Article

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], Chen-Ming Pan[2]

[1] Dept. of Communication Engineering, National Taipei University, Taiwan; cychiang@mail.ntpu.edu.tw (C.-Y. C.); c2333391@gmail.com (Y.-P. H.); henryyeh034@gmail.com (H.-Y. Y.)

[2] Telecommunication Laboratories, Chunghwa Telecom, Taoyuan Taiwan; snet@cht.com.tw (I.-B. L.); chenming@cht.com.tw (C.-M. P.)

* Correspondence: cychiang@mail.ntpu.edu.tw; Tel.: +886-2-8671-1111#67732

Abstract: This paper proposes two fully-automatic machine-extracted linguistic features from an unlimited text input for Mandarin prosody generation. One is the punctuation confidence (PC) which measures the likelihood of inserting a major punctuation mark (PM) at a word boundary. Another is the quotation confidence (QC) which measures the likelihood of a word string to be quoted as a meaningful or emphasized unit in text. Because a major PM in a text is highly correlated with a prosodic break, and a quoted word string plays an important role in human language understanding, the two features potentially could provide useful information for prosody generation. The idea is first realized by employing conditional random field (CRF)-based models to predict major PMs, quoted word string locations, and their associated confidences, i.e., the PC and the QC, for each word boundary. Then, the predicted punctuations and their confidences are combined with traditional contextual linguistic features to predict prosodic-acoustic features. Both objective and subjective tests showed that the prosody generation with the proposed linguistic features performed better than the one without the proposed features. So, the proposed PC and QC are promising features for Mandarin prosody generation.

Keywords: Mandarin; prosody generation; linguistic feature; break prediction; text-to-speech; punctuation confidence

---

1. Introduction

Prosody generation plays a crucial role in a text-to-speech system (TTS). We can regard prosody generation as a function mapping from linguistic feature to prosodic structures or prosodic-acoustic feature. In a practical implementation of an unlimited-text Mandarin text-to-speech system (MTTS), availability and reliability of linguistic features are highly dependent on performances of text analyzers. A basic text analyzer includes Chinese word segmenter, grapheme-to-phone (G2P) converter and part of speech (POS) tagger. Prosodic structures are abstract descriptions of speech prosody, and usually categorically represented by prosodic break tags, such as non-break, minor/major break, and so forth. A commonly agreed Mandarin prosody hierarchy is a four-layer prosodic structure with, from the lowest layer to the highest one, syllable (SYL) layer, prosodic word (PW) layer, intermediate phrase (or prosodic phrase, PPh) layer, and intonation phrase (IP) layer, which are demarked respectively by non-break, minor break, major break, and utterance boundary [1-3]. Prosodic-acoustic features are prosodic information numerically represented by values or vectors of log-F0 contour, duration, and energy of any linguistic domain, e.g., a phone, a syllable, an initial/final, or a word. Representative prosodic-acoustic features for Mandarin speech are syllable log-F0 contour, syllable duration, pause duration, and syllable energy level [4-6]. Besides, in the most popular speech synthesis method - HMM-based synthesis [7-10], prosodic-acoustic features are modeled in HMM state level, i.e., state duration, state logF0 value, and energy contour enclosed by spectral parameters.

No matter what the target (prosodic structure or prosodic-acoustic feature) of prosody generation is, studies of prosody generation focused on the following two issues: (1) design or utilization of prediction model, and (2) utilization of features. In the first issue, popular prediction methods for generating prosodic structure are hierarchical stochastic model [11], N-gram model [12], classification and regression tree (CART) [13,14], bottom-up/sifting hierarchical CART [13], Markov model [15], artificial neural networks [16], maximum entropy model [17], etc. As for generating prosodic-acoustic features, popular pattern recognition tools were utilized, such as multi-layer perceptron (MLP) [18-23], recurrent neural network (RNN) [4], CART [7-10,24], and decision tree plus hidden Markov model with multi-space distribution modeling of F0 contour [7-10], and so forth. In the second issue, conventional linguistic features, such as POS, word length, sentence length, position in a sentence, and so forth, are widely used in many existing MTTSs [4,12-14,17,22,24-27]. Some studies further improved the accuracy of prosodic structure prediction or prosodic-acoustic prediction by incorporating higher-level syntactic features, such as word chunk [16] and syntactic tree [16,26,27]. On the other hand, statistical linguistic features - connective degree [14], punctuation confidence (PC) [28-31] and quotation confidence (QC) [30,31] were proposed to neglect complex syntactic tree parsing and manual word chunking that causes impracticality in constructing an unlimited-text MTTS.

This paper focuses on the second issue to extend and elaborate on our previous works in the PC [28-31] and QC [30,31] features. More substantial analysis and modeling details are provided in this paper to give readers an insight into the proposed PC and QC features. The proposed PC and QC features are motivated by automatic Chinese punctuation generation [32] and linguistic characteristic of Chinese punctuation system [33]. The PC measures the likelihood of inserting a major punctuation mark (MPM) at a word boundary while the QC measures the likelihood of a word string quoted by brackets to emphasize the meaning of the quoted word strings. In [32], a maximum entropy (ME)-based automatic Chinese punctuation generation method was proposed to insert 16 types of punctuation mark (PM) to an un-punctuated text by using features of word and lexical-functional grammar features. The results in [32] showed that the punctuation generation model could generate alternative/acceptable insertions, deletions or substitutions of PMs. This phenomenon was also observed in a human punctuation experiment reported by Tseng [33] in which alternative punctuation strategies were found among different native Mandarin Chinese speakers. These observations reflect the fact that Chinese PMs serve as a loose reference to both syntactic structure and semantic domain, and therefore native Chinese writers would freely utilize PMs to delimit written Chinese into various linguistic elements, such as phrases and clauses, to clearly express the meaning of a text. Furthermore, punctuation generation of a speaker when reading written Chinese would reflect his/her prosodic phrasing strategy because pause break is highly correlated with some MPMs, such as period, comma, exclamatory mark, question mark, semicolon, and colon. Therefore, an automatic punctuation generation model predicting MPMs trained from a large text corpus can learn punctuation strategies for MPMs from various text contributors, to provide useful cues for both prosodic break [28,31] and prosodic-acoustic feature predictions [29-31].

On the other hand, a word strings sandwiched by brackets or quotes have essential or unique meanings in sentences. By our analysis on a large text corpus - the Academia Sinica Balanced Corpus of Modern Chinese (ASBC) V.4.0 [34] with 9,454,734 words (or 31,126 paragraphs), we found that the functions of the quoted word strings can be classified into several cases: (1) to add supplementary information to the proceeding words, (2) to represent the name of a particular person, place or institution, (3) to emphasize the meaning of a word string, or (4) to indicate a new derived compound word or a word chunk which compose a complex meaning. In the cases of (3) and (4), the quoted word strings which are called quoted phrases in this paper, from small to large linguistic units, may form new-derived words, compound words, base phrases, word chunks, syntactic phrases, and even sentences. The mentioned-above linguistic units are usually larger than common words in size, containing more complex meanings than a word, or even generating new meanings, and maybe constituting a higher-level unit in syntax than POSs of words. Since a quoted phrase exhibits richer linguistic information than just words, it plays a crucial role in human language understanding when reading a text. Moreover, it is generally agreed that a speaker can generate good prosody if he/she understands the meanings of a text. Thus, adding quotations to plain Chinese texts and then regarding the added brackets as linguistic features may help naturalness of machine-generated prosody. Note that in written Chinese, the use of quotations by adding brackets depends on writing styles or habits of text contributors. Unlimited Chinese input texts may already contain some brackets to exhibit the four functions illustrated previously. However, the remaining un-quoted words may also be emphasized, be regarded as larger syntactic units if they share similar contextual POS or word structures with the quoted phrases. For the case that Chinese texts contain no quotations, if quotations can be labeled with brackets by a machine automatically given the word and POS information, the features associated with the labeled brackets could provide richer linguistic information to enhance the performances of prosodic-acoustic feature predictions.

