Chiang et al. EURASIP Journal on Audio, Speech, and Music Processing (2019) 2019:4
https://doi.org/10.1186/s13636-019-0147-y

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Punctuation-generation-inspired linguistic features for Mandarin prosody generation

Chen-Yu Chiang[1*], Yu-Ping Hung[1], Han-Yun Yeh[1], I-Bin Liao[2] and Chen-Ming Pan[2]

Abstract

This paper proposes two novel linguistic features extracted from text input for prosody generation in a Mandarin text-to-speech system. The first feature is the punctuation confidence (PC), which measures the likelihood that a major punctuation mark (MPM) can be inserted at a word boundary. The second feature is the quotation confidence (QC), which measures the likelihood that a word string is quoted as a meaningful or emphasized unit. The proposed PC and QC features are influenced by the properties of automatic Chinese punctuation generation and linguistic characteristic of the Chinese punctuation system. Because MPMs are highly correlated with prosodic–acoustic features and quoted word strings serve crucial roles in human language understanding, the two features could potentially provide useful information for prosody generation. This idea was realized by employing conditional random-field-based models for predicting MPMs, quoted word string locations, and their associated confidences—that is, PC and QC—for each word boundary. The predicted punctuations and their confidences were then combined with traditional linguistic features to predict prosodic–acoustic features for performing speech synthesis using multilayer perceptrons. Both objective and subjective tests demonstrated that the prosody generated with the proposed linguistic features was superior to that generated without the proposed features. Therefore, the proposed PC and QC are identified as promising features for Mandarin prosody generation.

Keywords: Conditional random field, Multilayer perceptron, Text-to-speech system, Prosody generation, Linguistic feature, Speech synthesis, Punctuation confidence, Mandarin

1 Introduction

Prosody generation serves a crucial role in a text-to-speech system (TTS). Prosody generation can be regarded as a function mapping from linguistic features to prosodic structures or prosodic–acoustic features. In a practical implementation of an unlimited-text Mandarin TTS (MTTS), the availability and reliability of linguistic features are highly dependent on the performance of the text analyzer employed. A basic text analyzer comprises a Chinese word segmenter, grapheme-to-phone (G2P) converter, and part-of-speech (POS) tagger. Prosodic structures are abstract descriptions of speech prosody and are usually categorically represented using prosodic break tags, such as nonbreak and minor or major break. A Mandarin prosody hierarchy that is commonly agreed upon is a four-layer prosodic structure. The four layers are, from the lowest to the highest, the syllable (SYL) layer, prosodic word (PW) layer, intermediate phrase (or prosodic phrase; PPh) layer, and intonation phrase (IP) layer, which are demarked by nonbreaks, minor breaks, major breaks, and utterance boundaries, respectively [1–3]. Prosodic–acoustic features are prosodic information that is numerically represented by values or vectors of the log-F0 contour, duration, and the energy of any linguistic domain, for example, a phone, syllable, initial or final, or word. The representative prosodic–acoustic features for Mandarin speech are the syllable log-F0 contour, syllable duration, pause duration, and syllable energy level [4–6]. In hidden Markov model (HMM)-based synthesis, the most popular speech synthesis method [7–10], prosodic–acoustic features are modeled at the HMM state level, that is, modeled using the state duration, state log-F0 value, and energy contour enclosed by the spectral parameters.

* Correspondence: cychiang@mail.ntpu.edu.tw
[1]Department of Communication Engineering, National Taipei University, 151, University Rd., San Shia District, New Taipei City 23741, Taiwan
Full list of author information is available at the end of the article

Irrespective of the target (prosodic structure or prosodic–acoustic features) of prosody generation, studies on prosody generation have focused on the following two problems: (1) design or utilization of a prediction model and (2) utilization of features. For the first problem, the prediction methods popularly used for generating a prosodic structure are the hierarchical stochastic model [11], the N-gram model [12], classification and regression tree (CART) [13, 14], bottom-up/sifting hierarchical CART [13], the Markov model [15], artificial neural networks [16], and the maximum entropy model [17]. Moreover, the popular pattern recognition tools for generating prosodic–acoustic features are a multilayer perceptron (MLP) [18–23], a recurrent neural network [4], CART [7–10, 24], and a decision tree with the hidden Markov model with multispace distribution modeling of the F0 contour [7–10]. For the second problem, conventional linguistic features, such as POSs, word length, sentence length, and position of a word in a sentence, are widely used in many existing MTTSs [4, 12–14, 17, 22, 24–27]. Some studies have improved the accuracy of prosodic structure prediction or prosodic–acoustic feature prediction by incorporating higher-level syntactic features, such as word chunks [16] and syntactic trees [16, 26, 27]. Moreover, statistical linguistic features such as connective degree [14], punctuation confidence (PC) [28–31], and quotation confidence (QC) [30, 31] have been proposed to neglect complex syntactic tree parsing and manual word chunking that is impractical when constructing an unlimited-text MTTS.

This paper focuses on the second problem to extend and elaborate on our previous research pertaining to PC [28–31] and QC [30, 31] features. A more substantial analysis and modeling details are provided in this paper to provide readers with an insight into the proposed PC and QC features, the design of which is influenced by automatic Chinese punctuation generation [32] and the linguistic characteristic of the Chinese punctuation system [33]. PC measures the likelihood of inserting a major punctuation mark (MPM) at a word boundary, whereas QC measures the likelihood of using a word string that is quoted by Chinese quotation marks (or brackets) to emphasize the meaning of the quoted word string. In [32], a maximum-entropy-based automatic Chinese punctuation generation method was proposed to insert 16 types of PMs into unpunctuated text by using word features and lexical–functional grammar features. The results in [32] indicated that the punctuation generation model could generate alternative or acceptable insertions, deletions, or substitutions of PMs. A successful outcome was also obtained in a punctuation experiment involving human readers, as reported by Tseng [33], in which the alternative punctuation strategies of different native Mandarin Chinese speakers were found. These observations reflect that Chinese PMs serve as a loose reference to the syntactic structure and semantic domain. Therefore, native Chinese writers can freely utilize PMs to delimit written Chinese into various linguistic elements, such as sentences, phrases, and clauses, for clearly expressing the meaning of text. Furthermore, the punctuation generation of a speaker when reading written Chinese reflects the speaker's prosodic phrasing strategy because pause breaks are highly correlated with some MPMs, such as the period, comma, exclamatory mark, question mark, semicolon, and colon. Therefore, an automatic punctuation generation model that predicts MPMs and is trained by using a large text corpus can learn punctuation strategies for predicting MPMs from various contributors for providing useful cues for predictions of both prosodic breaks [28, 31] and prosodic–acoustic features [29–31].

Word strings enclosed by brackets or quotes have essential or unique meanings in sentences. In our analysis of a large text corpus, the Academia Sinica Balanced Corpus of Modern Chinese (ASBC) 4.0 [34], which contains 9,454,734 words (31,126 paragraphs), we discovered that the functions of quoted word strings can be classified into several cases: (1) adding supplementary information to the proceeding words; (2) representing the name of a particular person, place, or institution; (3) emphasizing the meaning of a word string; and (4) indicating a newly derived compound word or word chunk that has a complex meaning. In cases (3) and (4), the quoted word strings, which are named quoted phrases in this paper, from small to large linguistic units, may form newly derived words, compound words, base phrases, word chunks, syntactic phrases, or sentences. The aforementioned linguistic units are usually larger than common words, contain more complex meanings than a word or may even have new meanings, and may be a higher-level unit in terms of the syntax compared with the POSs of words. Because a quoted phrase exhibits richer linguistic information than only words, it plays a crucial role in human language understanding during the reading of a text. Moreover, it is generally agreed that a speaker can generate good prosody if they understand the meaning of a text. Thus, adding quotations to plain Chinese text and then regarding the added brackets as linguistic features may enable a system to generate prosody that sounds natural. Note that in written Chinese, the use of quotations by adding brackets depends on the writing style or habit of the text contributor. Chinese input texts may thus already contain some brackets for the four functions indicated previously. However, the remaining unquoted words may also be emphasized and be regarded as larger syntactic units if they share similar contextual POSs or word structures with the quoted phrases. For Chinese texts containing

no quotations, if quotations can be labeled with brackets automatically by a machine when the word and POS information are given, then the features associated with the labeled brackets could provide richer linguistic information and thus enhance the performance of prosodic–acoustic feature prediction.

To realize the use of automatic MPMs and for quotation predictions, we constructed two types of conditional random field [35, 36] (CRF)-based automatic punctuation generation models: the CRF-based MPM generation model and CRF-based quotation generation model. The CRF-based MPM generation model predicts MPMs and generates the associated confidence measures, which are referred to as the PC, through MPM-removed word or POS sequences. The PC can be regarded as a statistical linguistic feature measuring the likelihood of correctly inserting an MPM into a text. Word junctures in which MPMs are more likely to be inserted are, it is reasonable to assume, junctures in which pause breaks are more likely. We could, therefore, expect that the utilization of PC in prosody generation would improve the performance of prosodic–acoustic feature generation. The CRF-based quotation generation model predicts the structure of a quoted word string (hereafter referred to as the quoted phrase, or QP) from the bracket-removed word or POS sequences and calculates the associated confidence, which is referred to as the QC. The QC can also be considered a statistical linguistic feature used for measuring the likelihood of word strings that are quoted using left and right brackets. Because words in brackets constitute meaning, it is reasonable to assume that fewer prosodic breaks are inserted within quoted text and that quoted text may be emphasized using some variation in prosodic–acoustic features. Therefore, we inferred that the use of QC may assist in prosody generation.

To evaluate the usefulness of the proposed PC and QC in Mandarin prosody generation, experiments of prosodic–acoustic feature prediction were conducted, and the corresponding objective and subjective tests were evaluated. The experimental database used was a Mandarin speech corpus, the Treebank speech corpus, which contains 425 utterances with 56,237 syllables uttered by a professional female announcer. The corpus is further divided into three parts: a training set of 301 utterances with 41,317 syllables, a development set of 75 utterances with 10,551 syllables, and a test set of 44 utterances with 3898 syllables. The corpus used for training the CRF-based punctuation generator was the ASBC 4.0 [34] (hereafter denoted as the ASBC text corpus). For the prosodic–acoustic feature prediction, the proposed linguistic features combined with conventional linguistic features were employed as the input to directly predict four prosodic–acoustic features of the syllable log-F0 contour, syllable duration, syllable energy level, and intersyllable pause duration. Objective tests were evaluated using the root-mean-square error (RMSE). Subjective tests were then conducted on speech-synthesized utterances by using the predicted prosodic–acoustic features.

Several advantages of the approach were discovered. First, the PC and QC were conveniently determined from the features of word or POS sequences robustly obtained by performing segmentation of the current word and employing POS-tagging technologies without using complicated statistical syntactic parsing. This advantage makes the proposed approach suitable for practical online unlimited TTS. Second, because the CRF-based punctuation generation models were trained by using a large text corpus, the models could learn alternative punctuation strategies from numerous paragraphs by various writers to generate more reliable PCs and QCs. Third, compared with the size of an available text corpus for constructing a statistical syntactic parser, the size of the corpus used to train the CRF-based punctuation generator was considerably larger. Therefore, we infer that the obtained PC and QC are more robust than the syntactic features derived from an automatic syntactic parser.

The research process and corresponding section organization of this paper are summarized as follows:

- Section 2: Analysis of punctuations

We demonstrate the relationship between punctuations and prosodic structures by analyzing the Treebank speech corpus, which is labeled with prosodic break tags. The analyses that motivated our use of the proposed PC are explained. This section also analyzes the quoted phrases in the ASBC text corpus, thus identifying possible QC candidates for the training of the CRF-based quotation model.

- Section 3: Construction of the CRF-based MPM generation model

The CRF-based MPM generation model was trained by using the ASBC text corpus. The precision and recall of the MPM insertions are examined on the test dataset of the ASBC text corpus. The feasibility of using the proposed PC in prosody generation was examined by analyzing the relationship between the prosodic–acoustic features of the training dataset of the Treebank speech corpus and the associated PC generated using the CRF-based MPM generation model.

- Section 4: Construction of the CRF-based quotation generation model

The model was trained and examined using the ASBC text corpus. The feasibility of using the QC for prosody generation was determined using the Treebank speech corpus.

- Section 5: Prosody generation experiments

The prosody generation experiments were conducted on the Treebank speech corpus. The PC and QC features generated by the proposed automatic punctuation generation models by using the Treebank text corpus were combined with the conventional linguistic features to predict the prosodic–acoustic features of the syllable pitch contour, syllable duration, syllable energy level, and pause duration. Objective and subjective tests were conducted to verify the usefulness of the proposed PC and QC features.

