476125
Tone + Ini + Fin	6.771739	0.192073	2.744602	1.613471	1.383982
POS + Len	3.784235	0.790399	0.163139	0.759569	1.264691
POS + PM	3.829629	0.940603	0.163008	0.651837	0.715341
Len + PM	3.263237	0.23402	0.111337	0.719843	1.0488
Tone + POS	7.313634	1.274107	0.225383	0.981465	2.538744
Tone + Len	6.568548	0.316116	0.150855	0.881406	0.931149
Tone + PM	6.838903	0.109319	0.16344	0.855223	0.968483
Ini + POS	4.234988	0.545472	2.668594	1.182875	1.776917
Ini + Len	3.561855	0.309969	2.628509	1.002126	0.762015
Ini + PM	3.55621	0.198324	2.662113	1.017818	2.48475
Fin + POS	4.057612	0.880022	0.524518	1.010926	1.891361
Fin + Len	2.52902	0.090341	0.450022	0.82832	0.67397
Fin + PM	3.25448	0.090583	0.458897	0.944236	1.077918
L + POS	4.028324	0.952697	0.216917	0.992081	1.515554
L + Len	1.78791	0.15412	0.086597	0.487867	0.383718
L + PM	3.249672	0.212944	0.093472	0.763406	1.644841

linguistic feature. It means that more linguistic features will improve the performance on pitch and final duration. The 'Tone' has great contribution on the pitch. However, the 'Tone + POS' is the best LF than 'Tone + Ini + Fin'. The word-level LF 'POS' has more contribution than the syllable-level LF such as 'Ini' and 'Fin'. From above results, maybe the 'POS' will have a larger contribution on the decrease effect of pitch contour over a whole sentence. Moreover, 'POS' and 'Ini' make large contributions to the pause and initial durations, respectively. The word-level LF, such as 'POS' and 'PM', will make a larger contribution to the energy than the other LFs.

For the third topic, Table 5 lists the value of differential RMSE and the situation of classification on the final type of syllable and the POS type of word. 'Fin39' and 'Fin17' represent the different classification of 39 and 17 classes on the final type, respectively. 'POS43' and 'POS13' represent the 43 and 13 classes on the POS type. In Table 5, 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 in Table 5, a suitable classification for each LF for the prosody generator can be achieved. Figure 4 shows the RMSE with the original classification of LF (total) and the simple classification of 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.

According to above analysis results, an efficient TA with the best performance can be easily achieved. In total 35 243 Chinese characters are employed to test the performance of our TA. Only 79 ms CPU time for the PC (Pentium-IV, 1.4 GHz) is achieved. Moreover, correction rates of 97.5 and 93.2% are achieved for the word segmentation and POS tagging. It confirms that the performance of our text analyser is very good. A Mandarin text-to-speech system has been implemented to inspect the performance of text analysis and the contribution to the prosody generator. More natural and fluent speech is obtained with less computation. The MOS of prosody of the synthesised and original speech are 4.2 and 4.8, respectively, which is reasonably good.

Table 4: Type of cross-influence on prosodic information for each linguistic feature

	Pitch	Pause duration	Initial duration	Final duration	Energy
Tone + Ini	overlapped	overlapped	interference	overlapped	overlapped
Tone + Fin	overlapped	overlapped	co-operation	overlapped	co-operation
Tone + L	overlapped	co-operation	overlapped	overlapped	overlapped
Ini + Fin	co-operation	interference	overlapped	overlapped	overlapped
Ini + L	overlapped	co-operation	interference	overlapped	interference
Fin + L	co-operation	overlapped	co-operation	co-operation	overlapped
Tone + Ini + Fin	overlapped	overlapped	overlapped	overlapped	co-operation
POS + Len	overlapped	interference	interference	overlapped	interference
POS + PM	overlapped	interference	interference	overlapped	interference
Len + PM	overlapped	co-operation	overlapped	overlapped	interference
Tone + POS	overlapped	co-operation	overlapped	overlapped	co-operation
Tone + Len	overlapped	co-operation	interference	overlapped	co-operation
Tone + PM	overlapped	overlapped	interference	overlapped	interference
Ini + POS	overlapped	interference	interference	overlapped	overlapped
Ini + Len	overlapped	co-operation	interference	overlapped	overlapped
Ini + PM	overlapped	overlapped	interference	overlapped	co-operation
Fin + POS	overlapped	interference	overlapped	overlapped	overlapped
Fin + Len	co-operation	interference	co-operation	overlapped	overlapped
Fin + PM	overlapped	interference	co-operation	overlapped	interference
L + POS	overlapped	interference	co-operation	co-operation	interference
L + Len	overlapped	overlapped	overlapped	overlapped	overlapped
L + PM	co-operation	co-operation	co-operation	overlapped	overlapped