To realize the ideas of automatic MPM and quotation predictions, we construct two types of the conditional random field [35,36] (CRF)-based automatic punctuation generation models: the CRF-based MPM generation model and the CRF-based quotation generation model. The CRF-based MPM generation model predicts MPMs and generates the associated confidence measures, referred to as the punctuation confidence (PC), from major PM-removed word/POS sequences. The PC can be regarded as a statistical linguistic feature to measure the likelihood of inserting an MPM into a text. It is reasonable to hypothesize that word junctures which are more likely to be inserted with MPMs in text, are more likely to be inserted with pause breaks in an utterance. We could, therefore, expect that the utilization of the PC in prosody generation may improve the performance of prosodic-acoustic feature generation. The CRF-based quotation generation model predicts the structures of quoted word string (i.e., QP) from bracket-removed word/POS sequences and generates the associated confidence, referred to as the quotation confidence (QC). The QC can also be taken as a statistical linguistic feature to measure the likelihood of word strings being quoted by a left bracket and a right bracket. Since words in the brackets are closely related to constitute meanings, it is reasonable to assume that less prosodic breaks are inserted within a quoted text, and quoted text may be emphasized with some variations in prosodic-acoustic features. We therefore also expect the use of QC may also assist in prosody generation.

To evaluate the usefulness of the proposed PC and QC in Mandarin prosody generation, the experiments of prosodic-acoustic feature prediction were conducted, and the corresponding objective and subjective tests were then evaluated. The experimental database is a read Mandarin speech corpus – the Treebank speech corpus, containing 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 of 3,898 syllables. The corpus used for training the CRF-based punctuation generator was the Academia Sinica Balanced Corpus of Modern Chinese (ASBC) V.4.0 [34] (denoted as the ASBC text corpus thereafter). In the prosodic-acoustic feature prediction, the proposed linguistic features combined with conventional linguistic feature were taken as input to directly predict four prosodic-acoustic features of syllable log-F0 contour, syllable duration, syllable energy level, and inter-syllable pause duration. Objective tests were evaluated by root-mean-square error (RMSE). Subjective tests were then evaluated with speech-synthesized utterances with the predicted prosodic-acoustic features.

Several advantages of the approach can be found. First, the PC and the QC can be easily obtained from features of word/POS sequence which can be robustly obtained by current word segmentation and POS tagging technologies without using complicated statistical syntactic parsing. This makes the proposed approach more suitable for practical on-line unlimited TTS. Second, as being trained using a large text corpus, the CRF-based punctuation generation models can 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 corpus used to train the CRF-based punctuation generator can

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2018 doi:10.20944/preprints201802.0108.v1

be considerably larger. Therefore, we can expect that the PC and the QC would be more robust than syntactic features derived from an automatic syntactic parser.

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

- **Section 2: Analysis of Punctuations**

  We show the relationship between punctuations and prosodic structures via analyzing the Treebank speech corpus which is labeled with prosodic break tags. This analysis motives the proposed PC. This section will also analyze the quoted phrases observed in the ASBC text corpus, finding the 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 will be trained given with the ASBC text corpus. The precisions and recalls of the MPM insertions are examined on the test set of the ASBC text corpus. The feasibility of the proposed PC in prosody generation will be examined by analysis the relationship between the prosodic-acoustic features of the training set of the Treebank speech corpus and the associated PC generated by the CRF-based MPM generation model.
- **Section 4: Construction of the CRF-based Quotation Generation Model**

  The model will also be trained and examined on the ASBC text corpus. The feasibility of the QC for the prosody generation is also examined on the Treebank speech corpus.
- **Section 5: Prosody Generation Experiments**

  The prosody generation experiments will be conducted on the Treebank speech corpus. The proposed PC and QC features generated by the proposed automatic punctuation generation models with the texts of the Treebank text corpus are combined with the conventional linguistic features to predict the prosodic-acoustic features of 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 Works**

2. Analysis of Punctuations

Because prosodic-acoustic features are highly dependent on Mandarin prosodic structure and the prosodic structure are categorically represented by a finite set of prosodic break tags, it is easier to analyze the relationship between prosodic break types and PMs than to analyze the relationship between numerical prosodic-acoustic features and PMs. This section, therefore, analyzes the relationship between Chinese PMs and Mandarin prosodic structure. In the following subsections, the analyses will disclosure the motivations and the rationality of the proposed PC and QC features. The prosody labeling system for illustrating prosodic structures of utterances used in this study will be introduced in Section 2.1. The relationship between the labeled prosodic break types and PM types will be discussed in Section 2.2. Section 2.3 will experiment to let native Mandarin speakers insert MPMs manually given with PM-removed texts excerpted from the Treebank speech corpus. The relationships between the human-labeled MPMs by the native Mandarin speakers and the associated prosodic break types are analyzed, showing some evidence for the proposed PC. Section 2.4 will analyze the quoted phrases observed in the ASBC text corpus, finding the possible QC candidates for the training of the CRF-based quotation generation model.

2.1. Prosody Label System

Famous prosody labeling systems are the ToBI [37], TILT [38], and C-ToBI [39]. The mentioned-above prosody labeling systems require human labeling with linguistic expertise. To leverage the intensive human labor and to increase consistency of prosody labeling, Chiang et at. [40,41] proposed an unsupervised joint prosody labeling and modeling (PLM) method to construct a speaker-dependent statistical hierarchical prosodic model (HPM) and to label prosody tags for Mandarin speech. The PLM method was then successfully applied to construct a speaker-independent HPM to assist in a large vocabulary speech recognition task [42]. Hence, in this study, to avoid intensive human labeling and inconsistent labeling results, the corpus was labeled with seven break types by the PLM method [40,41] proposed by Chiang et al.. As shown in Figure 1, the seven break types, i.e. {*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4}, delimit an utterance into four types of prosodic units, namely syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breathe group/prosodic phrase group (BG/PG).

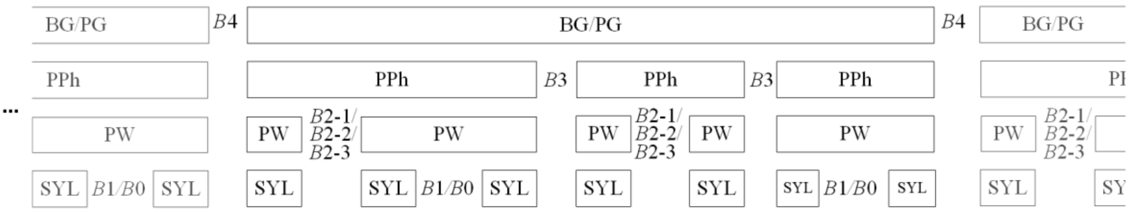


Figure 1. The prosody-hierarchy model of Mandarin speech used in this study [42]

In the labeling system, each defined break type is characterized by its specific juncture prosodic-acoustic features. *B*4 is defined as a major break accompanying long pause and apparent F0 reset across adjacent syllables; *B*3 is a major break with medium pause and medium F0 reset; *B*0 and *B*1 represent respectively non-breaks of tightly-coupling syllable juncture and normal syllable boundary, within a PW, which have no identifiable pauses between SYLs; and *B*2 is a minor break with three variants: F0 reset (*B*2-1), short pause (*B*2-2), or pre-boundary syllable duration lengthening (*B*2-3).

Among various types of prosodic-acoustic features, pause duration is the most salient cue to specify boundaries of prosodic units. Figure 2 displays the distributions of pause durations for the seven break types. As can be seen from the figure, the higher-level break types were generally associated with more prolonged pause duration. Note that *B*4, *B*3, and *B*2-2 have apparent pause duration (>30ms), while *B*0, *B*1, *B*2-1 and *B*2-3 all have very short pause duration (<30ms). By the above analysis on the pause duration of the seven break types, this study categorizes four break classes to ease the following analysis in Section 2.2, including (i) *B*4, (ii) *B*3, (iii) *B*2-2, and (iv) non-pause break type (NPB) which is a grouping of *B*0, *B*1, *B*2-1 and *B*2-3.

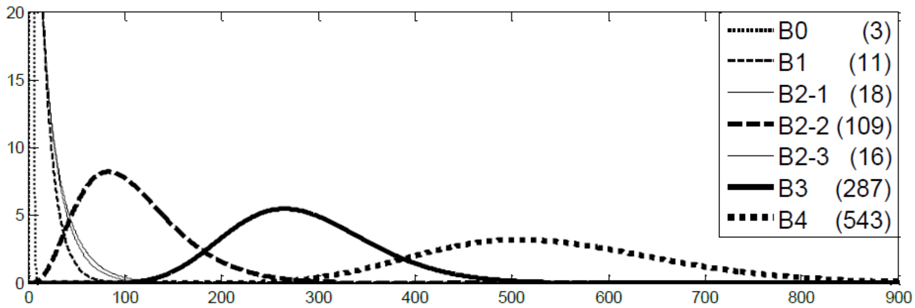


Figure 2. The distributions of pause durations (ms) for the seven break types. The average pause duration (ms) for each of the prosodic break type is displayed within the brackets.