- Section 6: Conclusions and future work

2 Analysis of punctuations

Because prosodic–acoustic features are highly dependent on Mandarin's prosodic structure and the prosodic structure is categorically represented by a finite set of prosodic break tags, it is more convenient to analyze the relationship between prosodic break types and PMs than to analyze the relationship between numerical prosodic–acoustic features and PMs. Therefore, the relationship between Chinese PMs and Mandarin prosodic structure is analyzed in this section. The following subsections present the analyses that provided the motivations and rationality for using the proposed PC and QC features. The prosody labeling system for determining the prosodic structures of utterances is introduced in Section 2.1. The relationship between the labeled prosodic break types and PM types is discussed in Section 2.2. Section 2.3 presents the experimental process wherein native Mandarin speakers were allowed to manually insert MPMs in PM-removed texts excerpted from the Treebank speech corpus. The relationships between the manually inserted MPMs by the native Mandarin speakers and the associated prosodic break types are analyzed, thus providing evidence for the proposed PC. An analysis of the quoted phrases in the ASBC text corpus is presented in Section 2.4, identifying the possible QC candidates for the training of the CRF-based quotation generation model.

2.1 Prosody labeling system

The widely used prosody labeling systems are ToBI [37], TILT [38], and C-ToBI [39]. These prosody labeling systems require manual labeling by humans with linguistic expertise. To reduce the human labor required and to increase the consistency of prosody labeling, Chiang et al. [40, 41] proposed an unsupervised joint prosody labeling and modeling (PLM) method for constructing a speaker-dependent statistical hierarchical prosodic model and labeling prosody tags for Mandarin speech. The PLM method was then successfully applied to construct a speaker-independent hierarchical prosodic model for use in a large vocabulary speech recognition task [42]. Hence, in this study, to avoid the need for intensive human labeling and inconsistent labeling results, the corpus was labeled with seven break types using the PLM method [40, 41] proposed by Chiang et al. As illustrated in Fig. 1, the seven break types—*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, and *B*4—delimit an utterance into four types of prosodic units: a SYL, PW, PPh, and breathe group or prosodic phrase group (BG/PG).

In the labeling system, each defined break type is characterized by its specific juncture's prosodic–acoustic features. *B*4 is defined as a major break and contains a long pause and apparent F0 reset across adjacent syllables. *B*3 is a major break with a medium pause and medium F0 reset. *B*0 and *B*1 are nonbreaks of a tightly coupled syllable juncture and a normal syllable boundary within a PW, respectively, which have no identifiable pauses between SYLs. Moreover, *B*2 is a minor break with three variants—an F0 reset (*B*2-1), short pause (*B*2-2), and preboundary syllable duration lengthening (*B*2-3).

Among the various types of prosodic–acoustic features, pause duration is the most salient cue for specifying the boundaries of prosodic units. Figure 2 shows probability density functions (pdfs) of Gamma distributions for the seven break types and reveals that the higher-level break types were generally associated with longer pause durations. According to the pdfs of pause durations for each of the break type shown in Fig. 2, the long pause of *B*4 has pause duration $>= 400$ ms, the medium pause of *B*3 has the pause duration in the interval of 200 ~ 400 ms, and the short pause of *B*2-2 has the pause duration in the interval of 30–200ms. On the other hand, *B*0, *B*1, *B*2-1, and *B*2-3 have very short

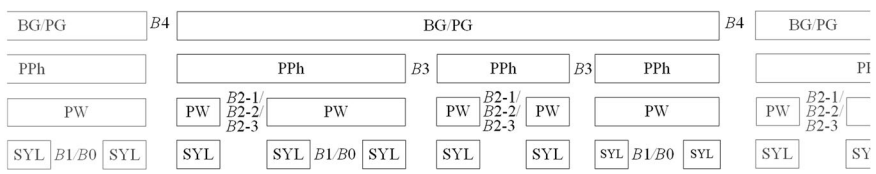

Fig. 1 Hierarchical prosodic model of Mandarin speech used in this study [42]

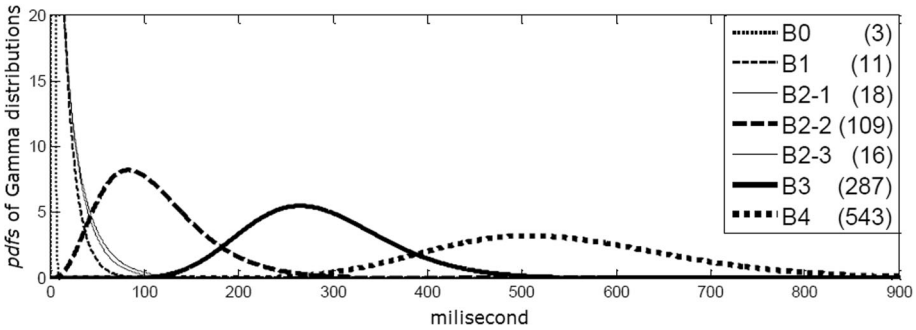

Fig. 2 Distribution (*pdf*) of pause durations (ms) for the seven break types. The average pause duration (ms) for each of the prosodic break types is displayed in parentheses

pause durations (< 30 ms). Specifically, on the basis of this analysis of the pause duration of the seven break types, this study defined four break classes for conveniently conducting the analysis presented in Section 2.2: (i) *B*4, (ii) *B*3, (iii) *B*2-2, and (iv) the nonpause break (NPB) type that comprises *B*0, *B*1, *B*2-1, and *B*2-3.

2.2 Relationship between the labeled break types and PM types

In general, pause breaks are considered to co-occur with PMs. Most TTSs cautiously insert pauses only for MPMs, such as commas and periods. This cautious strategy of pause insertion can make synthesized speech very clear but may sound unnatural because the input sentence can be very long and contain complicated syntactic structures. Table 1 displays the co-occurrence matrix of the four break classes and three syllable juncture types, calculated using the training dataset of the Treebank speech corpus. The table reveals that most PM locations co-occur with breaks of the pause-related type (*B*2-2, *B*3, and *B*4), whereas most intraword locations map to NPBs. Non-PM interword locations co-occur with NPBs, *B*2-2, and *B*3. Approximately 40% of prosodic phrase boundaries (*B*3s) and more than 94% of *B*2-2s occur at non-PM interword junctures. By conducting a more detailed analysis, we found that 60% of non-PM *B*3s coincide with the depth-1 node boundary of a fully parsed syntactic tree. These results imply that inserting pauses only at PM locations would be unsatisfactory.

Table 2 displays the co-occurrence matrix of the four break classes and eight PM types that exist in the Treebank speech corpus. The PM types in the MPM set with sufficient samples—that is, {comma ",", period "。", semicolon ";", question mark "?", exclamation mark "!"}—are highly correlated with the major breaks of *B*3s and *B*4s. This implies that a word juncture that inserts an MPM in a text is more likely to be a major break in an utterance. This motivated us to propose a CRF-based automatic MPM generator in this study to predict the insertion of MPM and its likelihood (i.e., PC) for each word juncture and to use the predicted MPMs and PC for prosody generation.

Table 1 Co-occurrence matrix of four target break types and three syllable juncture types

	NPB	B2–2	B3	B4
Intraword	21,970	14	2	0
Non-PM interword	20,288	3148	1391	30
PM	30	169	2130	2320

In the texts of the training dataset, the Treebank speech corpus, no word strings were quoted in Chinese brackets. Thus, we could not directly analyze the relationship between Chinese brackets and labeled break types. In this study, we directly analyzed the characteristics of the brackets and their associated quoted phrases from the ASBC text corpus, as presented in Section 2.4.

2.3 Human-labeled PMs versus prosodic break types

From the results displayed in Table 2, we concluded that occurrences of *B*3s and *B*4s are highly correlated with periods, exclamation marks, question marks, semicolons, colons, and commas. Therefore, we assumed that automatic punctuation generation models that predict MPMs and are trained using a large text corpus can learn strategies for inserting MPMs from texts by various contributors to provide informative cues for prosodic–acoustic feature prediction. To access the feasibility of this idea, we conducted an experiment in which 10 native Mandarin speakers were asked to insert periods and commas in the same 30 PM-deleted short paragraphs. These 30 paragraphs were chosen from the Treebank speech corpus texts labeled with prosodic breaks, as stated in Section 2.1. The longest and shortest paragraphs were 270 and 80 characters, respectively, and the average length was 138 characters. The frequencies (or probabilities) of word junctures that were added with periods or commas were regarded as PCs labeled manually by humans (or text contributors). An analysis of the relationship

Table 2 Correlation matrix of the four break types and the eight PM types

	Comma "，"	Period "。"	Semicolon "；"	Question mark "？"	Exclamation mark "！"	Colon "："	Chinese back-sloping comma "、"	Chinese back-sloping comma "、"	Partition sign "　"
NPB	4	1	0	0	0	0	25	25	1
B2-2	88	2	0	1	1	1	75	75	1
B3	1901	42	9	7	1	2	168	168	0
B4	1523	606	63	58	39	0	30	30	1

between these frequencies (PCs labeled by humans) and labeled prosodic breaks provided some evidence that the proposed method is feasible.

Figures 3 a–c display the average percentage for each prosodic break type with respect to the number of times that a word juncture was inserted with a comma, period, and comma or period, respectively. Here, the number of times that a comma or period was inserted is analogous to the proposed PC. The percentage of NPBs decreased considerably when the frequency of MPM insertions increased. Figure 3a displays that the percentage for *B*4 increased as the frequency of comma insertion increased. The percentage for *B*3 was highest when two or three commas were inserted and then decreased and maintained a constant level for when more than four commas were inserted. The percentage for *B*2-2 exhibited a similar trend to that for *B*3 but at a lower magnitude. Figure 3b displays that *B*4 dominated when more than three insertions of periods were observed for each word juncture. These results indicate that a word juncture was more likely to be inserted with pause-related break types—*B*2-2, *B*3, and *B*4—when the PC was larger. Moreover, the break types of the higher prosodic units (i.e., larger break types) were associated with a larger PC. Figure 3c can be viewed as the combined result of Fig. 3a and b. Because commas and periods are major constituents of the MPM set, the result displayed in Fig. 3c is analogous to the distributions of the prosodic break types pertaining to the PC values. We observed more evident trends for the percentages of the four break classes in Fig. 3c than the trends shown in Fig. 3a and b and found that these trends were informative for prosody generation.

2.4 Analysis of quotations

Table 3 displays the 26 Chinese quotation mark pairs that are used in the ASBC text corpus [34]. We categorized these quotation mark pairs into ten types according to the functions of the enclosed words, i.e., QPs. Table 4 lists the types of the quotation mark pairs, their statistics, and the associated exemplar QPs in sentences with word-by-word Chinese–English translations. The

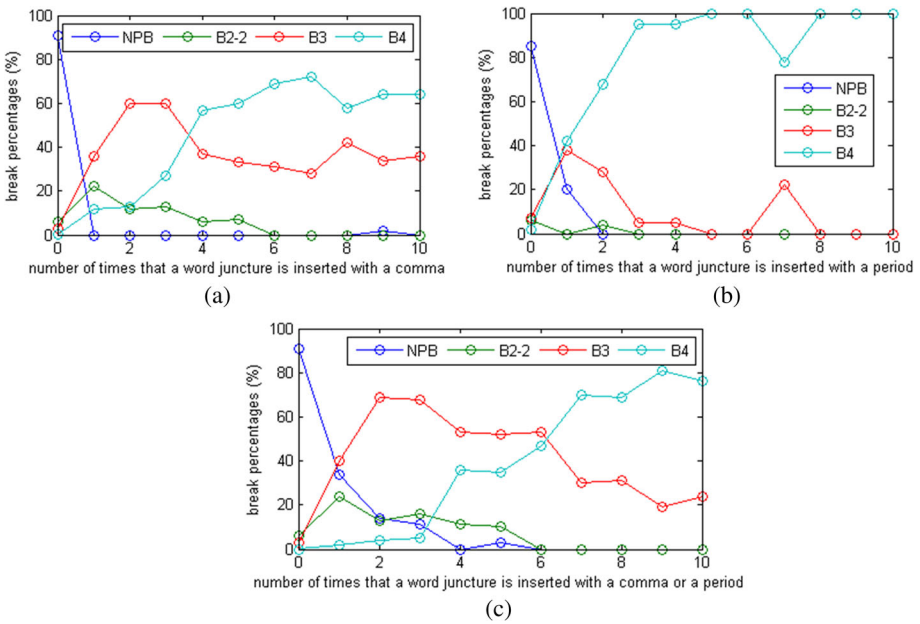

Fig. 3 Average percentage of prosodic break types dependent on the number of times that a word juncture was inserted with a comma (**a**), period (**b**), and comma or period (**c**)

Table 3 Types of Chinese quotation marks

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Quotes	(	)	(	)	{	}	{	}	〔	〕	〔	〕	「	」	『	』	〈	〉	【	】	《	》	〝	〞	〝	〞
Type	1				2				3				4		5		6		7		8		9		10	

characteristics of the types of QPs are introduced as follows:

Type 1 (): They are mostly used for enumeration. Therefore, we did not regard Type 1 as prediction targets for QPs.
Type 2 {}: They are mostly used for titles of books or articles; thus, we regarded this type as a prediction target.
Type 3 〔〕: They mostly function as captions of articles. This type was not included in our prediction target.
Types 4 and 5 「」and『』: These types contributed most samples (68%) for the QP predictions because they generally enclosed word chunks or base phrases. Single-word QPs of these types are usually emphasized nouns, verbs, or idioms. Most two- to four-word QPs are noun phrases. QPs that are longer than four words are generally long noun phrases or sentences.
Types 6, 7, and 8 〈〉【】《》: These types are similar to Type 2 and thus were included in the QP prediction.
Type 9 " ": We included the samples of this type in the QP prediction. In this type, single-word QPs are generally proper nouns. The two- to four-word QPs are frequently used phrases, and five- to six-word QPs are similar to sentences.
Type 10 〝〞: This type is similar to types 4 and 5. We employed this type as a QP prediction target, although the sample size was very small.