Table 5: Normalised differential RMSE and situation of classification of 'final' and the 'POS' types for each prosodic information

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)

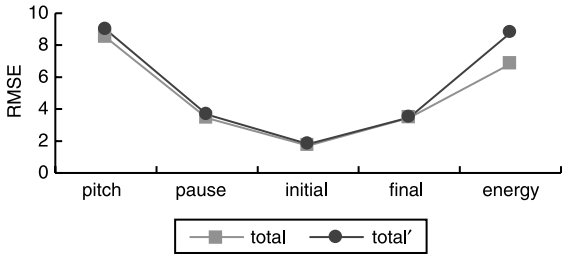

Fig. 4 *RMSE of each prosody for original and simple classification of linguistic feature*

5 Conclusions

A new approach to the efficient text analysis driven by the prosody generator for Mandarin TTS is proposed. The text analysis and the prosody generator are considered and designed together. Each linguistic feature is employed to analyse the contribution to the prosody generator. 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 for suitable linguistic features. Moreover, the best prosody generator can be achieved and more natural speech can be achieved for the TTS system.

In our approach, the performance of linguistic features for TTS is well examined by theoretical and experimental methods. Complex analysis by a linguist on high-level linguistic features is avoided. Moreover, the difficult analysis on the phrase [8, 28, 29], the preposition [30], the sub-sentence, the prosodic boundary [31], the breathable point or the prosodic phrase [32, 33] can also be avoided. The large requirement on computation is reduced. The cost and TTS system on chip (SOC) become possible.

From the analysis results, a suitable prosody generator is more important than linguistic features. The best speech comes from a suitable prosodic parameter. The major part of the prosodic parameter is a time-variant function. The time-causal model, such as the RNN model or the *n*-order Markov chain, will have good performance. The minor part of the prosodic parameter is influenced by some low-level linguistic features. Thus, more attention must be paid to the prosody generator. Analysis of high-level linguistic features can be avoided.

6 References

1 Klatt, D.H.: 'Review of text-to-speech conversion for English', *J. Acoust. Soc. Am.*, 1987, **82**, pp. 137–181
2 Lee, L.S., Tseng, C.Y., and Ming, O.Y.: 'The synthesis rules in a Chinese text-to-speech system', *IEEE Trans. Acoust. Speech Signal Process.*, 1989, **37**, (9), pp. 1309–1320