2.2. Relationship Between the Labeled Break Types and PM Types

It is generally agreed that pause breaks co-occur with PMs. Most TTSs cautiously insert pause only on major PMs, such as comma and period. This cautious strategy of pause insertion can make the synthesized speech very stable but may be unnatural as the input sentence is very long and constituted in complicated syntactic structures. Table 1 shows the co-occurrence matrix of four break classes and three syllable juncture types calculated from the training set of the Treebank speech corpus. It can be seen from the table that most PM locations co-occur with pause-related break type (*B*2-2, *B*3, and *B*4), while most intra-word locations map to NPB. In-between of PM and intra-word, non-PM inter-word locations co-occur with NPB, *B*2-2, and *B*3. About 40% of prosodic phrase

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2018 doi:10.20944/preprints201802.0108.v1

boundaries (*B*3s) and over 94% of B2-2 come from non-PM inter-word junctures. By more detail analysis, we find that 60% of non-PM *B*3s coincides with depth-1 node boundary of the full parsed syntactic tree. The above discussions imply that it would be unsatisfactory to insert pause only at PM locations.

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

	NPB	B2-2	B3	B4
Intra-word	21,970	14	2	0
Non-PM inter-word	20,288	3,148	1,391	30
PM	30	169	2,130	2,320

Table 2 shows the co-occurrence matrix of four break classes and eight high-frequency PM types in the Treebank speech corpus. It can be found from the table that the MPM set {period '。', exclamation mark '！', question mark '？', semicolon '；', colon '：', comma '，'} is highly correlated with major breaks, i.e., *B*3 and *B*4. This implies that a word juncture which tends to insert an MPM in a text is more likely to be a major break in an utterance. This motivates us in this study to propose a CRF-based automatic MPM generator to predict the insertion of MPM (i.e., punctuation) and its likelihood (i.e., punctuation confidence, PC) for each word juncture, and use them to help the prosody generation.

Table 2. Correlation matrix of 4 break types and 8 PM types

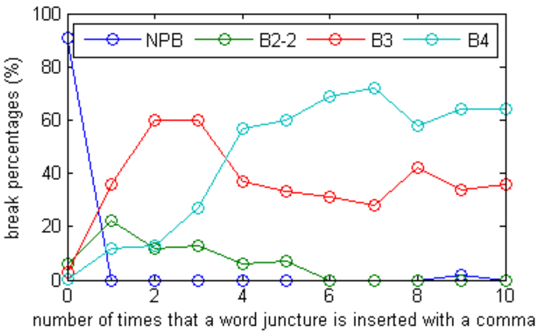
	。	！	？	；	：	，	、	・
NPB	1	0	0	0	0	4	25	1
B2-2	2	1	1	0	1	88	75	1
B3	42	1	7	9	2	1,901	168	0
B4	606	39	58	63	0	1,523	30	1

Note that in the texts of the training set of the Treebank speech corpus, no word string was quoted by Chinese brackets. This means we cannot directly analyze the relationship between Chinese brackets and labeled break types. In this paper, we directly analyze the characteristics of the brackets and their associated quoted phrases from the ASBC text corpus in Subsection 2.4.

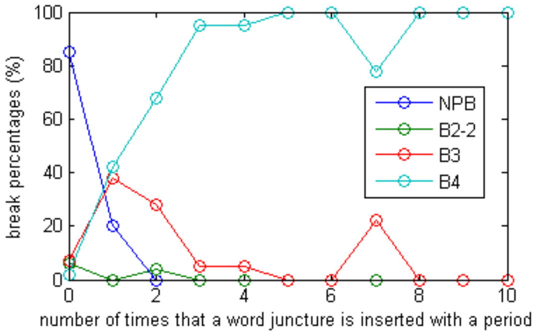
2.3. Human Labeled PMs vs. Prosodic Break Types

Evidently, we may conclude from the results shown in Table 2 that the occurrences of *B*3 and *B*4 are highly correlated with MPMs of periods, exclamation marks, question marks, semicolons, colons, and commas. We, therefore, assume that an automatic punctuation generation model predicting MPMs trained from a large text corpus can learn punctuation strategies for MPMs from various text contributors to provide informative cues for prosodic-acoustic feature predictions. To access the feasibility of the proposed idea, we conduct an experiment in which ten native Mandarin speakers are asked to insert periods and commas to the same 30 PM-deleted short paragraphs. These 30 paragraphs were chosen from the Treebank speech corpus which is labeled with prosodic breaks as stated in Section 2.1. The maximum and minimum lengths of the paragraphs are 270 and 80 characters, and the average length is 138 characters. The frequencies of word junctures being added with periods or commas can be regarded as the PCs made by human labelers (or text contributors). The analysis of the relationship between these frequencies (PCs by humans) and labeled prosodic breaks would provide some evidence that the proposed method is feasible.

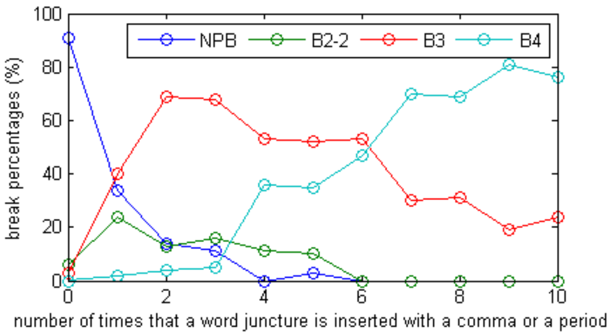
Figures 3(a)-(c) show average percentages of prosodic break types with respect to the number of times that a word juncture is inserted with a comma (Figure 3(a)), a period (Figure 3(b)), and a comma or a period (Figure 3(c)), respectively. Here, the number of the time that a comma or a period inserted is analogous to the proposed PC. We can find in Figures 3(a)-(c) that the percentages of NPB drop rapidly when the frequencies of MPM insertions increase. In Figure 3(a), it is found that percentages for *B*4 increase as the frequency of comma insertion increases. The percentage for *B*3 reaches the highest value around two/three comma insertions, and then decreases and keeps a level for more than four insertions. The percentage for *B*2-2 has a similar trend with the one for *B*3 but in a lower level. As can be seen from Figure 3(b), *B*4 dominates when more than three insertions of periods are observed for each word juncture. These results indicate that a word juncture is more likely to be inserted with pause-related break types (*B*2-2, *B*3, and *B*4) when the PC values are larger. It is also found that the break types of the higher prosodic units (i.e., larger break types) are likely to be associated with larger PC values. Figure 3(c) can be view as the result combined with Figures 3(a) and (b). Because commas and periods are major populations in the MPM set, the result shown in Figure 3(c) is analogous to the distributions of the prosodic break types concerning the PC values. We can observe more evident trends for the percentages of four break classes in Figure 3(c), and these trends would be informative for prosody generation.



(a)



(b)



(c)

Figure 3. Average percentages of prosodic break types concerning the number of times that a word juncture is inserted with (a) a comma, (b) a period and (c) a comma or a period

2.4. Analysis of Quotations

Table 3 shows 26 types of Chinese quotation marks existing in the ASBC text corpus [34]. We categorize words sandwiched by quotation mark into ten types according to their functions, and we called these sandwiched words 'quoted phrase' (QP). Table 4 shows the types of QPs, their statistics, and examples. In the following, we describe the characteristics of the QPs:

<u>Type 1</u> - （） : They mostly function as enumerating. Therefore, we do not regard Type 1 as our prediction targets for QP.

<u>Type 2</u> - ｛｝ : They are mostly titles of books or article, so we regard this type as our prediction targets.

<u>Type 3</u> - 〔〕 : They mostly function as captions of articles. This type is not included in our prediction target.

<u>Type 4 and 5</u> - 「」and『』: This type contributes most samples (68%) for the QP predictions since their properties are generally like word chunks or base phrases. For the single-word QPs of this type, they usually are emphasized nouns, verbs, or idioms. Most two- to four-word QPs are noun phrases. For QPs that longer than four words are generally long noun phrases or even sentences.

Types 6, 7 and 8 - 〈 〉【】《》: these types are similar to the Type 2 and therefore included in the QP prediction.

Type 9 - “”: We include 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 mostly are frequently-used phrases, and five- to six-word QPs are similar to sentences.

Type 10 - 〝〞: This type is similar to the types 4 and 5. We take this type as the QP prediction target though the sample size is very small.

Table 5 shows statistics of lengths of QPs in word. It is found that most QPs are single-word to four-word QPs. Single-word QPs are usually emphasized nouns or verbs. Two- to four-word QPs are mostly base-phrase like word strings (or word chunks). The QPs longer than four words are mostly sentence-like units.