Table 5 displays statistics on the numbers of words in QPs. Most QPs were found to be single- to four-word QPs. Single-word QPs are usually emphasized nouns or verbs. Two- to four-word QPs are mostly base phrases such as word strings (or word chunks). QPs longer than four words are mostly sentence-like units.

3 Proposed PC

3.1 CRF-based MPM generator

PC [28] is obtained by a CRF-based MPM generator. The CRF-based MPM generator overcomes the label-tagging problem because it labels each lexical word juncture with a sequence of types of PMs—for example, presence or absence of an MPM $\boldsymbol{Y}$ by using some linguistic feature sequence $\boldsymbol{X}$. The function of the CRF-based MPM generator is formulated as follows:

Table 4 Types of QPs, their statistics, and examples in sentences. Note that words in the examples are delimited by slashes and word-by-word English translations of Chinese words are provided

Type		Count (%)	Examples of QPs in sentences
1	()	14,131 (25.13%)	**(**一**)/**不/抽菸/、**/(**二**)/**不/熬夜/, (1)/no/smoking/,/(2)/no/staying up night/,
2	{}	34 (0.06%)	**{**桃花源記**}/**是/陶淵明/的/作品/, **{Tao-Hua-Yuan-Ji}**/is/Mr. Yuan-Ming Tao/possession indicator/work/,
3	〔〕	101 (0.17%)	〔本報訊〕**/**全省/橄欖球/錦標賽/快/開始/了/, 〔**News**〕/whole province/football/ championship contest/soon/start/modal particle/,
4	「」	37,197 (66.17%)	「十八歲**/**的**/**約定」**/**是/2002年/台灣/的/一部/偶像劇/, 「**Eighteen years old/possession indicator/commitment**」/is/year 2002/Taiwan/possession indicator/one/idol drama/,
5	『』	1223 (2.17%)	瓊瑤/哀痛/寫下/『與**/**夫**/**訣別**/**書』**/**宣布/關閉/臉書/,Chung-Yao/grief-stricken/write/『**to/husband/farewell/letter**』/announce/shut down/Facebook account/,
6	〈〉	562 (0.99%)	〈銀鬚**/**上**/**的**/**春天〉**/**是/黃春明/先生/的/小說/中/較/特殊/的/一篇/。 〈silver whisker/upon/possession indicator/spring〉/is/Mr. Chueng-Ming Huang/possession indicator/novel/among/more/special/possession indicator/one piece/.
7	【】	314 (0.55%)	【族群**/**與**/**文化**/**政策**/**綱領】**/**已經/上線/發布/, 【**ethnic group/and/culture/policy/principle**】/already/online/publish/,
8	《》	2523 (4.48%)	《屋頂**/**上**/**的**/**小孩**》/**中/濃厚/的/南方/氣息/, 《**Roof/top/possession indicator/child**》/among/deep/possession indicator/southern/atmosphere/,
9	""	105 (0.18%)	美國/作曲家/柯維爾/用/ **"**新**/**音樂**"/**為/主題/, American/composer/Cowell/take/**"new/music"**/as/title/,
10	〝〞	22 (0.04%)	最近/有/一部/名為/〝第一**/**夫人〞**/**的/電影/上映/, Recently/there/one/so-called/〝**first/lady**〞/possessive indicator/movie/be on/,

Table 5 Statistics on the lengths of QPs

No. of words	No. of example	Percentage
1	26,791	41%
2	16,749	25%
3	10,933	17%
4	5847	9%
5	3415	5%
6 or larger	1988	3%

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{N(\mathbf{X})} \exp\left(\sum_{t=1}^{T}\sum_{i=1}^{I}\lambda_i f_i(Y_t = y, Y_{t-1}, \mathbf{X})\right) \quad (1)$$

where $N(\boldsymbol{X})$ is the normalization factor ensuring that $\Sigma_{\mathbf{Y}} P(\mathbf{Y}|\,\boldsymbol{X}) = 1$; t represents the lexical word index; Y_t represents the prediction target, that is, the type of PM between the t-th and $(t+1)$-th lexical words; I represents the number of feature functions; and $f_i(Y_t = y, Y_{t-1}, \mathbf{X})$ is a feature function defined by

$$f_i(Y_t = y, Y_{t-1}, \mathbf{X}) = \begin{cases} 1, \text{if } \mathbf{X} = h_j \text{ is satisfied and } y = y_k \\ 0, \text{otherwise} \end{cases} \quad (2)$$

where h_j represents the j-th possible linguistic feature context and y_k is the k-th possible tag (i.e., PM type) to be predicted. Feature contexts are generally organized into several groups, referred to as "feature templates." The predicted PM sequence can be obtained by the Viterbi search as follows:

$$Y_1^*, Y_2^*, \cdots, Y_T^* = \arg\max_{Y_1, Y_2, \cdots, Y_T} P(\mathbf{Y}|\mathbf{X}) \quad (3)$$

The PC is obtained through forward or backward calculation and is equal to the marginal probability of the k-th type of PM for the t-th word:

$$\phi_{t,k}(\mathbf{X}) = P(Y_t = y_k|\mathbf{X}) \quad (4)$$

3.2 Design of the prediction targets

Two types of prediction targets were designed—the basic PC (bPC) and improved PC (iPC). The bPC is generated by considering two prediction targets—the presence of an MPM, y_1, and the absence of an MPM, y_0. The iPC is produced by considering the structures of sentences accompanied by MPMs. For the bPC, the MPMs include "。", "!", "?", ";", ":", and ",". The PC, $\phi_{t,\,k}(\mathbf{X})$, generated by the target setting $\{y_1, y_0\}$ is known as bPC. Figure 4a displays the original text with word or PM tokens, and Fig. 4b presents the corresponding target-labeling example for the training of the bPC.

Note that the bPC only considers modeling the insertion of MPMs, and the MPMs serve as delimiters for sentences, phrases, or clauses for Chinese. For convenience's sake, a linguistic unit between the two closest MPMs is defined as a sentence-like unit in this study. Therefore, modeling structures of sentence-like units could be equivalent to modeling insertions of MPMs and could even provide superior prediction of MPM insertion. We proposed the iPC to model structures of sentence-like units and optional MPMs in a sentence-like unit. The two types of iPCs are improved PC with structure of sentence-like unit (iPCst) and improved PC with enforced major punctuation mark insertion (iPCef). Here, iPCst is designed for modeling structures of sentence-like units, whereas iPCef is defined for modeling enforced MPM insertion into a sentence-like unit. For the prediction of iPCst, the prediction targets for the CRF-based MPM generator were labeled for each word and designed to represent structures of sentence-like units regarding word position in a sentence-like unit. The targets "B," "I," "M," "S," and "E" represent beginning, intermediate, middle, single, and ending words, respectively, in a sentence-like unit. To further precisely label the word order information in a sentence-like unit, numbers 1 to 4 were added to the targets "B" and "E" for indicating forward and backward word order, e.g., B2 for the second word from the beginning of a sentence-like unit, and B3 for the third last word from the end of a sentence-like unit. On the basis of the statistics pertaining to the length of sentence-like unit (unit: words) for the ASBC text corpus, the length of sentence-like units is mostly (84%) between four and nine words. Therefore, the target-labeling schemes were designed differently for sentence-like units with ≤ 9 and < 9 words. The complete targets for iPCst are listed in Table 6. Specifically, four rules were employed to guide target tagging:

1. "B1," "B2," "B3," and "B4" represent the first, second, third, and fourth word in a sentence-like unit, respectively, whereas "E1," "E2," "E3," and "E4" represent the last, penultimate, antepenultimate, and preantepenultimate word in a sentence-like unit, respectively.
2. If the length of a sentence-like unit was < 9 words, then we used "B1"–"B4" and "E4"–"E1" to tag targets from the beginning and end of a sentence-like unit and used "M" to tag the other intermediate words in the sentence-like unit.
3. If the length of a sentence-like unit was ≤ 9 words and even, we used "B1"–"B*k*" and "E1"–"E*k*" to tag targets from the beginning and end of a sentence-like unit for $k = 1$–4 and k = length of sentence-like unit in words/2.
4. If the length of a sentence-like unit was ≤ 9 words and odd, we employed "B1"–"B*k*" and "E1"–"E*k*" to tag targets from the beginning and end of a

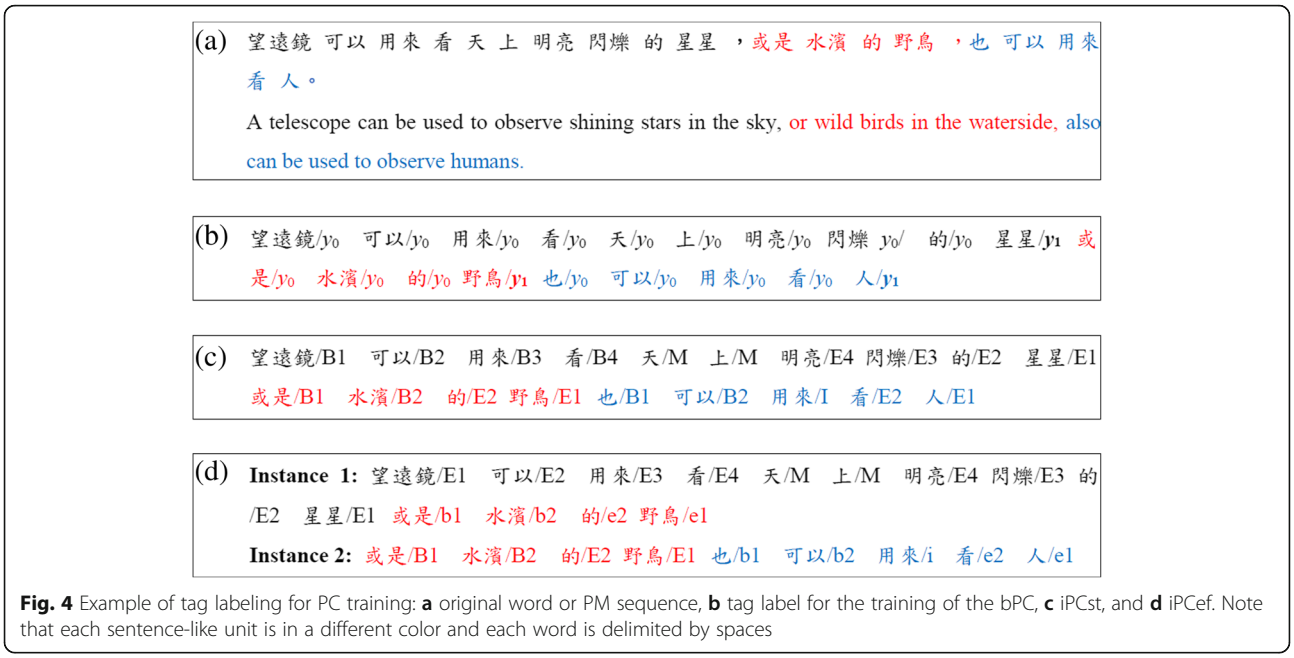

Fig. 4 Example of tag labeling for PC training: **a** original word or PM sequence, **b** tag label for the training of the bPC, **c** iPCst, and **d** iPCef. Note that each sentence-like unit is in a different color and each word is delimited by spaces

sentence-like unit for $k = 1–4$ and k = length of sentence-like unit in words/2. The remaining words were labeled with "I" to indicate that they are intermediate words in the sentence-like unit.

Figure 4c displays an example of tag labeling for the iPCst training.

The prediction of iPCef is to enforce the insertion of an MPM in a sentence-like unit. This idea is motivated by the observation of the ASBC text corpus [34] that optional MPMs could be inserted into some long sentence-like units without a loss of understanding. In the training of iPCef, two consecutive sentence-like units were considered as one training instance for an enforced MPM insertion. The target set for iPCef is similar to that for iPCst, as shown in Table 6. However, the target set for iPCef uses upper- and lower-case letters for the distinction between tags for the first and second sentence-like units, respectively. Figure 4d shows two training instances, i.e., Instance 1 and Instance 2, extracted from the text displayed in Fig. 4a. Instance 1 of Fig. 4d is made of the first and the second sentence-like units of Fig. 4a while Instance 2 of Fig. 4d is made of the second and the third ones of Fig. 4a. In the testing phase that generates the PM type labels and the iPCefs for prosody prediction, i.e., $\phi_{t,\,k}(\mathbf{X})$, each sentence-like unit is labeled with the sequences of the PM type and the associated iPCefs by the CRF-based punctuation generator and the given linguistic features of each sentence-like unit. The CRF-based punctuation generator surely inserts an MPM into each sentence-like unit. Therefore, we call this target labeling and testing for generating iPCef "enforced MPM insertion." This enforced MPM may provide informative cues for inserting a pause or cause preboundary syllable duration lengthening for word junctures in a long sentence-like unit.