3 Cheng, S.H., Hwang, S.H., and Wang, Y.R.: 'A Mandarin text-to-speech system'. Proc. ICSLP, 1996, Vol. 3, pp. 1421–1424
4 Wang, R.H., Liu, Q.F., and Tang, D.F.: 'A new Chinese text-to-speech system with high naturalness'. Proc. ICSLP, 1996, Vol. 3, pp. 1441–1444
5 Lai, G.G., Min, H., and Qin, Z.S.: 'The research and implementation of Mongolian text-to-speech system'. 6th Int. Conf. on Signal Processing, August 2002, Vol. 1, pp. 472–475
6 Sef, T., and Gams, M.: 'Govorec (speaker) – Slovenian text-to-speech system for telecommunication applications'. 6th Int. Conf. on Signal Processing, August 2002, Vol. 1, pp. 504–507
7 Chang, L.L., *et al.*: 'Part of speech (POS) analysis on Chinese language'. Tech. Rep. Inst. Inform. Sci., Academia Sinica, Taiwan, Republic of China, 1989
8 Chen, K.J.: 'The identification of thematic roles in parsing Mandarin Chinese'. Proc. ROCLING II, 1989, pp. 121–146
9 Sef, T.: 'Text analysis for the Slovenian text-to-speech system'. Proc. ICECS, September 2001, Vol. 3, pp. 1355–1358
10 Sproat, R.: 'Multilingual text analysis for text-to-speech synthesis'. Proc. ICSLP, October 1996, Vol. 3, pp. 1365–1368
11 Tzoukermann, E., and Faubert, V.M.: 'Text analysis for French text-to-speech synthesis: optimization Bell Labs system'. Proc. IEEE Workshop on Speech Synthesis, 2002, pp. 195–198
12 Olive, J.P.: 'Fundamental frequency rules for the synthesis of simple declarative English sentences', *J. Acoust. Soc. Am.*, 1975, **57**, pp. 476–482
13 Scordilis, M.S., and Gowdy, J.N.: 'Neural network based generation of fundamental frequency contours'. Proc. ICASSP, 1989, Vol. 1, pp. 219–222
14 Sagisaka, Y.: 'On the prediction of global F0 shape for Japanese text-to-speech'. Proc. ICASSP, 1990, Vol. 1, pp. 325–328
15 Traber, C.: 'F0 generation with a database of natural F0 patterns and with a neural network', 'Talking machines: theories, models and applications' (Elsevier, Amsterdam, The Netherlands, 1992)
16 Hwang, S.H., and Chen, S.H.: 'Neural network synthesiser of pause duration for Mandarin text-to-speech', *Electron. Lett.*, 1992, **28**, pp. 720–721
17 Lee, L.S., Tseng, C.Y., and Hsieh, C.J.: 'Improved tone concatenation rules in a formant-based Chinese text-to-speech system', *IEEE Trans. Speech Audio Process.*, 1993, **1**, (3), pp. 287–294
18 Hwang, S.H., and Chen, S.H.: 'Neural network-based F0 text-to-speech synthesiser for Mandarin', *Proc. IEE, Vis. Image Signal Process.*, 1994, **141**, pp. 384–390
19 Riedi, M.: 'A neural-network-based model of segmental duration for speech synthesis'. Proc. EUROSPEECH, 1995, pp. 599–602
20 Chou, F.C., Tseng, C.Y., and Lee, L.S.: 'Automatic generation of prosodic structure for high quality Mandarin speech synthesis'. Proc. ICSLP, 1996, Vol. 3, pp. 1624–1627
21 Chen, S.H., Hwang, S.H., and Wang, Y.R.: 'An RNN-based prosodic information synthesizer for mandarin text-to-speech', *IEEE Trans. Speech Audio Process.*, 1998, **6**, (3), pp. 226–239
22 Ying, Z., and Shi, X.: 'An RNN-based algorithm to detect prosodic phrase for Chinese TTS'. Proc. ICASSP, 2001, Vol. 2, pp. 809–812
23 Jokisch, O., Ding, H., and Kruschke, H.: 'Towards a multilingual prosody model for text-to-speech'. Proc. ICASSP, May 2002, Vol. 1, pp. 421–424
24 Takano, S.: 'A Japanese TTS system based on multiform units and a speech modification algorithm with harmonics reconstruction', *IEEE Trans. Speech Audio Process.*, 2001, **9**, (1), pp. 3–10
25 Bigorgne, D., Boeffard, O., *et al.*: 'Multilingual PSOLA text-to-speech system'. Proc. ICASSP, April 1993, Vol. 2, pp. 187–190
26 Yunbo, Z., Zhao, Z., Xu, Y., and Niimi, Y.: 'A Chinese text-to-speech system based on TD-PSOLA'. Proc. ICASSP, 2002, Vol. 1, pp. 204–207
27 Hwang, S.H., and Yeh, C.Y.: 'An efficient text analyzer with prosody generator-driven approach for Mandarin text-to-speech'. Proc. ICASSP, April 2003, Vol. 1, pp. 488–491
28 Tseng, C.Y., and Chen, D.D.: 'The interplay and interaction between prosody and syntax: Evidence from Mandarin Chinese'. Proc. ICSLP, 2000, pp. 95–97
29 Chiang, T.H., Chang, J.S., and Lin, M.Y.: 'Statistical models for word segmentation and unknown word resolution'. Proc. of ROCLING V, 1992, pp. 121–146
30 Chang, J.S., Shu, R.H., and Chen, H.C.: 'The automatic detection method on the translation rule of preposition'. Proc. ROCLING IX, 1996, pp. 295–320
31 Chou, F.C., Tseng, C.Y., Chen, K.J., and Lee, L.S.: 'A Chinese text-to-speech system based on part-of-speech analysis, prosodic modeling and non-uniform units'. Proc. ICASSP, 1997, Vol. 2, pp. 923–926
32 Kim, Y.J., and Oh, Y.H.: 'Prediction of prosodic phrase boundaries considering variable speaking rate'. Proc. of ICSLP, October 1996, Vol. 3, pp. 1505–1508
33 Bachenko, J., and Fitzpatrick, E.: 'Prosodic phrasing for speech synthesis of written telecommunications by the deaf'. Proc. of GLOBECOM, December 1991, Vol. 2, pp. 1391–1395
34 Ma, X., Zhang, W., Shi, Q., Zhu, W., and Shen, L.: 'Automatic prosody labeling using both text and acoustic information'. Proc. ICASSP, April 2003, Vol. 1, pp. 516–519
35 Morin, J.Y.: 'Theoretical and effective complexity in natural language processing'. Proc. of ROCLING VIII, 1995, pp. 155–173
36 Luk, W.P.R.: 'Chinese-word segmentation based on maximal-matching and bigram techniques'. Proc. of ROCLING VII, 1994, pp. 273–282
37 Chen, K.J., Liu, S.H., and Chang, L.P.: 'A practical tagger for Chinese corpora'. Proc. of ROCLING VII, 1994, pp. 111–126