Table 3. Types of Chinese quotations

NO.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Quotes	（	）	(	)	｛	｝	{	}	〔	〕	〔	〕	「	」	『	』
Type	1				2				3				4		5	

NO.	17	18	19	20	21	22	23	24	25	26
Quotes	〈	〉	【	】	《	》	“	”	〝	〞
Type	6		7		8		9		10	

Table 4. Types of QPs, their statistics, and examples. Examples are delimited by commas, and words are delimited by slashes for each example.

type		count (percentage)	Examples
1	（）	14131 (25.13%)	S，圖/一，見/左/圖，本/報/資料/照片，一種/蔗糖/做成/的/蘭姆酒，不/合/者/恕/不/退件
2	｛｝	34 (0.06%)	桃花源記，山居/筆記，松花江/的/浪
3	〔〕	101 (0.17%)	本報訊，其他/功能，美麗/與/哀愁，草地/上/的/午餐，倫飛/電腦/公司/應對/之/道
4	「」	37197 (66.17%)	人，企業/改造，十八歲/的/約定，戀愛/中/的/寶貝，全/國/原住民/教育/會議，臺北市/土地/使用/分區/管制/規則
5	『』	1223 (2.17%)	他，新/民族，廣島/什錦/煎餅，與/夫/訣別/書，大家/來/寫/村/史，羅浮宮/博物館/珍藏/名/畫/特展
6	〈〉	562 (0.99%)	夾竹桃，芝蘭室/記，馬難/明白/了，銀鬚/上/的/春天，一隻/米蘭/夜梟/的/報告，瘂叔/和/他/的/孫子/們
7	【】	314 (0.55%)	宗教，趙/６８，救主/的/使命，對/你/的/忠告，族群/與/文化/政策/綱領，男人/的/一半/還/是/男人
8	《》	2523 (4.48%)	芝蘭室圖，黃色/壁紙，存有/與/時間，屋頂/上/的/小孩,在/我/墳/上/起舞，我/和/我/豢養/的/宇宙
9	“”	105 (0.18%)	蒼蠅，新/音樂，助人/之/服務，只要/信/不要/怕，創造/海/中/的/動物，人/死/後/靈魂/仍然/存在
10	〝〞	22 (0.039138%)	善有善報，第一/夫人，女人/的/私家/珍藏

Table 5. Statistics of lengths of QPs in word.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2018 doi:10.20944/preprints201802.0108.v1

Length in word	# of example	percentage
1	26791	41%
2	16749	25%
3	10933	17%
4	5847	9%
5	3415	5%
5	1988	3%

3. The Proposed Punctuation Confidence

3.1. The CRF-Based MPM Generator

The Punctuation Confidence (PC) [28] is produced by a CRF-based MPM generator. The task of the CRF-based MPM generator can be viewed as a label-tagging problem that labels each lexical word juncture with a sequence of types of PMs, e.g., presence or absence of an MPM, **Y**, by using some linguistic feature sequence, **X**. It is formulated by

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

where $N(\boldsymbol{X})$ is a normalization factor to ensure that $\sum_{\mathbf{Y}} P(\mathbf{Y}\,|\,\boldsymbol{X})=1$; t stands for lexical word index; Y_t represents prediction target, i.e., 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

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

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. Generally, feature contexts are organized into several groups, referred to as ‘feature templates.’ The predicted PM sequence can be obtained by the Viterbi search:

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

Moreover, the PC is given by the forward/backward calculation:

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

which is the marginal probability of the k-th type of PM for the t-th word.

3.2. The Design of Prediction Targets

Two types of prediction targets are designed: the basic PC (bPC) and the improved PC (iPC). The bPC is generated by considering the two prediction targets: the presence of an MPM, y_1, and the absence of an MPM, y_0. The iPC is produced by considering structures of sentences accompanying with MPMs. For the bPC, the MPMs includes ‘。’, ‘！’, ‘？’, ‘；’, ‘：’, and ‘，’. The PC, $\varphi_{t,k}(\mathbf{X})$, generated by the target setting {y_1, y_0} is called the basic PC (bPC). Figure 4(a) shows the original text with word/PM tokens and Figure 4(b) shows the corresponding target-labeling example for the training of bPC.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2018 doi:10.20944/preprints201802.0108.v1

(a) 望遠鏡 可以 用來 看 天 上 明亮 閃爍 的 星星 ，或是 水濱 的 野鳥 ，也 可以 用來 看 人。

(b) 望遠鏡/y_0 可以/y_0 用來/y_0 看/y_0 天/y_0 上/y_0 明亮/y_0 閃爍 y_0/ 的/y_0 星星/$\mathbf{y_1}$ 或是/y_0 水濱/y_0 的/y_0 野鳥/$\mathbf{y_1}$ 也/y_0 可以/y_0 用來/y_0 看/y_0 人/$\mathbf{y_1}$

(c) 望遠鏡/B1 可以/B2 用來/B3 看/B4 天/M 上/M 明亮/E4 閃爍/E3 的/E2 星星/E1 或是/B1 水濱/B2 的/E2 野鳥/E1 也/B1 可以/B2 用來/I 看/E2 人/E1

(d) **Instance 1:** 望遠鏡/E1 可以/E2 用來/E3 看/E4 天/M 上/M 明亮/E4 閃爍/E3 的/E2 星星/E1 或是/b1 水濱/b2 的/e2 野鳥/e1

Instance 2: 或是/B1 水濱/B2 的/E2 野鳥/E1 也/b1 可以/b2 用來/i 看/e2 人/e1

Figure 4. An exemplary tag labeling for the PC training: original word/PM sequence is shown in pane (a), the tag labeling for the training of bPC (b), iPCst (c), and iPCef (d). Note that each sentence is in a different color and each word is delimited by spaces.

Note that the bPC only considers modeling the insertion of the MPMs and the MPMs serve as delimiters for sentences. Therefore, modeling structures of sentences could be equivalent modeling insertion of MPMs and even could give a better prediction of MPM insertion. Besides, by an analysis on the ASBC text corpus [34], it is found that many long sentences could be inserted with some optional MPMs without losing understanding. These optional inserted MPMs may correspond to insertion of pause breaks. We hence proposed so-called the improved PC (iPC) to model sentence structures and optional MPMs in a sentence. Two types of the iPC are designed: iPCst and iPCef. The iPCst is designed for modeling of sentence structure while iPCef is for modeling of an enforced MPM insertion in a sentence. For the prediction of the iPCst, the prediction targets for the CRF-based MPM generator are labeled for each word and designed to represent sentence structures regarding word position in a sentence. The targets 'B', 'I', 'M', 'S', and 'E' respectively present beginning, intermediate, middle, single and ending words in a sentence. To further precisely label the word order information in a sentence, numbers 1 to 4 are added to the targets 'B' and 'E' for indicating forward and backward word order. According to the statistics about sentence length in word for the ASBC text corpus, the length of sentences mostly (84%) distributes from 4 to 9 words. The target labeling schemes, therefore, are designed differently for sentences with ≤ 9 and > 9 words. The complete targets for iPCst are listed in Table 6. Specifically, there are four rules to guide the tagging of targets:

1. 'B1', 'B2', 'B3', and 'B4' represent the first, second, third, and fourth word in a sentence respectively while 'E1', 'E2', 'E3', and 'E4' represent respectively the first last, second last, third last, and fourth last word in a sentence.
2. If sentence length is > 9 words, we use 'B1'~'B4' and 'E4'~'E1' to tag targets from the beginning and the ending of a sentence and use 'M' to tag the other intermediate words in a sentence.
3. If sentence length is ≤ 9 words and even, we use 'B1'~'B*k*' and 'E1'~'E*k*' to tag targets from the beginning and the ending of a sentence for *k*=1~4 and *k*= (length of sentence in word)/2.
4. If sentence length is ≤ 9 words and odd, we use 'B1'~'B*k*' and 'E1'~'E*k*' to tag targets from the beginning and the ending of a sentence for *k*=1~4 and *k*= (length of sentence in word)/2. The rest of the words are labeled with 'I' to indicate the intermediate words in a sentence.

Figure 4(c) shows an exemplary tag labeling for the iPCst training.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2018 doi:10.20944/preprints201802.0108.v1

The idea of the prediction of the iPCef is to enforce inserting an MPM in a sentence. This enforced MPM may provide informative cues for inserting a pause or exhibit a pre-boundary syllable duration lengthening for word junctures in a long sentence. To realize this enforced MPM insertion, the prediction targets are designed to learn to insert an MPM given instances of two consecutive sentences whose sandwiched MPM are removed. The target set for iPCef is similar to the one for iPCst shown in Table 6 but using upper- and lower-case letters for the distinction between tags respectively for first and second sentences. This idea is motivated by observing frequent pause insertions in long sentences as shown in Section 2. Figure 4(d) shows an example of prediction target labeling for iPCef. Noted that in the training of iPCef, two consecutive sentences are taken as one training instance for an enforced MPM insertion.