3.3 Design of features and templates

The linguistic features used in the CRF training are lexical words (W_t), POSs (S_t), and word length (L_t). Therefore, the linguistic feature sequence for the CRF model is.

$$\mathbf{X} = \{X_1, X_2 \cdots, X_T\} \text{ and } X_t = \{W_t, S_t, L_t\} \tag{5}$$

The linguistic features are generated by the NCTU Chinese parser [43, 44]. The significance of these linguistic features is summarized in Table 7.

Table 6 Targets for iPCst

Target tag: position in a sentence-like unit		
B1: 1st word	I: intermediate word if length of sentence-like unit in word is odd and less than 9	E4: 4th last word
B2: 2nd word	M: intermediate word if length of sentence-like unit in word is equal or more than 9	E3: 3rd last word
B3: 3rd word		E2: 2nd last word
B4: 4th word,		E1: 1st last word
		S: single word

Table 7 Significance of the linguistic features

Feature	Definition	Description
W_t	t-th lexical word	The smallest meaningful linguistic unit
S_t	Part of speech (POS) of t-th lexical word	Basic syntactic role of t-th lexical word; 47 categories [45]
L_t	Length of t-th lexical word in syllable	Longer words are more likely to be followed by PMs

The feature templates for the training of the CRF-based MPM generator for PCs considered the contextual word, POSs, length of the word, and combinations of these features. We designed four templates for PC generation, as shown in Table 8. All the templates consider the same POSs, lexical word POSs, and word length contexts. The difference between templates 1 and 2 is that template 2 considers more varied word contexts. Templates 3 and 4 are similar to templates 1 and 2 but templates 3 and 4 add a combination of the previous target Y_{t-1} (i.e., bigram templates) and the POS of the current word S_t. The reason for this combination is that the type of current PM, Y_t, depends on the joint factors of the previous PM type, Y_{t-1}, and the current POS, S_t.

3.4 Experiment of PC generation

The CRF models were trained using the ASBC text corpus [34] training dataset containing 6,625,277 words, and the optimal feature templates were then tuned by the results obtained using the training set with 2,817,785 words. The tool used for the training was CRF++, a CRF toolkit [36]. Table 9 displays the precision and recall of the predicted MPM insertions trained by setting the prediction targets bPC, iPCst, and iPCef using templates 1–4. The optimal precision and recall are achieved using template 4, followed by templates 3, 2, and 1. This indicates that wider feature contexts and joint factors of (Y_{t-1}, S_t) improve MPM prediction. The optimal precision and recall of MPM generations in the test set for bPC, iPCst, and iPCef were 94.1% and 93.1%, 96.9% and 96.1%, and 95.7% and 95.5%, respectively. We selected the results obtained using template 4 for the following analysis and prosody generation experiments. The values were reasonably high and thus suitable for modeling the characteristics of MPM insertion and structures of sentence-like units.

We then examined the interplay between the proposed PC values—that is, $\phi_{t,\ k}(\mathbf{X})$—and the distributions of prosodic–acoustic features in the training set of the Treebank speech corpus, as displayed in Figs. 5, 6, and 7. Figure 5 displays the average syllable log-F0 corresponding to the prediction targets for (a) bPC, (b) iPCst, and (c) iPCef in different levels of PC values. Note that the PC values are divided into 10 equal-width intervals from 0 to 1 for bPC in Fig. 5a. As can be seen from Fig. 5, the average syllable log-F0 decreases as the bPC for MPM—that is, $\phi_{t,\ k}(\mathbf{X})$—for the prediction target y_1 increases, whereas the bPC for y_0 exhibits a contrary trend. This indicates that a syllable had a lower log-F0 value because the syllable was more likely to be followed by an MPM. Figure 5b displays the average syllable log-F0 of the prediction targets in the three representative levels of iPCst values—the high level: iPCst = 0.9–1.0; median level: iPCst = 0.5–0.6; and low level: iPCst = 0.0–0.1. Note that the prediction targets are listed in a forward position order in the sentence-like unit on the x-axis; that is, "B1," "B2," "B3," "B4," "I"/"M," "E4," "E3," "E2," and "E1." A clear declining trend of log-F0 was found for the high-level iPCst. By contrast, the average syllable log-F0 is flat for the low-level iPCst. The average syllable log-F0 for the median-level iPCst displays a moderate log-F0 declining trend. Figure 5c displays the average syllable log-F0 of the prediction targets in the three representative levels of iPCef values. The prediction targets in Fig. 5c are also listed in a forward position order in a sentence-like unit on the x-axis. The log-F0 declination is also clearly observed for the cases of the high and median levels of iPCef values. These findings may indicate that the proposed PCs provided informative cues for modeling the decrease in log-F0 during prosody generation. Furthermore, iPCst and iPCef (especially iPCef) exhibited a higher and lower log-F0 at the beginning and end of a sentence-like unit, respectively, indicating that the proposed iPCst and iPCef may provide more useful cues for prosody generation than those provided by bPC.

Table 8 Feature templates for PC. The notation represents a sequence: $W_{t-l}, W_{t-l+1} \ldots W_t \ldots W_{t+u-1}, W_{t+u}$

	Template 1	Template 2	Template 3	Template 4
Lexical word context	W_t	$\{W_{t+\tau}\}_{\tau=-1\sim+1}, \{W^{t+\tau}_{t-1+\tau}\}_{\tau=0,1} W^{t+1}_{t-1}$	W_t	$\{W_{t+\tau}\}_{\tau=-1\sim+1}, \{W^{t+\tau}_{t-1+\tau}\}_{\tau=0,1} W^{t+1}_{t-1}$
POS context	$\{S_{t+\tau}\}_{\tau=-3\sim+3}$, $\{S^{t+\tau}_{t-1+\tau}\}_{\tau=0,1}, \{S^{t+\tau}_{t-2+\tau}\}_{\tau=0\sim2}, \{S^{t+\tau}_{t-3+\tau}\}_{\tau=0\sim3}, \{S^{t+1+\tau}_{t-3+\tau}\}_{\tau=0\sim3}, \{S^{t+2+\tau}_{t-3+\tau}\}_{\tau=0,1}$			
Lexical word and POS context	$\{(W_t, S_{t+\tau})\}_{\tau=-3\sim+3}, \{(W_t, S^{t+\tau}_{t-1+\tau})\}_{\tau=0,1}, \{(W_t, S^{t+\tau}_{t-2+\tau})\}_{\tau=0\sim2}, \{(W_t, S^{t+\tau}_{t-3+\tau})\}_{\tau=0\sim3}, \{(W_t, S^{t+1+\tau}_{t-3+\tau})\}_{\tau=0\sim3}$			
Lexical word length	$\{L_{t+\tau}\}_{\tau=-1\sim+1}$			
Previous Target & POS context	Y_{t-1}	Y_{t-1}	(Y_{t-1}, S_t)	(Y_{t-1}, S_t)

Table 9 Precision and recall of the MPM generations, as obtained using target-labeling methods for bPC, iPCst, and iPCef

	bPC		iPCst		iPCef	
	Precision	Recall	Precision	Recall	Precision	Recall
Template 1	0.902	0.867	0.961	0.949	0.940	0.937
Template 2	0.919	0.890	0.962	0.951	0.942	0.938
Template 3	0.905	0.869	0.967	0.959	0.955	0.953
Template 4	0.941	0.931	0.969	0.961	0.957	0.955

Figure 6 displays the average syllable duration corresponding to the prediction targets for (a) bPC, (b) iPCst, and (c) iPCef in different levels of PC values. Figure 6a reveals that the average syllable duration was shorter for the two extreme cases—bPC for $y_1 < 0.1$ and bPC for $y_0 > 0.9$. This indicated that bPC provides cues to shorten (lengthen) syllable durations when it is very unlikely (likely) that an MPM will be inserted after the target syllable. Figure 6b displays the average syllable durations of the prediction targets in the high, median, and low levels of iPCst. Note that the prediction targets are also listed in a forward position order in a sentence-like unit on the x-axis. Significantly long average syllable durations were found for the high and median iPCst levels at the prediction target of "E1," which represents a syllable followed by an MPM. A slightly longer average syllable duration was observed for the target "M" because this target represents an intermediate location in a long sentence-like unit and "M" is more likely to be inserted with a prosodic break in a long sentence-like unit. The average syllable durations for the predictions of the low-level iPCst are almost in the same level. These results indicate that the proposed iPCst models the preboundary syllable duration lengthening effect with various iPCst values. Moreover, for the prediction target "S," which represents a word enclosed by preceding and following MPMs, the syllable is lengthened because iPCst is high. The prediction targets "B1" (the first syllable in a sentence-like unit) and "I" (the intermediate syllable in a short sentence-like unit) have shortened average syllable durations compared with their nearby syllables in a sentence-like unit. These results are in agreement with the findings of a previous study [46] concerning syllable durations in a PPh. In the study [46], the first syllable in a PPh and the intermediate syllable in a short PPh were found to be shortened. The shortened syllable duration for the target "E2" (the second last syllable in a sentence-like unit) significantly contradicted the following preboundary syllable duration lengthening cue for the prediction target "E1." Figure 6c

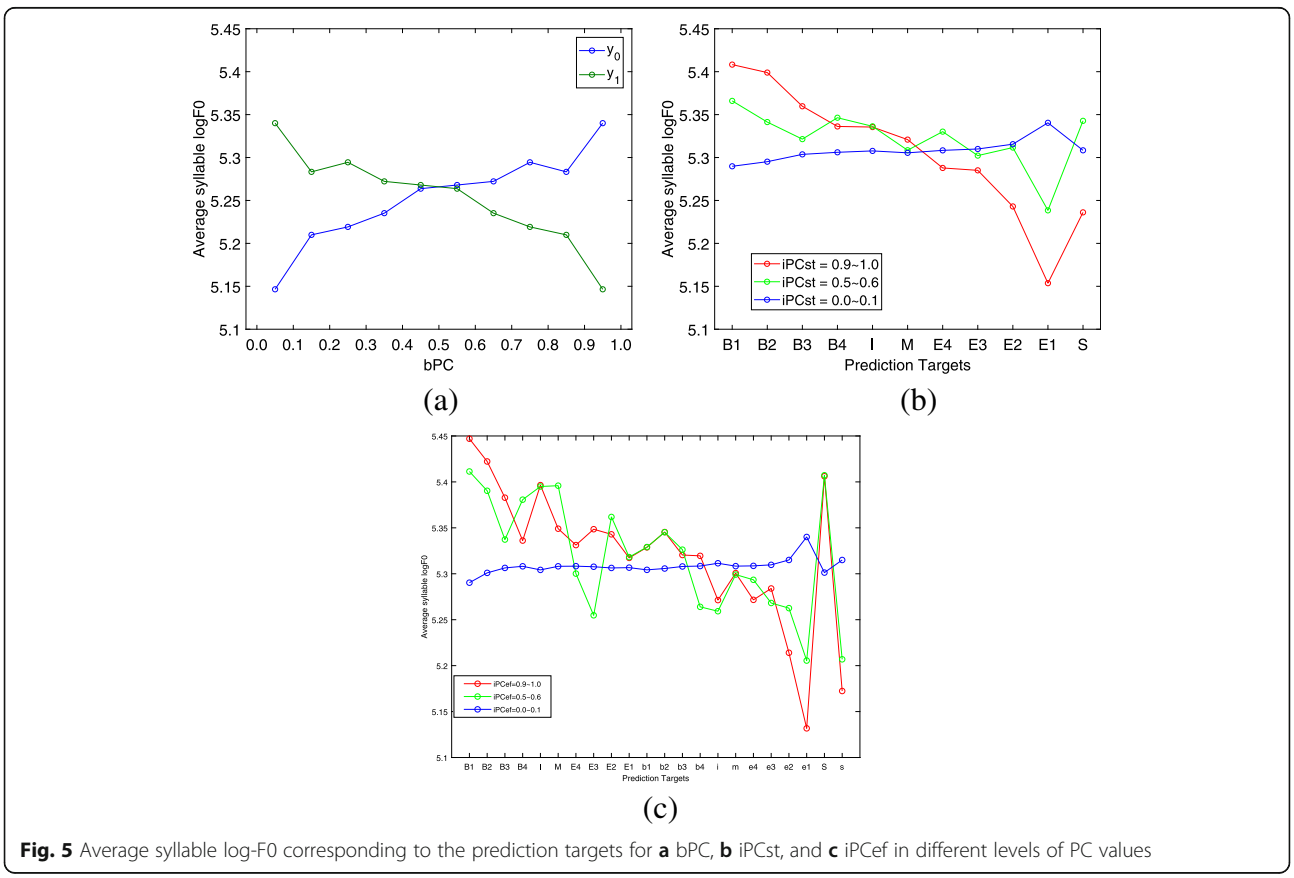

Fig. 5 Average syllable log-F0 corresponding to the prediction targets for **a** bPC, **b** iPCst, and **c** iPCef in different levels of PC values

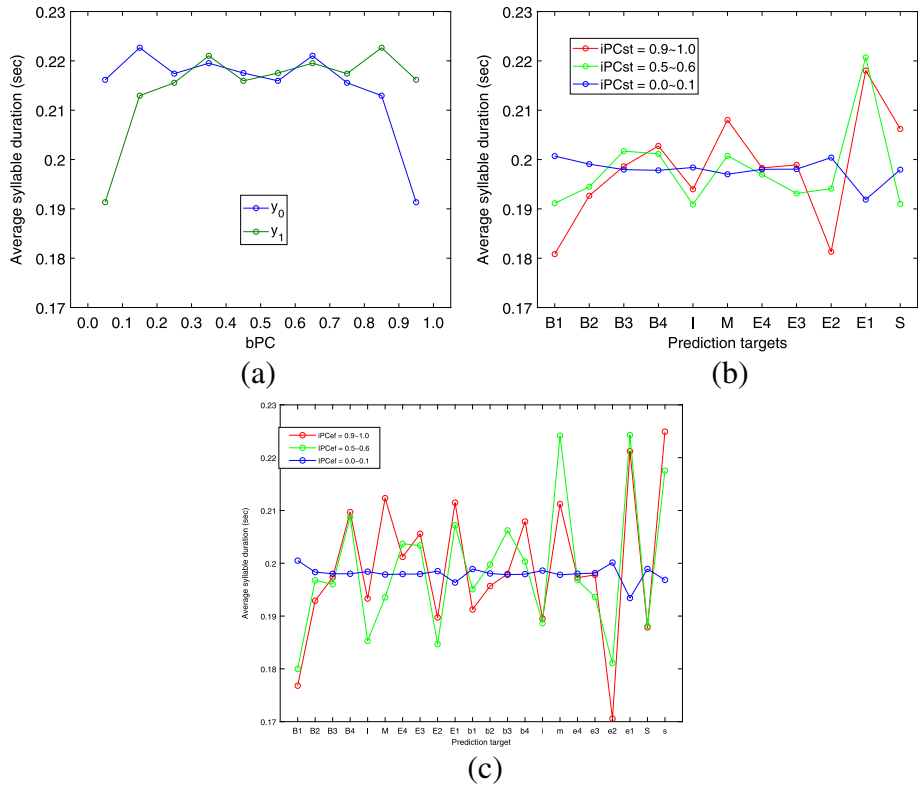

Fig. 6 Average syllable durations corresponding to the prediction targets for **a** bPC, **b** iPCst, and **c** iPCef in different levels of PC values

demonstrates that the trends in average syllable duration for the prediction targets for the first sentence-like unit and those for the second sentence-like unit are similar. Slightly longer average syllable durations for the targets "B4," "M," "b4," and "m" were observed, which was reasonable because these targets were distant from the beginning and the end of sentence-like units, resulting in a more probable prosodic break insertion. Note that the CRF-based MPM generator for iPCef predicts an enforced MPM for each sentence-like unit. Words of each sentence-like unit are therefore labeled with the prediction targets of {"B1," "B2,"…"E2," "E1," "S," "b1," "b2,"…"e2," "e1," "s"} to represent the delimitation of one sentence-like unit into two (the first and second sentence-like units). The prediction target "E1" in this case indicates that there exists an enforced inserted MPM in a sentence-like unit. The similar trends in the average syllable durations of the first and second sentence-like units indicated that the proposed iPCef could sophisticatedly model syllable duration patterns for a long sentence-like unit that may be delimited into two PPhs. As stated in Section 2.2, 40% of prosodic phrase boundaries (*B*3s) occur from non-PM interword junctures. Therefore, it was encouraging to observe these syllable duration patterns caused by the enforced insertion of MPMs through modeling of iPCef. The superiority of the proposed iPCef over the proposed iPCst and bPC for the prediction of syllable duration was partially confirmed by the prosody generation experiment presented in Section 5.3.

Figure 7 displays the pause duration corresponding to the prediction targets for (a) bPC, (b) iPCst, and (c) iPCef in different levels of PC values. Figure 7a illustrates that the average pause duration increases as bPC for MPM—that is, $\phi_{t,\ k}(\mathbf{X})$—for the prediction target y_1 increases, whereas bPC for y_0 exhibits a contrary trend. Long pause durations were found for the prediction targets of "E1" and "S" for the high and median levels of iPCst. We may conclude from these observations that higher bPC or iPCst results in longer pause durations for the predicted MPM locations. The pause duration trend in Fig. 7c for the prediction targets of the second sentence-like unit is similar to the trend of pause durations in Fig. 7b. The prediction target "E1" for the first sentence-like unit only displays a slightly longer pause duration compared with the nearby targets. The pause durations for "E1" are at the same level as the pause durations for the prediction targets that represent intermediate locations in a long sentence-like unit, that is, "B4," "M," and "m." This result indicates that the iPCef features cannot be used as salient cues for pause duration prediction, unlike the iPCst features. The objective

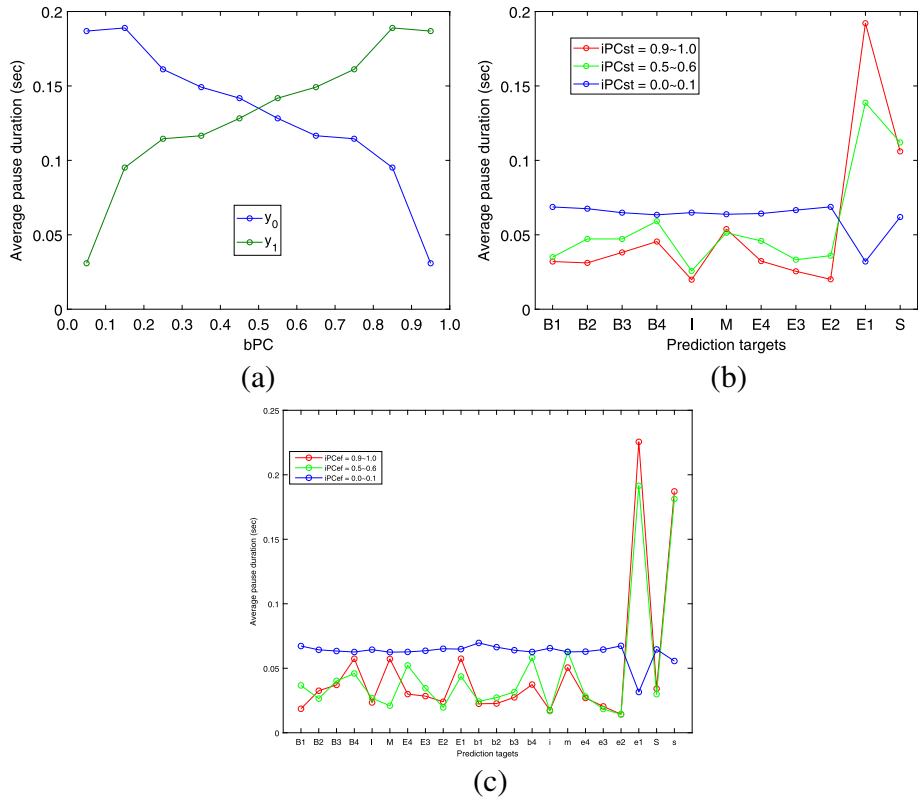

Fig. 7 Average pause durations corresponding to the prediction targets for **a** bPC, **b** iPCst, and **c** iPCef in different levels of PC values

evaluations of the prosody generation experiment presented in Section 5.3 partially confirmed this indication.

4 QC

4.1 Design of prediction targets

The prediction of QPs was also conducted using the CRF model, as described in Section 3. The target, y_k, is the k-th possible tag representing the word position in a QP. The optimal QPs, $Y_1^*, ..., Y_T^*$, can be predicted using Eq. (3), and the marginal probability of the k-th tag of the t-th word, $\phi_{t,\,k}(\mathbf{X})$, is known as the QC and is generated using Eq. (4). Two types of QCs were designed in this study: basic QC (bQC) and sentence-like unit structure QC (sQC). The bQC is generated by predicting structures of QPs, whereas sQC is generated by predicting both structures of QPs and their position in a sentence-like unit. As shown in Table 10, an eight-tag set was designed for modeling bQC. An additional tag "O" was used to represent non-QP words. Figure 8b displays a target-labeling example for the training of bQC, and the original word or PM tokens of bQC are presented in Fig. 8a. Moreover, sQC can be regarded as an improved version of bQC that uses additional tags to represent the positions of non-QP words in a sentence-like unit. These additional tags have a two-character format: xy. Here,$x \in \{B, M, F\}$ represents a word string before a QP ("B"), in-between two QPs ("M"), or following a QP ("F"), and $y \in \{b, m, e, s\}$ represents the beginning (b), intermediate (m), last (e), or a single word in a word string (s). Figure 8c displays a tag example used in the sQC training. The complete set of the prediction target for sQC is shown in Table 11.

4.2 Design of features and templates

As shown in Table 12, the features used for the prediction of QPs are similar to those ones used for the prediction of PC. The newly added PM features are used to indicate information concerning boundaries of sentence-like units. Table 13 presents the five templates for the QP prediction in this study. In template 1, we employed a three-POS context, that is, from $(t-1)$-th to $(t+1)$-th in the POS field. The word-and-POS field contains the combined features of a three-POS context and the current word (W_t). Templates 2 and 3 use a five-POS and seven-POS context,

Table 10 Tag format for the labeling of target QPs for bQC

Length in word	Tag format	Length in word	Tag format
1	S	4	B B2 M E
2	B E	5	B B2 M M E
3	B I E	6	B B2 B3 M M E

Fig. 8 **a** Original word or PM tokens, **b** example of tag labeling for the bQC training, and **c** example for the sQC training

respectively, and the combination of this context and the current word comprises the word-and-POS field. Templates 4 and 5 are identical to templates 2 and 3, respectively, in all feature fields except for the lexical word context field. We use a five-lexical word context for templates 4 and 5.

4.3 Experiment of QC generation

Only 0.69% of the ASBC text corpus contributes instances of QPs, that is, it includes only 65,723 QP token examples. To ensure that the CRF models for QC focus on predicting QPs, we only selected the sentence-like units with QPs for training and testing. The numbers of QP tokens for training and testing were 57,824 and 8439, respectively. Table 14 displays the precision and recall for bQC and sQC. Table 14 reveals that the five templates resulted in similar precision and recall. The optimal results were achieved using template 5 for bPC and template 4 for sQC. Therefore, we selected the optimal models trained using templates 4 and 5 for the following analysis and prosody generation experiments. The precision and recall for predicting bQC were approximately 60.7% and 39.0%, respectively, whereas the precision and recall for sQC were approximately 55.6% and 52.2%. These results demonstrate that modeling both structures of QPs and their position in a sentence-like unit improved the prediction of QPs. Although the precision and recall for predicting QP were considerably lower than those for predicting PC, QC enables more interesting analysis of the interplay between the prosodic–acoustic features and QC values—that is, $\phi_{t,\,k}(\mathbf{X})$.

Table 11 Tag format for labeling of target QPs for sQC

Target	Description
Pb	Presence the first word in a word string which is before a quoted phrase
Pm	Presence of the middle word in a word string which is before a quoted phrase
Pe	Presence of the end word in a word string which is before a quoted phrase
Ps	Presence of the single word in a word string which is before a quoted phrase
Mb	Presence of the first word in a word string which is between two quoted phrases
Mm	Presence of the middle word in a word string which is between two quoted phrases
Me	Presence of the end word in a word string which is between two quoted phrases
Ms	Presence of the single word in a word string which is between two quoted phrases
Fb	Presence of the first word in a word string which is after a quoted phrase
Fm	Presence of the middle word in a word string which is after a quoted phrase
Fe	Presence of the end word in a word in the word string which is after a quoted phrase
Fs	Presence of the single word in a word string which is after a quoted phrase
B/B2/B3/I/M/E/S	The same definitions as shown in Table 10

Figure 9a displays the average syllable log-F0 of the prediction targets in the three representative levels of bQC values—the high level: bQC = 0.9–1.0; median level: bQC = 0.4–0.5; and low level: bQC = 0.0–0.1. Note that the prediction targets are positioned in a forward order in a quoted phrase on the x-axis—"B," "B2," "B3," "I"/"M," and "E." We can observe a clear log-F0 declining trend for the high and median bQC levels within a QP. The average log-F0 for the single-word QPs and non-QPs are at around the medium levels. By contrast, the average syllable log-F0 is flat for the low-level iPCst. Thus, we conclude that a string of words may have log-F0 reset at the beginning of the string and then decline gradually as the string is more likely to be labeled as a QP. The log-F0 declination within a QP can also be observed in Fig. 9b for the median and high levels of sQC values. Note that some values of the average log-F0 of the prediction targets for the high-level sQC—that is, "Mb," "Mm," "Me," "B3," and "Ms"—are missing because high sQC values were not generated by the CRF-based quotation generator for these prediction targets. Additionally, log-F0 declination can also be observed for the word string preceding ("Pb," "Pm," and "Pe") and following ("Fb," "Fm," and "Fe") a QP. Therefore, we expect that the sQC features provide more informative cues for log-F0 generation than the cues provided by the bQC features. The objective evaluations of the log-F0 generation experiment presented in Section 5.3 partially meet this expectation.