Table 6. Targets for iPCst

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

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 7. The 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 generators for PCs considered the contextual word, POSs, length of the word, and the combinations of the above features. In this study, we design four templates for the PC generation as shown in Table 8. All the templates consider the same POS, lexical word-POS and word length contexts. The difference between the templates 1 and 2 is that the template 2 considers wider word contexts. The templates 3 and 4 are similar to the template 1 and 2 but different in that the 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 we observe that the types of the current PM, Y_t, depend on the joint factor of the previous PM type, Y_{t-1}, and the current POS, S_t.

3.5. The Experiment of PC Generation and Evidence

The CRF models were trained by the ASBC [34] training set with 6,625,277 words, and the best feature templates were tuned by the results on the training set with 2,817,785 words. The tool for the training is CRF++: Yet Another CRF toolkit [36]. Table 9 shows precisions and recalls of predicted MPM insertions trained by setting prediction targets of bPC, iPCst and iPCef with the templates 1 to 4. It is observed that the best precision and recall are achieved by the template 4, followed by the templates 3, 2 and 1, indicating that the wider feature contexts and joint factors of (Y_{t-1}, S_t) could

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2018 doi:10.20944/preprints201802.0108.v1

improve the MPM prediction. The best precision/recall of MPM generations on the test set for bPC, iPCst and iPCef are respectively 94.1%/93.1%, 96.9%/96.1%, and 95.7%/95.5%. We choose the results made by the template 4 for the following analysis and prosody generation experiments. The results were reasonably high to model the characteristics of MPM insertion and sentence structures.

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	$\{W_{t+\tau}\}_{\tau=-1\sim+1}, \{W_{t-1+\tau}^{t+\tau}\}_{\tau=0,1}$ W_{t-1}^{t+1}	W_t	$\{W_{t+\tau}\}_{\tau=-1\sim+1}, \{W_{t-1+\tau}^{t+\tau}\}_{\tau=0,1}$ W_{t-1}^{t+1}
POS context	$\{S_{t+\tau}\}_{\tau=-3\sim+3}$, $\{S_{t-1+\tau}^{t+\tau}\}_{\tau=0,1}, \{S_{t-2+\tau}^{t+\tau}\}_{\tau=0\sim2}, \{S_{t-3+\tau}^{t+\tau}\}_{\tau=0\sim3}, \{S_{t-3+\tau}^{t+1+\tau}\}_{\tau=0\sim3}, \{S_{t-3+\tau}^{t+2+\tau}\}_{\tau=0,1}$			
Lexical word and POS context	$\{(W_t, S_{t+\tau})\}_{\tau=-3\sim+3}, \{(W_t, S_{t-1+\tau}^{t+\tau})\}_{\tau=0,1}$, $\{(W_t, S_{t-2+\tau}^{t+\tau})\}_{\tau=0\sim2}$, $\{(W_t, S_{t-3+\tau}^{t+\tau})\}_{\tau=0\sim3}$, $\{(W_t, S_{t-3+\tau}^{t+1+\tau})\}_{\tau=0\sim3}$			
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. The precisions and recalls of the MPM generations by 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**

We then examine the interplay between the proposed PC values, i.e., $\varphi_{t,k}(\mathbf{X})$, and distributions of prosodic-acoustic features on the training set of the treebank speech corpus in Figures 5, 6 and 7. Figure 5 shows the average syllable logF0s corresponding to the prediction targets for bPC (a), iPCst (b) and iPCef (c) in different levels of PC values. Note that the PC values are divided into ten even intervals from 0 to 1 for the bPC in Figure 5(a). As can be seen from Figure 5(a), the average syllable logF0 decrease as the bPC for MPM, i.e., $\varphi_{t,k}(\mathbf{X})$ for the prediction target y_1, increases while the bPC for y_0 exhibits a contrary trend. This indicates that a syllable would have lower logF0 value as the syllable is more likely to be followed by an MPM. Figure 5(b) shows the average syllable logF0 of the prediction targets in the three representative levels of iPCst values, i.e., the high level: iPCst = 0.9~1.0, the 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 a sentence on the x-axis, i.e., 'B1', 'B2', 'B3', 'B4', 'I'/'M', 'E4', 'E3', 'E2', and 'E1'. A clear trend of logF0 declination can be found for the high-level iPCst. On the contrary, the average syllable logF0s are flat for the low-level iPCst. The average syllable logF0s for the median-level iPCst shows a moderate logF0 declination trend. Figure 5(c) shows the average syllable logF0 of the prediction targets in the three representative levels of iPCef values. The prediction targets in Figure 5(c) are also listed in a forward position order in a sentence on the x-axis. The logF0 declination effects are also clearly observed for the cases of the high and median levels of iPCef values. These findings may indicate that the proposed PCs could provide informative cues for modeling logF0 declination effect in prosody generation. Besides, iPCst and iPCef (especially iPCef) exhibited a higher and lower logF0s in the beginning and end of a sentence, respectively, indicating the proposed iPCst and iPCef may provide more significant cues than bPC for prosody generation.

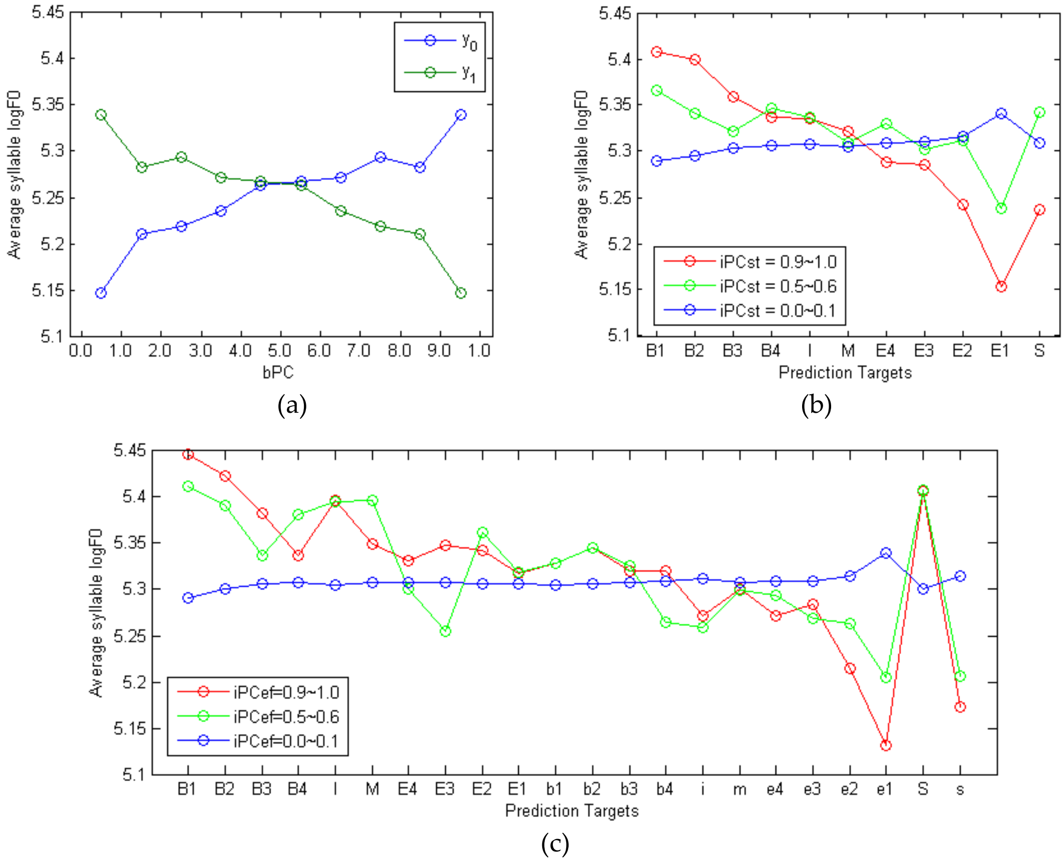
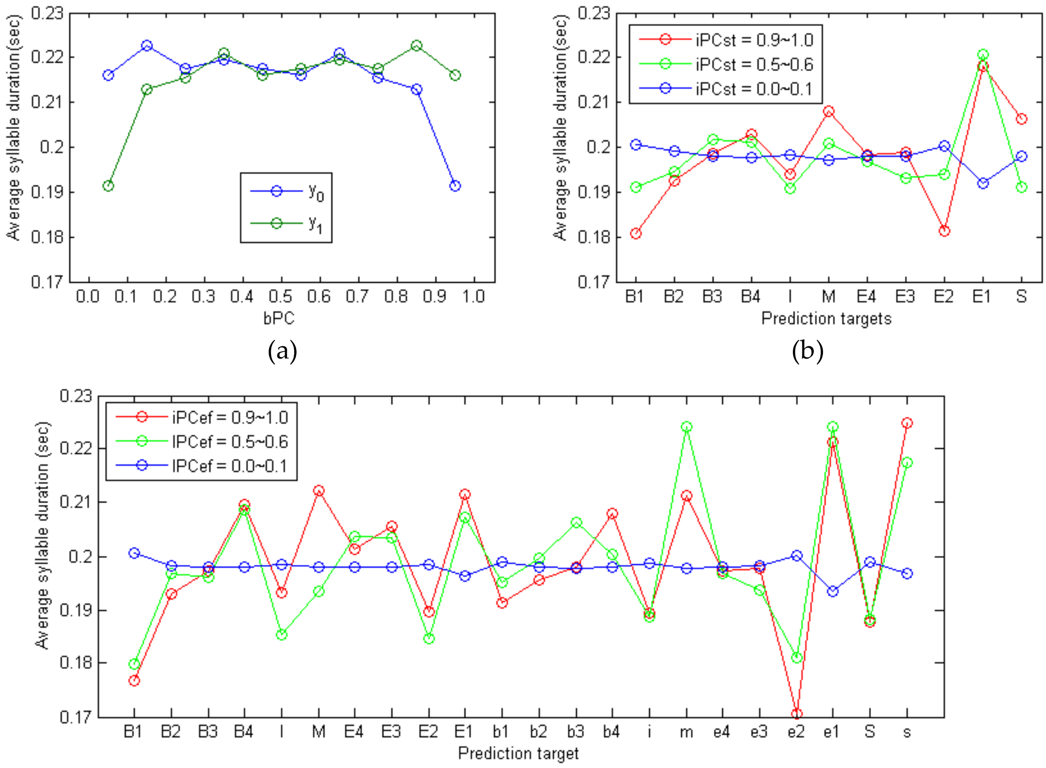