Table 12 Significance of the linguistic features

Feature	Definition	Description
W_t	t-th lexical word	The smallest meaningful linguistic unit
S_t	Part of speech of t-th lexical word	Basic syntactic role of t-th lexical word; 47 categories [45]
P_t	Major PM following t-th lexical word	Major PM as boundary of sentence-like units
L_t	Length of t-th lexical word in syllable	The structure of a QP is related to word length combinations

Figure 10 displays the average syllable duration of the prediction targets in the three representative levels of bQC values. The prediction targets are also positioned in a forward order in a quoted phrase on the x-axis. The pre- or postboundary duration lengthening effect may be modeled by the trend in the QCs that is shown in Fig. 10 a and b because the average syllable duration for predicting targets "B," "B2," and "E" increased as QCs increased. Moreover, the syllable duration for the target "S," which represents a single-word QP, became longer as the corresponding QC increased. Note that some of the average syllable durations of the prediction targets for the high- and median-level QCs are missing because we did not have syllable duration samples corresponding to those cases. For the non-QP cases, significant syllable shortening and lengthening are observed for the first ("Fb") and last words ("Fe") in a word string that is followed by a QP, respectively. The objective evaluations of the syllable duration generation experiment presented in Section 5.3 indicate that these QC features can cause the RMSE of the synthesized prosody to be lower than the RMSE when the conventional linguistic features are used, thus confirming that the QC features are useful in prosody generation.

Figure 11a and b show that a word that is more likely to be at the end of QPs—that is, with the tags "E" and "S"—was more likely to be followed by a long pause. However, the other tags, except for the tag "Fe," exhibited a contrary trend. Because the sQC features provide more sophisticated structures of QPs and their contexts, we inferred that the sQC features generate pause durations with lower RMSEs than the durations generated by the bQC features.

5 Prosody generation experiments

Figure 12 displays the flowchart of the prosody generation experiments. First, the texts were fed into the text analysis modules to generate the linguistic feature sets for the following prosody generation and speech synthesis. Here, the text analysis modules included the conventional linguistic processors commonly used in an MTTS and the proposed advanced PC and QC generators. Next, the four independent MLPs were trained using the conventional linguistic feature sets and the proposed PC and QC features for predicting the syllable log-F0 contour (lf0), syllable duration (Dur), syllable energy level (Eng), and intersyllable pause duration (Pau). Subsequently, we conducted an objective test to calculate the RMSEs between the predicted and true prosodic–acoustic features. Here, the predicted prosodic–acoustic features were generated using the given different settings of linguistic features to prove the usefulness of the proposed PC and QC features. Finally, we utilized an

Table 13 Feature templates for bQC and sQC

	Template 1	Template 2	Template 3	Template 4	Template 5
Lexical word context	$\{W_{t+\tau}\}_{\tau=-1\sim+1}$, $\{W_{t-1+\tau}^{t+\tau}\}_{\tau=0,1}$, W_{t-1}^{t+1}			$\{W_{t+\tau}\}_{\tau=-2\sim+2}$, $\{W_{t-1+\tau}^{t+\tau}\}_{\tau=0,1}$, $\{W_{t-2+\tau}^{t+\tau}\}_{\tau=0,1,2}$	
POS context	$\{S_{t+\tau}\}_{\tau=-1\sim+1}$, $\{S_{t-1+\tau}^{t+\tau}\}_{\tau=0,1}$, S_{t-1}^{t+1}	$\{S_{t+\tau}\}_{\tau=-2\sim+2}$, $\{S_{t-1+\tau}^{t+\tau}\}_{\tau=0,1}$, $\{S_{t-2+\tau}^{t+\tau}\}_{\tau=0\sim2}$, $\{S_{t-2+\tau}^{t+1+\tau}\}_{\tau=0,1}$, S_{t-2}^{t+2}	$\{S_{t+\tau}\}_{\tau=-3\sim+3}$, $\{S_{t-1+\tau}^{t+\tau}\}_{\tau=0,1}$, $\{S_{t-2+\tau}^{t+\tau}\}_{\tau=0\sim2}$, $\{S_{t-3+\tau}^{t+\tau}\}_{\tau=0\sim3}$, $\{S_{t-3+\tau}^{t+1+\tau}\}_{\tau=0\sim3}$, $\{S_{t-3+\tau}^{t+2+\tau}\}_{\tau=0,1}$	The same as template 2	The same as template 3
Lexical word and POS context	$\{(W_t, S_{t+\tau})\}_{\tau=-1\sim+1}$, $\{(W_t, S_{t-1+\tau}^{t+\tau})\}_{\tau=0,1}$, (W_t, S_{t-1}^{t+1})	$\{(W_t, S_{t+\tau})\}_{\tau=-1\sim+1}$, $\{(W_t, S_{t-1+\tau}^{t+\tau})\}_{\tau=0,1}$, $\{(W_t, S_{t-2+\tau}^{t+\tau})\}_{\tau=0\sim2}$, $\{(W_t, S_{t-2+\tau}^{t+1+\tau})\}_{\tau=0\sim1}$, (W_t, S_{t-2}^{t+2})	$\{(W_t, S_{t+\tau})\}_{\tau=-1\sim+1}$, $\{(W_t, S_{t-1+\tau}^{t+\tau})\}_{\tau=0,1}$, $\{(W_t, S_{t-2+\tau}^{t+\tau})\}_{\tau=0\sim2}$, $\{(W_t, S_{t-3+\tau}^{t+\tau})\}_{\tau=0\sim3}$, $\{(W_t, S_{t-3+\tau}^{t+1+\tau})\}_{\tau=0\sim2}$	The same as template 2	The same as template 3
PM	P_t				
Lexical word length	L_t				
Previous Target	Y_{t-1}				

Table 14 QC model prediction results

	bQC		sQC	
	Precision	Recall	Precision	Recall
Template 1	0.603	0.369	0.557	0.520
Template 2	0.603	0.380	0.552	0.520
Template 3	0.597	0.389	0.548	0.518
Template 4	0.606	0.384	0.556	0.522
Template 5	0.607	0.390	0.551	0.518

HMM-based speech synthesizer and the predicted prosodic–acoustic features to generate synthesized speeches. These synthesized speeches were used to conduct subjective tests and demonstrate that the proposed PC and QC features improved the naturalness of the synthesized speeches.

5.1 Text analysis and linguistic feature sets

Figure 12 also displays the linguistic processors used and the associated linguistic features generated in this study. To perform experiments with various settings, the processors were categorized into two classes: the baseline processor and proposed advanced processor. The baseline processor performed the functions of word segmentation, POS tagging, and G2P conversion. The features generated by the baseline processor were linguistic information of phonetics, lexical words, and POSs. Because the features extracted by the baseline processor are prevalent in most MTTSs [4, 12–14, 17, 22, 24–27], we regarded them as the base linguistic features for prosody generation. In this study, we adopted the NCTU Speech Lab Traditional Chinese Parser [43, 44] as the baseline processor. The parser is an online CRF-based word tagger and generates information concerning word boundaries and the associated categories of POSs. An F-measure of 96.72% for the word segmentation and an accuracy of 94.16% for the POS tagging were reported [44]. This study employed two advanced processors—the CRF-based MPM generator and CRF-based quotation generator, described in Sections 3 and 4, respectively. These two advanced processors were cascaded after the baseline processor. The features used in the prosody generation experiments were organized into several sets according to the corresponding linguistic processors. They are summarized as follows.

5.1.1 Raw

The features in subset *Raw* can be simply extracted from raw texts. The most obvious feature in a raw text is the type of PM. PMs are the most salient feature for predicting pause break because PMs serve as delimiters for both syntax and intonation in Mandarin Chinese. Because boundaries of sentence-like units of Chinese can be identified by the type of PM, a contextual feature of syllable position in a sentence-like unit can also be extracted from the raw text. The positional features are highly related to rhythmic patterns of the syllable duration and syllable F0 contour; for example, syllables at the end of a sentence-like unit usually exhibit both syllable duration lengthening and F0 declination. Therefore, the features in the subset *Raw* include the types of PMs and syllable position in a sentence-like unit.

5.1.2 WordSeg

The features in the subset *WordSeg* are extracted after the word segmentation and include word length, syllable position in a word, and word position in a sentence-like unit. Regarding word length, the lengths of neighboring words are conventionally included because PWs are usually composed of several words with some length constraints. Most studies consider a window of five words [16, 25]; thus, the current word, two words to the left, and two words to the right, are

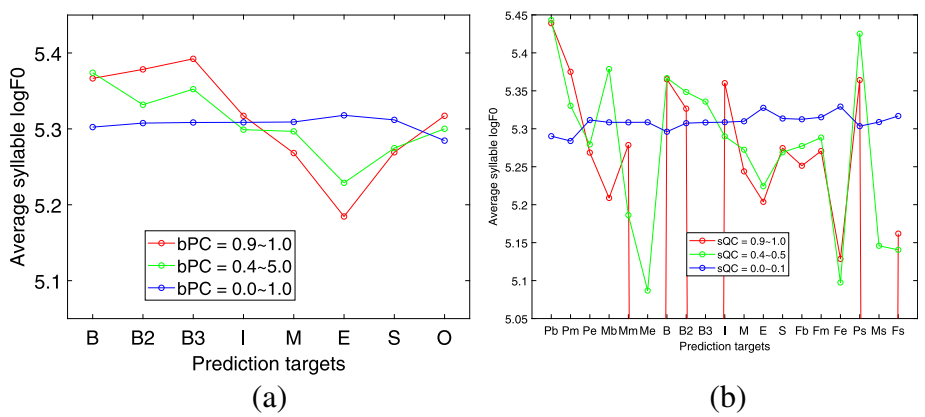

Fig. 9 Average syllable log-F0 corresponding to the prediction targets for **a** bQC and **b** sQC in different levels of QC values

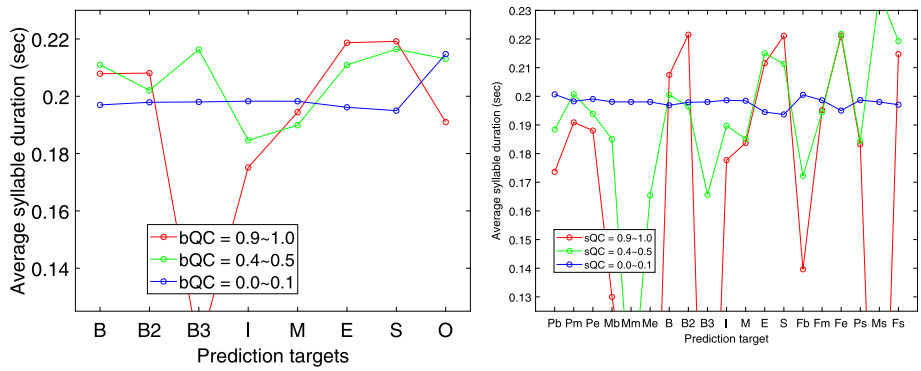

Fig. 10 Average syllable durations corresponding to the prediction targets for **a** bQC and **b** sQC in different levels of QC values

included. In this study, we extended the window to seven words; thus, the current word, three words to the left, and three words to the right were included. The positional features in this subset are also essential to syllable duration patterns. The most significant evidence for this is that syllable position in a word affects the degree of syllable duration lengthening [4].

5.1.3 WordPos

The features in the subset *WordPos* are POS tags for the associated words and are obtained after the POS-tagging process. PWs are generally composed of one to three words with some POS combinations [12, 13, 38] based on the word length constraints. Moreover, prosodic breaks and pause insertions are generally agreed to be related to some POS pairs at word junctures [12, 13, 38]. Therefore, POS and word length are the most frequently used and crucial features for predicting prosody structures from texts. In this study, we adopted a 47-POS tag set [45] that is used by the NCTU Speech Lab Traditional Chinese Parser. Similar to the usage of word length, the analysis window for POSs is set to seven words or fewer; this involves the current word, three words to the left, and three words to the right.

5.1.4 G2P

The *G2P* set comprises important features characterizing properties of Mandarin prosody: tone, base-syllable type, and initial–final type. There are five tones in Mandarin Chinese. To account for a high amount of prosodic variation resulting from contextual tones, the tones of the current, following, and previous syllables are considered for prosody generation. There are approximately 411 base-syllable types in Mandarin Chinese, and a base syllable can be further decomposed into two parts—an initial and a final. To reduce the number of features, we consider initial and final types as features to account for the information of the base-syllable type. In this study, we define 23 initial types and 40 final types. The initial and final types of the current syllable, initial type of the following syllable, and final type of the previous syllable are also considered for prosody generation.