Figure 5. Average syllable logF0s corresponding to the prediction targets for bPC (a), iPCst (b) and iPCef (c) in different levels of PC values.

Figure 6 shows the average syllable duration corresponding to the prediction targets for bPC (a), iPCst (b) and iPCef (c) in different levels of PC values. It is found in Figure 6(a) that the average syllable durations are shortened for the two extreme cases: bPC for $y_1 < 0.1$ and bPC for $y_0 > 0.9$. This result indicated that the bPC could provide cues to shorten or lengthen the syllable durations when it is very unlikely or likely to insert an MPM following the target syllable. Figure 6(b) shows 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 on the x-axis. Significant long average syllable durations can be found at the prediction target of 'E1' which represents a syllable followed by an MPM for the high and median iPCst levels. It is reasonable to observe a slightly longer average syllable duration for the target 'M' because the target 'M' represents an intermediate location in a long sentence where is more likely to be inserted with a prosodic break. The average syllable durations for all the prediction of the low-level iPCst are almost in the same level. These results indicate that the proposed iPCst can model the pre-boundary syllable duration lengthening effect with various degrees of the iPCst values. It is also found that in the case of the prediction target 'S' which represent a word sandwiched by preceding and following MPMs, the syllable is lengthened as the iPCst value is high. The prediction targets 'B1' (the first syllable in a sentence) and 'I' (the intermediate syllable in a short sentence) have shortened average syllable durations compared with their nearby syllable locations in a sentence. These results coincide the findings in the previous studies [46] about syllable durations in a PPh. In the paper [46], it was found that first syllable in a PPh and intermediate syllable in a short PPh is shortened. The shortened syllable duration for the target 'E2' (the second last syllable in a sentence) manifested a significant contrast for the following pre-boundary syllable duration lengthening cue by the prediction target 'E1'. In Figure 6(c), the trends of average syllable durations of the prediction targets for the first sentence and the ones of the second sentence are similar. It is also reasonable to observe a slightly longer average syllable duration for the targets of 'B4', 'M', 'b4', and 'm' because these targets are distant to the beginning and the ending of a sentence, resulting in a more probable insertion of a prosodic break. Note that the CRF-based MPM generator for the iPCef predicts an enforced MPM for each sentence. Words of each sentence are therefore labeled with the prediction targets of {'B1', 'B2', … 'E2', 'E1', 'S', 'b1', 'b2', …'e2', 'e1', 's'} to represent delimiting one sentence into two (the first and second sentences). The prediction target 'E1' in this case indicates that there exists an enforced inserted MPM in a sentence. The similar trends for the average syllable durations of the first and second sentences indicated that the proposed iPCef could more sophisticatedly model syllable duration patterns for a long sentence which may be delimited into two PPhs. Recall that as stated in Section 2.2, 40% of prosodic phrase boundaries (*B*3s) come from non-PM inter-word junctures. It is, therefore, encouraging to observe this syllable duration patterns made by the enforced insertion of MPM by modeling of iPCef. The superiority of the proposed iPCef over the proposed iPCst and bPC in the prediction of syllable duration is partially confirmed by the prosody generation experiment shown later in this paper (Section 5.3).



(a) (b)

(c)

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

Figure 7 shows the pause durations corresponding to the prediction targets for bPC (a), iPCst (b) and iPCef (c) in different levels of PC values. Figure 7(a) shows a trend that the average pause durations increase as the bPC for MPM, i.e., $\varphi_{t,k}(\mathbf{X})$ for the prediction target y_1, increases while the bPC for y_0 exhibits a contrary trend. Long pause durations can be found for the prediction targets of 'E1' and 'S' for the high and median levels of iPCst. We may conclude from the mentioned-above observations that the higher bPC or iPCst values would result in longer pause durations for the predicted MPM locations. In Figure 7(c), the trend of pause durations for the prediction targets of the second sentence is similar to the ones in Figure 7(b). The prediction target 'E1' for the first sentence only shows a slightly longer pause duration compared with the nearby targets. The pause durations for 'E1' is at the same level for the prediction targets that represent intermediate locations of a long sentence, i.e., 'B4', 'M', and 'm'. This result indicates that the iPCef features would not provide as salient cues for pause duration prediction as the iPCst features would. The objective evaluations of the prosody generation experiment shown later in this paper (Section 5.3) partially confirm this indication.

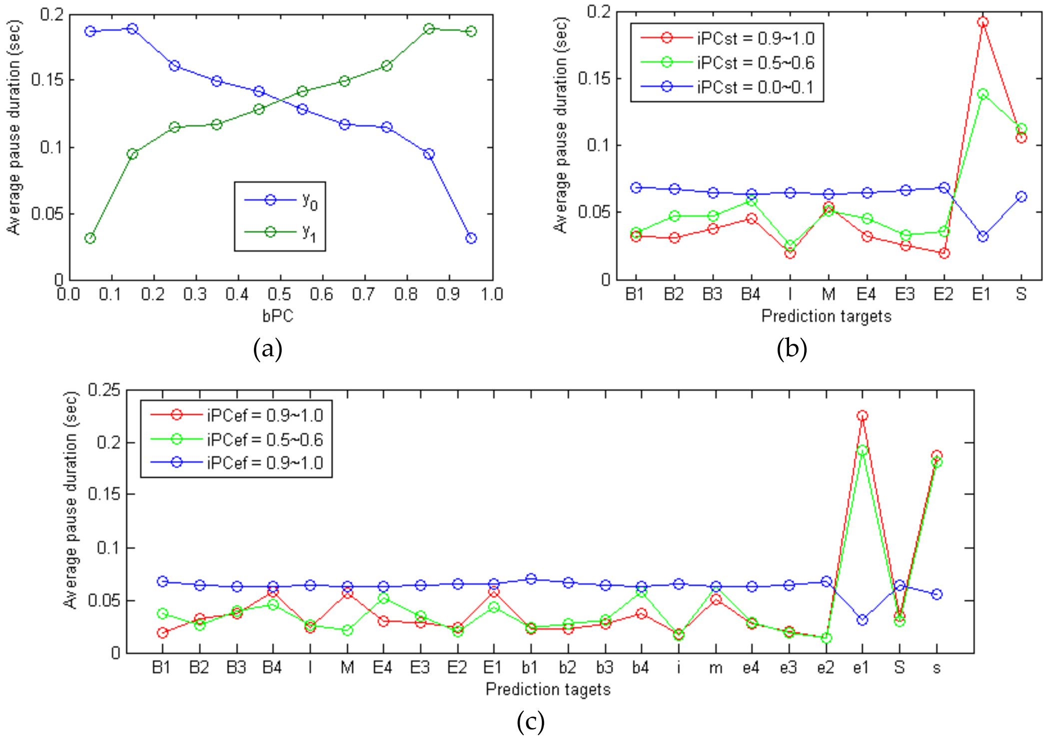


Figure 7. Average pause durations corresponding to the prediction targets for bPC (a), iPCst (b) and iPCef (c) in different levels of PC values.

4. The Quotation Confidence

4.1. The Design of Prediction Targets

The prediction of QPs is also developed by the CRF model as described in Section 3. The target, y_k, is the k-th possible tag representing word position in a QP. The optimal QPs, $Y_1^*,\ldots,Y_T^*$, can be predicted by Eq. (3), and the marginal probability for the k-th tag of the t-th word, $\varphi_{t,k}(\mathbf{X})$, is called the Quotation Confidence (QC) generated by Eq. (4). Two types of QCs are designed in this study: basic QC (bQC) and sentence structure QC (sQC). The bQC is generated by predicting structures of QPs while sQC is generated by predicting both structures of QPs and their position in a sentence. As shown in Table 10, an 8-tag set is designed for modeling bQC. Besides, an additional tag 'O' is used to represent non-QP words. Figure 8(b) shows a target labeling example for the training of the bQC whose original word/PM tokens are shown in Figure 8(a). The sQC can be regarded as an improved version of bQC that use additional tags to represent positions of non-QP words in a sentence. These additional tags are designed in a two-alphabet format: xy where $x \in \{\mathrm{B,M,F}\}$ represents a word string before a QP (B), in-between two QPs (M), or following a QP (F); $y \in \{\mathrm{b,m,e,s}\}$ represents beginning (b), intermediate (m), the last (e), or a single word in a word string (s). Figure 8(c) shows a tag example for the sQC training. The complete set of the prediction target for sQC is shown in Table 11.

Table 10. Tag format for labeling of target QP 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2018 doi:10.20944/preprints201802.0108.v1

(a) 其實 〔 中醫 理論 〕 中 最 有 〔 特色 之 處 〕 就是 氣 行血 ，

(b) 其實/O 中醫/B 理論/E 中/O 最/O 有/O 特色/B 之/I 處/E 就是/O 氣/O 行血/O

(c) 其實/Ps 中醫/B 理論/E 中/Mb 最/Mm 有/Me 特色/B 之/I 處/E 就是/Fb 氣/Fm 行血/Fe

Figure 8. (a) Original word/PM tokens, (b) an exemplary tag labeling for the bQC training, and (c) an exemplar for the sQC training

Table 11. Tag format for labeling of target QP for bQC

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

4.2. Design of Features and Templates

As shown in Table 12, the features used for the prediction of QP are similar to the ones used for the prediction of PC. The newly-added PM features are used to indicate information about sentence boundaries. Table 13 shows the five templates for the QP prediction in this study. In the template 1, we use a 3-POS context, i.e., from (*t-1*)-th to (*t+1*)-th in the POS field. The word-and-POS field contains the combined features of a 3-POS context and current word (W_t). The templates 2 and 3 respectively use a 5-POS context and a 7-POS context, and their combination with the current word. The templates 4 and 5 are identical to the templates 2 and 3 respectively in all feature fields except for the lexical word context field. We use a five-lexical word context for the templates 4 and 5.

Table 12. The 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 sentence boundary
L_t	Length of *t*-th lexical word in syllable	The structure of a QP is related to word length combinations

Table 13. Feature templates for bQC and sQC

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

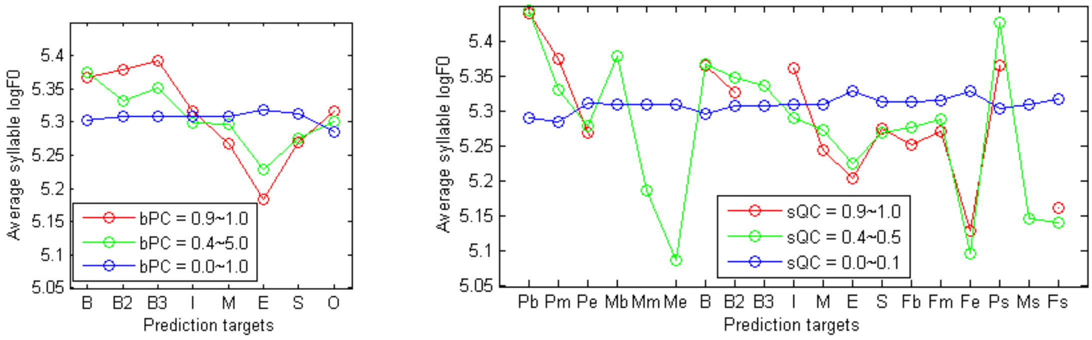
4.3. The Experiment of QC Generation and Evidence

Notice that only 0.69% of the ASBC text corpus contributed instances of QPs, i.e., only 65,723 QP token examples. To make the CRF models for QC concentrate more on predicting QPs, we only selected the sentences with QPs for training and testing. The numbers of QP tokens for training and testing are respectively 57,824 and 8,439. Table 14 shows the precisions and recalls for bQC and sQC. It can be seen from the tables that the five templates result in similar precisions and recalls. The best results are achieved by the template 5 for bPC and the template 4 for sQC. We, therefore, choose the best models trained by the templates 4 and 5 for the following analysis and prosody generation experiments. The precision and recall for predicting bQC are respectively around 60.7% and 39.0% while the precision and recall for sQC are respectively around 55.6% and 52.2%. These results show that modeling both structures of QPs and their position in a sentence could improve the prediction of QPs. Though the precision and recall are relatively much lower than the ones of the prediction of the PC, it is more interesting to analyze the interplay between the prosodic-acoustic features and the QC values, i.e., $\varphi_{t,k}(\mathbf{X})$.

Table 14. QC model predictions 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

Figure 9(a) shows the average syllable logF0 of the prediction targets in the three representative levels of bQC values, i.e. the high level: bQC = 0.9~1.0, the median level: bQC = 0.4~0.5, and the slow 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, i.e., 'B', 'B2', 'B3', 'I'/'M', and 'E'. We can observe a clear logF0 declination trend for the high and median bQC levels within a QP. The average logF0s for the single-word QP and non-QP are at around the average levels. On the contrary, the average syllable logF0s are flat for the low-level iPCst. We may conclude from the mentioned-above observation that a string of words may have logF0 reset at the beginning of the string and then decline gradually as the string is more likely to be labeled as a QP. The logF0 declination within a QP can also be observed in Figure 9(b) for the median and high levels of sQC values. Note that some of the average logF0 of the prediction targets for the high-level sQC, i.e., 'Mb', 'Mm', 'Me', 'B3' and 'Ms', are missing because the high sQC values were not generated by the CRF-based quotation generator for these prediction targets. Besides, logF0 declination can also be observed for the word string preceding to ('Pb', 'Pm' and 'Pe') and following ('Fb', 'Fm' and 'Fe') a quoted phrase. We, therefore, expect the sQC features provide more informative cues for logF0 generation than the bQC features d. The objective evaluations of the logF0 generation experiment shown later in this paper (Section 5.3) partially meet this expectation.



(a) (b)

Figure 9. Average syllable logF0s corresponding to the prediction targets for bQC (a), sQC (b) in different levels of QC values.

Figure 10 shows the average syllable durations 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/post-boundary duration lengthening effect may be modeled by the trends of the QCs shown in Figures 10(a) and (b) because the average syllable durations for prediction targets of 'B', 'B2', and 'E' increase as the QCs increases. It is also interesting to find that the syllable durations for the target 'S' which represent a single-word QP are longer as the corresponding QC values increase. Note that some of the average syllable durations of the prediction targets for the high and median level QCs are missing because we do 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 the last words ('Fe') in a word string which is followed by a QP, respectively. The objective evaluations of the syllable duration generation experiment shown later in this paper (Section 5.3) show that these QC features can make the RMSE of the synthesized prosody lower than the RMSE by the conventional linguistic features, confirming the QC features are useful in prosody generation.