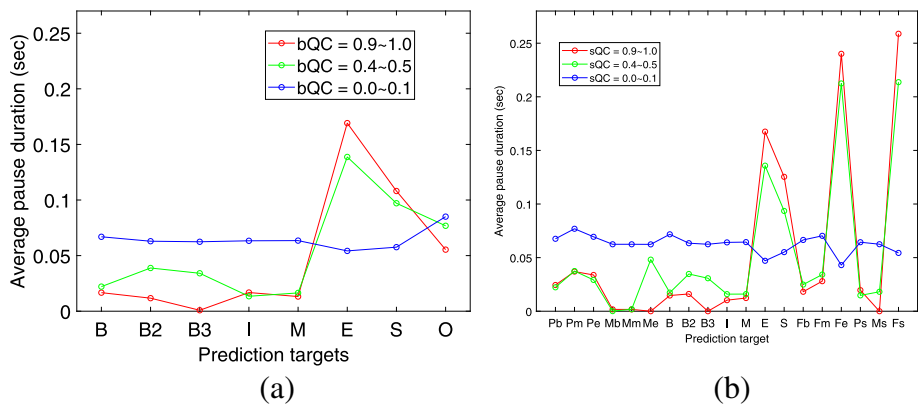

Fig. 11 Average pause durations corresponding to the prediction targets for **a** bQC and **b** sQC in different levels of QC values

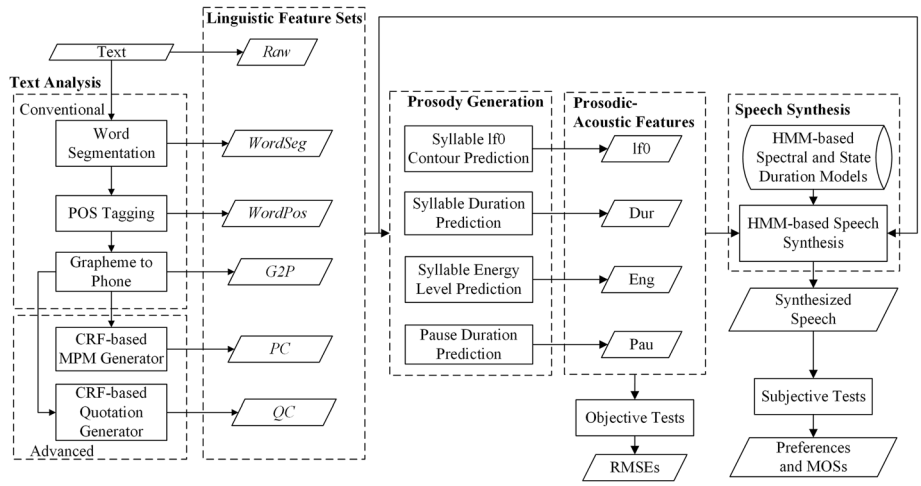

Fig. 12 Flowchart of the prosody generation experiments

5.1.5 Advanced feature set—PCs and QCs

The set comprises *PC*s and *QC*s generated using the proposed CRF-based MPM generator and the proposed CRF-based quotation generator, respectively. The subset *PC* consists of the predicted punctuation sequence given in Eq. (3)— $Y_1^*, Y_2^*, \cdots, Y_T^*$ —and the PC given in Eq. (4)—that is, $\phi_{t,\ k}(\mathbf{X})$—with target settings of bPC, iPCst, and iPCef. The subset *QC* consists of the predicted quotation label sequence, $Y_1^*, Y_2^*, \cdots, Y_T^*$, and the QC—that is,$\phi_{t,\ k}(\mathbf{X})$—with target settings of bQC and sQC.

5.2 MLP-based prosody generation

The prosody generation experiments were conducted using four independent MLPs to train prediction models for syllable log-F0 contours (lf0) represented by four-dimensional discrete orthogonal expansion coefficients [47], the syllable duration (Dur) in seconds, the syllable energy level (Eng) in dB, and the intersyllable pause duration (Pau) in seconds. The feature vectors for the input layer of the MLPs were categorized as follows for comparison: (1) baseline (*BSL*); (2) the proposed bPC, iPCst, and iPCef (*PCset*); and (3) the proposed bQC and sQC (*QCset*). *BSL* contained the most basic linguistic feature sets: *Raw*, *G2P*, *WordSeg*, and *WordPos*. It is noted that the features for predicting PC and QC include the POS (the subset of *WordPos*), word length (the subset of *WordSeg*), major PM features (the subset of *Raw*), and lexical word features. The lexical word features are not used in *BSL* because it is very hard for an MLP to capture the input information formed by high-dimensional categorical features of lexical words with limited input–output pairs (linguistic feature-prosodic feature pairs). The CRF-based MPM generator and the CRF-based quotation generator, however, potentially can be robustly trained by using a large text corpus to provide useful prosodic information that is highly correlated with major punctuations and quoted phrases. The *G2P* features and the syllable position in a word features in *WordSeg* are not included in the features for predicting PC/QC because these features are not related to the occurrence of punctuation marks. The word position in a sentence-like unit features in *WordSeg* are not included in the features for predicting PC because the word position in a sentence-like unit features themselves are the prediction targets in the PC prediction task. For the prediction of QC, the word position in a sentence-like unit features are implicitly represented by the features of major PMs following lexical words. Therefore, the results of prosody generation experiments using the feature vectors formed by PCset and QCset may show the effect of using predicted PC and QC for prosody generation.

There were 28 and 67 features in the set *Raw* and *G2P*, respectively. The feature sets bPC, iPCst, iPCef, bQC, and sQC were respectively composed of 4, 22, 44, 16, and 38 numerical features representing the marginal probabilities $\phi_{t,\ k}(\mathbf{X})$ and the predicted MPMs/quotations for some k-th target tags of PC or QC at the t-th word. The optimal numbers of nodes in the hidden layer of the MLPs and contextual analysis windows for the features of *WordSeg* or *WordPos* were tuned using the development set.

5.3 Objective tests

Table 15 displays the calculated RMSEs for the prosodic–acoustic features obtained using various linguistic feature sets. In general, the proposed *PCSet* and *QCset* improved the RMSEs with respect to *BSL*. For the lf0 prediction, the feature sets with the proposed PCs or QCs generally performed better than those without the PCs or QCs. The optimal RMSE for lf0 was achieved by using the set QC2 = *BSL3* + *sQC*. This may have been due to sQC modeling the syntactic structures of base

Table 15 RMSEs for the four prosodic–acoustic features

Feature set combinations		lf0 (logHz)	Dur (ms)	Eng (dB)	Pau (ms)
BSL	*BSL1 = Raw + G2P*	.191	43.77	3.72	*71.73*
	BSL2 = BSL1 + WordSeg	.182	39.93	3.53	*64.62*
	BSL3 = BSL2 + WordPos	.186	39.23	3.50	59.56
PCset	*PC1 = BSL3 + bPC*	.185	38.33	3.48	58.29
	PC2 = BSL3 + iPCst	.175	37.82	3.43	57.29
	PC3 = BSL3 + iPCef	.174	37.34	3.47	58.72
	PC4 = BSL2 + iPCst	.173	38.39	3.46	63.93
	PC5 = BSL2 + iPCef	.174	38.05	3.48	62.56
QCset	*QC1 = BSL3 + bQC*	.170	37.70	3.52	58.66
	QC2 = BSL3 + sQC	.169	37.83	3.52	57.95
	QC3 = BSL2 + bQC	.176	39.83	3.44	64.50
	QC4 = BSL2 + sQC	.172	39.30	3.54	63.33

phrases or word chunks that are highly correlated with the structures of PWs. The feature sets with sQC resulted in a larger RMSE improvement than for the feature sets with bQC because sQC describes not only the structures of QPs but also the structures of their contexts. The proposed iPCst and iPCef generally outperformed the proposed bPC because they could model the structures of sentences that are highly correlated with the structures of PPhs or IPs.

For the predictions of Dur and Pau, the feature sets containing *WordPos* generally outperformed those not containing *WordPos*. This partially confirms that the POS combination features are essential for prediction of the structures of PWs, PPh, and IPs. When the proposed QCs and PCs were added, further improvements were achieved because the QCs and PCs provided information that may have correlated with the structures of PWs, PPh, and IPs. The iPCef performed slightly better than the iPCst, bQC, and sQC in the prediction of Dur. This was perhaps because the iPCef models the forced insertion of an MPM in a sentence-like unit to provide more information for preboundary syllable duration lengthening. That iPCst resulted in the optimal performance in the prediction of Pau is reasonable because iPCst models the structures of sentence-like units that highly correlate with PPhs or IPs.

5.4 Subjective tests

The mean opinion score (MOS) test and preference test were performed simultaneously by 15 participants by using 15 synthesized long utterances with lengths in the range of 64 to 125 syllables (99 on average) for each prosody generation method. The feature combinations resulting in the smallest RMSEs for *BSL*, *QCset*, or *PCset*, as shown in Table 15, were selected to generate prosodic–acoustic features for speech synthesis by an HMM-based synthesizer [7–10]. Three types of proposed feature sets were compared with *BSL*: *QCset*, *PCset*, and *QCset + PCset*. As shown in Table 15, the optimal feature combination for *BSL* was the combination of *BSL2* for lf0 and *BSL3* for Dur, Eng, and Pau. The optimal combination for *QCset* was the combination of *QC2* for lf0 and Pau, *QC1* for Dur, and *QC3* for Eng. Moreover, the optimal combination for *PCset* was the combination of *PC4* for lf0, *PC3* for Dur, and *PC2* for Eng and Pau. The feature sets for *QCset + PCset* were *QC2* for lf0, *PC3* for Dur, and *PC2* for Eng and Pau. Before listening to the utterances synthesized using *BSL* and those using the proposed method, the participants were asked to listen to the true utterances in the test speech corpus corresponding to the synthesized speeches for reference. The order of the synthesized utterances in the preference test was randomly set. Table 16 reveals that the proposed *QCset*, *PCset*, and *QCset + PCset* generally yielded slightly more natural speech than that yielded by *BSL*. The synthesized utterances with prosody generated using *QCset + PCset* achieved the most significant difference in MOS from *BSL*. These results again confirm the usefulness of the proposed PC and QC features.

6 Conclusions and future work

This paper proposed two fully automatic machine-extracted linguistic features from an unlimited-text input

Table 16 Preferences (%) and MOSs (numbers in brackets ± standard deviation) for the two subjective tests

Pairs	The proposed	BSL	No prefer.
QCset vs. *BSL*	34% (3.45 ± 0.42)	25% (3.40 ± 0.45)	41%
PCset vs. *BSL*	37% (3.55 ± 0.41)	21% (3.34 ± 0.48)	42%
QCset + PCset vs. *BSL*	38% (3.57 ± 0.41)	22% (3.29 ± 0.48)	40%

for Mandarin prosody generation. The first feature is PC, which measures the likelihood that an MPM can be inserted at a word boundary. The second feature is QC, which measures the likelihood that a word string is quoted as a meaningful or emphasized unit in text. The rationale of these proposed punctuation-generation-inspired linguistic features was illustrated by analyses of the relationship of the prosodic structures and PM types with the structures of QPs. The usefulness of the proposed PC and QC features in Mandarin prosody generation was proved by both objective and subjective tests. The proposed features improved the performance of Mandarin prosody generation. The PC and QC features have the following advantages over the conventional linguistic features:

1. The features for predicting the PC include the POS (the subset of WordPos), word length (the subset of WordSeg), and lexical word features while the features for predicting the QC include the POS, word length, MPM (the subset of Raw), and lexical word features. The lexical word features, which inherently provide much richer linguistic information than word length and POS, however, are not used in the baseline prosody generation system because it is hard for an MLP to capture the input information formed by high-dimensional categorical features of lexical words with limited input-output pairs (linguistic feature-prosodic feature pairs). The CRF-based MPM generator and the CRF-based quotation generator, however, potentially can be robustly trained by using a large text corpus to provide useful prosodic information that is highly correlated with major punctuations and quoted phrases.
2. The PC and QC features are numerical features in a range of [0 1] while most of the conventional linguistic features are categorical and represented by the values of 0 or 1 in high-dimensional vectors. The numerical features inherently are more suitable than the categorical features in regression tasks. The relationship between the numerical prosodic features and the values of the PC and QC are easy to be analyzed as shown in Section 3.3 (Experiment of PC generation) and Section 4.3 (Experiment of QC generation). These analyses showed that the proposed PC and QC values are correlated with the prosodic features.
3. Most of the QPs were found to be single- to four-word QPs. Single-word QPs are usually emphasized nouns or verbs. Two- to four-word QPs are mostly base phrases such as word strings (or word chunks). The QPs longer than four words are mostly sentence-like units. Since the QC is a measure to show how likely a word string to be quoted as a word chunk or a base phrase and word chunks and base phases are larger and more meaningful linguistic units than words are, prosody generation with the QC is inherently advantageous over prosody generation with the baseline features, i.e., word lengths and POSs.
4. In this study, the PC not only models the likelihoods of word boundaries to be inserted with major punctuation marks but also models structures of sentence-like units. These sentence-like units can be sentences, phrases, or clauses, and they are larger linguistic units than words are. These properties infer that the PC features are inherently more powerful than the conventional word length and POS features in prosody generation.

In summary, the PC and QC features provide richer syntactic and partially semantic information than the conventional baseline linguistic features to prosody generation.

It is known that prosody is affected by linguistic, para-linguistic, and non-linguistic features [48]. Linguistic features include lexical, syntactic, semantic, and pragmatic features. Para-linguistic features include intentional, attitudinal, and stylistic features. Non-linguistic features include physical and emotional features. It is generally agreed that speaker styles are associated with para-linguistic (stylistic) and non-linguistic (physical) features while utterance styles (spoken words) are not only affected by para-linguistic and non-linguistic features but also biased by text content (represented by linguistic features). The proposed PC and QC features can provide syntactic and partially semantic information, i.e., linguistic features, to prosody. In this study, we utilize the conventional linguistic features (baseline features) and the proposed PC and QC to predict prosody of read speech style. To predict prosody of other speaker or utterance styles, recording speech utterances of various speakers and utterance styles are necessary while retraining of new text may be unnecessary because PC and QC features are trained by large text corpus with various text styles. Investigating the use of PC and QC for prosody generation of various speaker and utterance styles is beyond the scope of this study, but is worth doing in the future.

With the fast growth of deep learning technologies, it will be worthwhile to incorporate CRF-based punctuation generation models into neural network models; for example, the long short-term memory recurrent neural network [49]. Neural-network-based punctuation models can be easily integrated with the related neural-network-based prosody generator or speech synthesizer in the training phase. Under this integrated framework, the transfer learning technique can also be applied [50] to enable a neural network to learn prosody generation based on a different neural network that generates punctuations.

Abbreviations

ASBC: Academia Sinica Balanced Corpus of Modern Chinese; BG/PG: Breathe group or prosodic phrase group; bPC: Basic punctuation confidence; bQC: Basic quotation confidence; BSL: Baseline linguistic features; CART: Classification and regression tree; CKIP: Chinese knowledge and information processing; CRF: Conditional random field; Dur: Syllable duration; Eng: Syllable energy level; G2P: Grapheme-to-phone; HMM: Hidden Markov model; IP: Intonation phrase; iPC: Improved punctuation confidence; iPCef: Improved punctuation confidence with enforced major punctuation mark insertion; iPCst: Improved punctuation confidence with structure of sentence-like unit; lf0: Log-F0 contour; MLP: Multilayer perceptron; MPM: major punctuation mark; MTTS: Mandarin text-to-speech system; NPB: Nonpause break; Pau: Intersyllable pause duration; PC: Punctuation confidence; PCset: The linguistic feature set that contains bPC, iPCst, and iPCef; PLM: Prosody labeling and modeling; PM: Punctuation mark; POS: Part-of-speech; PPh: Prosodic phrase; PW: Prosodic word; QC: Quotation confidence; QCset: The linguistic feature set that contains bQC and sQC; *Raw*: Linguistic features extracted simply from raw text; RMSE: Root-mean-square error; sQC: Sentence-like unit structure quotation confidence; SYL: Syllable; TTS: Text-to-speech system; *WordPos*: Linguistic features extracted after part-of-speech tagging; *WordSeg*: Linguistic features extracted after word segmentation

Acknowledgements

The authors deeply thank Prof. Yih-Ru Wang of NCTU, Hsinchu, Taiwan, for providing the NCTU Speech Lab Traditional Chinese Parser. The authors also wish to thank Academia Sinica, Taiwan, for providing the Treebank Corpus, Academia Sinica Balanced Corpus of Modern Chinese 4.0, and online CKIP Parser. This manuscript was edited by Wallace Academic Editing.

Funding

This work was primarily supported by a grant from Chunghwa Telecom under contract no. TL-102-8202. This work was also supported in part by the Ministry of Science and Technology (MOST) of Taiwan under contract no. MOST-106-2221-E-305-010.

Availability of data and materials

Please contact author for data requests.

Authors' contributions

C-YC wrote the paper and conceived and designed the experiments. Y-PH and H-YY performed the experiments. Y-PH analyzed the data, and I-BL and C-MP contributed reagents, materials, and analysis tools. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Communication Engineering, National Taipei University, 151, University Rd., San Shia District, New Taipei City 23741, Taiwan.
[2]Telecommunication Laboratories, Chunghwa Telecom, No.99, Dianyan Rd., Yangmei District, Taoyuan City 32661, Taiwan.

Received: 28 March 2018 Accepted: 5 February 2019
Published online: 21 February 2019

References

1. Li, A. J., Zu, Y. Q., & Li, Z. Q. (1999). A national database design and prosodic labeling for speech synthesis. In *Proc. Oriental Co-ordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA) Workshop, Taipei, Taiwan* (pp. 13–16).
2. Li, A. J., & Lin, M. C. (2000). Speech corpus of Chinese discourse and the phonetic research. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP), Beijing, China* (pp. 13–18).
3. Cao, J. F. (2000). Rhythm of spoken Chinese—Linguistic and paralinguistic evidences. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP), Beijing, China* (pp. 357–360).
4. Chen, S. H., Hwang, S. H., & Wang, Y. R. (1998). An RNN-based prosodic information synthesizer for Mandarin text-to-speech. *IEEE Trans. Speech Audio Process., 6*, 226–239.
5. Chen, S. H., Lai, W. H., & Wang, Y. R. (2005). A statistics-based pitch contour model for Mandarin speech. *J. Acoust. Soc. Am., 117*, 908–925.
6. Chen, S. H., Lai, W. H., & Wang, Y. R. (2003). A new duration modeling approach for Mandarin speech. *IEEE Trans. Speech Audio Process., 11*, 308–320.
7. Tokuda, K., Yoshimur, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Istanbul, Turkey* (pp. 1315–1318).
8. Yoshimura, T. (2002). *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems*. Nagoya: Dissertation, Nagoya Institute of Technology.
9. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., & Tokuda, K. (2007). The HMM-based speech synthesis system version 2.0. In *Proceedings of the Sixth ISCA Workshop on Speech Synthesis (SSW6), Bonn, Germany* (pp. 294–299).
10. The HTS working group, HTS-2.3 source code, and demonstrations. http://hts.sp.nitech.ac.jp/?Download. Accessed 26 Jan 2018.
11. Ostendorf, M., & Veilleux, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Comput. Linguist., 20*, 27–52.
12. Peng, H. J., Chen, C. C., Tseng, C. Y., & Chen, K. J. (2004). Predicting prosodic words from lexical words—a first step towards predicting prosody from text. In *Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong* (pp. 173–176).
13. Chu, M., & Qian, Y. (2001). Locating boundaries for prosodic constituents in unrestricted mandarin texts. *Comput. Linguist. Chin. Lang. Process., 6*, 61–82.
14. Xu, D. W., Wang, H. F., Li, G. H., & Kagoshima, T. (2006). Parsing hierarchical prosodic structure for Mandarin speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toulouse, France* (pp. 14–19).
15. Black, A. W., & Taylor, P. (1997). Assigning phrase breaks from part-of-speech sequences. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Rhodes, Greece* (pp. 995–998).
16. Sheng, Z., Tao, J. H., & Jiang, D. L. (2003). Chinese prosodic phrasing with extended features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong* (pp. 492–495).
17. Li, J. F., Hu, G. P., & Wang, R. H. (2004). Chinese prosody phrase break prediction based on maximum entropy model. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP), Jeju Island, Korea* (pp. 729–732).
18. Riedi, M. (1995). A neural-network-based model of segmental duration for speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Madrid, Spain* (pp. 599–602).
19. Sagisaka, Y. (1990). On the prediction of global F0 shape for Japanese text-to-speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Albuquerque, New Mexico, USA* (pp. 325–328).
20. Traber, C. (1992). In G. Bailly & C. Benoit (Eds.), *Talking Machines: Theories, Models and Designs*. Amsterdam: Elsevier.
21. Scordilis, M. S., & Gowdy, J. N. (1989). Neural network based generation of fundamental frequency contours. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Glasgow, Scotland* (pp. 219–222).
22. GP Chen, G. B., Liu, Q. F., & Wang, R. H. (2004). A superposed prosodic model for Chinese text-to-speech synthesis. In *Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong* (pp. 117–120).
23. Bailly, G., & Holm, B. (2006). SFC: a trainable prosodic model. *Speech Comm., 46*, 348–364.
24. Wen, M. M., Wang, M. M., Hirose, K., & Minematsu, N. (2010). Improved Mandarin segmental duration prediction with automatically extracted syntax features. In *Proceedings of the 10th IEEE International Conference on Signal Processing (ICSP), Beijing, China* (pp. 621–624).

25. CC Hsia, C. H. W., & Wu, J. Y. (2010). Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis. *IEEE Trans. Audio Speech Lang. Process., 18*(8), 1994–2003.
26. Wang, M. M., Wen, M. M., Hirose, K., & Minematsu, N. (2010). Improved Generation of Prosodic Features in HMM-based Speech Synthesis. In *Proceedings of the Seventh ISCA Workshop on Speech Synthesis (SSW7), Kyoto, Japan* (pp. 359–336).
27. Wang, M. M., Wen, M. M., Hirose, K., & Minematsu, N. (2010). Improved generation of fundamental frequency in HMM-based speech synthesis using generation process model. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Makuhari, Chiba, Japan* (pp. 2166–2169).
28. Chiang, C. Y., Wang, Y. R., & Chen, S. H. (2012). Punctuation generation inspired linguistic features for Mandarin prosodic boundary prediction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan* (pp. 4597–4600).
29. Hung, Y. P., Yeh, H. Y., Liao, I. B., Pan, C. M., & Chiang, C. Y. (2014). An investigation on linguistic features for Mandarin prosody generation. In *Proceedings of the 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), Phuket, Thailand* (pp. 1–5).
30. Chiang, C. Y., Hung, Y. P., Liou, G. T., & Wang, Y. R. (2016). Improvements on punctuation generation inspired linguistic features for Mandarin prosody generation. In *Proceedings of the 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China* (pp. 1–5).
31. Hung, Y. P. (2015). *Punctuation Generation Inspired Linguistic Features for Mandarin Prosody Generation*. New Taipei City: Master, National Taipei University.
32. YQ Guo, H. F., Wang, J. V., & Genabith, A. (2010). Linguistically inspired statistical model for Chinese punctuation generation. *ACM Trans. Asian Lang. Process., 9*(2), 6.
33. Tseng, C. Y. (2003). Mandarin speech prosody: issues, pitfalls and directions. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland* (pp. 2341–2344).
34. http://www.aclclp.org.tw/use_asbc.php. Accessed 15 Mar 2018.
35. Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williamstown, MA, USA* (pp. 282–289).
36. CRF++: Yet Another CRF toolkit. https://taku910.github.io/crfpp/. Accessed on 26 Jan 2018.
37. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: a standard for labeling English prosody. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP), Banff, Alberta, Canada* (pp. 867–870).
38. Taylor, P. A. (1998). The tilt intonation model. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP), Sydney, Australia* (pp. 1383–1386).
39. Li, A. J. (2002). Chinese prosody and prosodic labeling of spontaneous speech. In *Proceedings of the ISCA International Conference on Speech Prosody (Speech Prosody), Aix-en-Provence, France* (pp. 39–46).
40. Chiang, C. Y., Chen, S. H., Yu, H. M., & Wang, Y. R. (2009). Unsupervised joint prosody labeling and modeling for Mandarin speech. *J. Acoust. Soc. Amer., 125*(2), 1164–1183.
41. Chiang, C. Y., Chen, S. H., & Wang, Y. R. (2009). Advanced unsupervised joint prosody labeling and modeling for Mandarin speech and its application to prosody generation for TTS. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brighton, UK* (pp. 504–507).
42. SH Chen, J. H., Yang, C. Y., Chiang, M. C., Liu, Y. R., & Wang, A. (2012). New prosody-assisted mandarin ASR system. *IEEE Trans. Audio Speech Lang. Process., 20*(6), 1669–1684.
43. The NCTU Speech Lab Traditional Chinese Parser. http://parser.speech.cm.nctu.edu.tw/ Accessed on 26 Jan 2018.
44. Lin, A. H., Wang, Y. R., & Chen, S. H. (2013). In *Proceedings of the 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), Gurgaon, India* (pp. 1–5).
45. Chen, K. J., & Huang, C. R. (1993). Part of speech (POS) analysis on Chinese language. In *CKIP Technical Report No.93–05; Institute of Information Science, Academia Sinica: Taiwan, R.O.C.*
46. Chiang, C. Y., Yu, H. M., Wang, Y. R., & Chen, S. H. (2008). Exploration of high-level prosodic patterns for continuous Mandarin speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, USA (pp. 4381–4384).*
47. Chen, S. H., & Wang, Y. R. (1990). Vector quantization of pitch information in Mandarin speech. *IEEE Trans. Commun., 38*(9), 1317–1320.
48. Fujisaki, H. (2003). Prosody, information, and modeling: with emphasis on tonal features of speech. In *Proceedings of the Workshop on Spoken Language Processing, Mumbai, India* (pp. 5–14).
49. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput., 9*(8), 1735–1780.
50. Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In *Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, Bellevue, Washington, USA* (pp. 17–36).