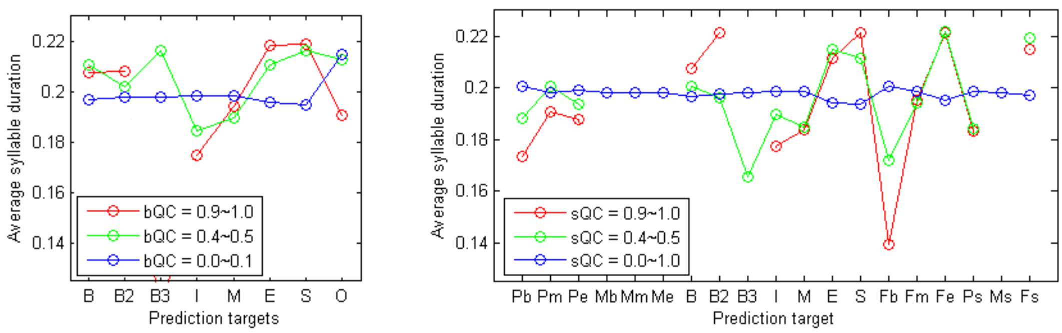


Figure 10. Average syllable durations corresponding to the prediction targets for bQC (a), sQC (b) in different levels of QC values.

doi:10.20944/preprints201802.0108.v1

Figures 11(a) and (b) shows the trends that a word which is more likely to be the end of QPs, i.e., the tags 'E' and 'S', is more tentative to be followed by a long pause while the other tags except for the tag 'Fe' exhibit a contrary trend. Because the sQC features provide more sophisticated structures of QPs and their contexts, we expect that the sQC features generate pause durations with lower RMSEs than the bQC features do.

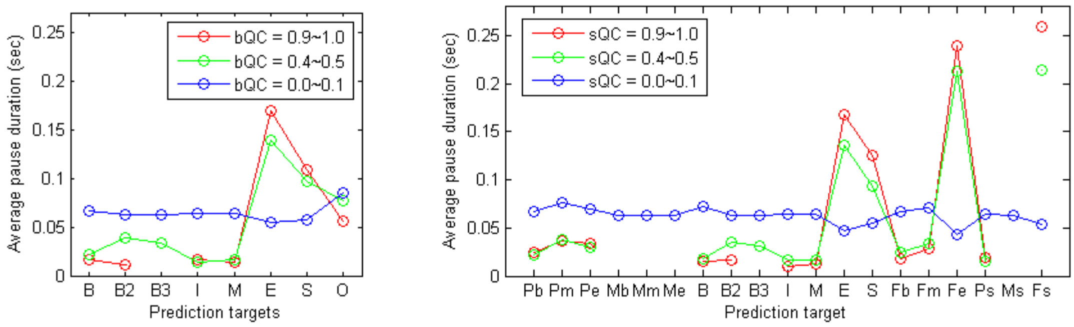


Figure 11. Average pause durations corresponding to the prediction targets for bQC (a), sQC (b) in different levels of QC values.

5. Prosody Generation Experiments

Figure 12 shows the flowchart for the experiments of prosody generation. First, the texts are 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 include the conventional linguistic processors commonly used in MTTS and the proposed advanced PC and QC generators. Next, the four independent MLPs are trained with the conventional linguistic feature sets and the proposed PC and QC features to predict syllable logF0 contour (lf0), syllable duration (Dur), syllable energy level (Eng), and inter-syllable pause duration (Pau). Then, we conduct some objective tests to evaluate the RMSEs between the predicted prosodic-acoustic features and the true prosodic-acoustic features. Here, the predicted prosodic-acoustic features are generated by the given different settings of linguistic features to prove the usefulness of the proposed PC and QC features. Last, we utilize an HMM-based speech synthesizer with the predicted prosodic-acoustic features to generate synthesized speeches. These synthesized speeches are used to conduct subjective tests, showing that the proposed PC and QC features could improve the naturalness of the synthesized speeches.

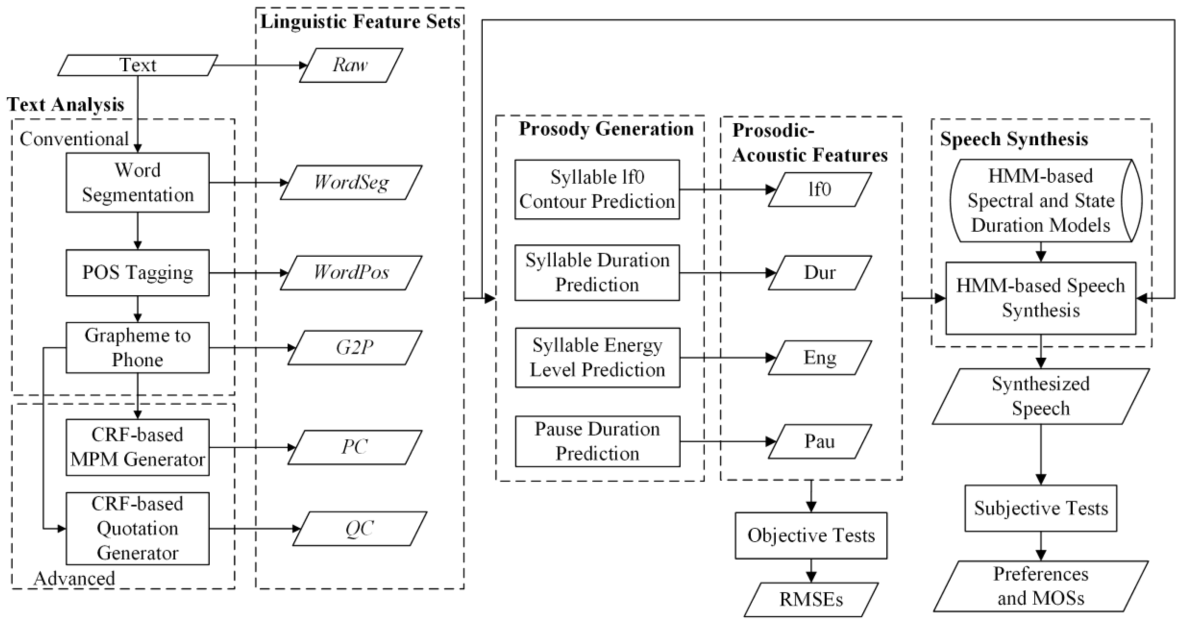


Figure 12. The flowchart for the experiments of prosody generation.

5.1 Text Analysis and Linguistic Feature Sets

Figure 12 also shows the linguistic processors used and the associated linguistic features generated in this study. To set up various settings of experiments, the processors are categorized into two classes: 1) baseline processor and 2) the proposed advanced processor. The baseline processor contains functions of word segmentation, POS tagging, and grapheme to phone (G2P). Basically, features generated from the baseline processor cover linguistic information of phonetics, lexical word, and POS. Since the features extracted by the baseline processor are prevalent in most MTTSs [4,12-14,17,22,24-27], we regard the features generated from the baseline processor as the base linguistic features for prosody generation. In this study, we adopt NCTU Speech Lab Traditional Chinese Parser [43,44] as the baseline processor. It is an online CRF-based word tagger and generates information about word boundaries and the associated categories of POS. The F-measure of 96.72% for the word segmentation and the accuracy of 94.16% for the POS tagging are reported [44]. This study includes two advanced processors: the CRF-based MPM generator and the CRF-based quotation generator which were described in Section 3 and Section 4, respectively. These two advanced processors are cascaded after the baseline processor. The features used in the prosody generation experiments are 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 from a raw text is the type of PM. PM is the most salient feature for predicting pause break because PMs serve as a delimiter in both syntax and intonation in Mandarin Chinese. Since sentence boundaries in Chinese can be identified by types of PMs, a contextual feature of syllable position in a sentence can also be extracted from the raw text. The positional features are highly related to rhythmic patterns of syllable duration and syllable F0 contour, e.g., syllables at the end of a sentence usually exhibit both syllable duration lengthening and F0 declination.

5.1.2. *WordSeg*

The features in subset *WordSeg* are extracted after the word segmentation, including word length, syllable position in a word, and word position in a sentence. For the feature of word length, it is conventional to include lengths of neighboring words because PWs are usually composed of several words with some length constraints. Most studies consider a window of five words [16,25] with the current word, two words to the left and the right. In this study, we extend the window to seven words, i.e., the current word, three words to the left and the right. The positional features in this subset are also essential to syllable duration patterns. The most significant evidence is that syllable position in a word affects the degree of syllable duration lengthening [4].

5.1.3. *WordPos*

The features in subset *WordPos* are POS tags for the associated words and are obtained after the POS-tagging process. It was found that PWs were generally composed of 1-3 words with some POS combinations [12,13,38] given by word length constraints. Also, it is generally agreed that prosodic breaks or pause insertion were related to some POS pairs on word junctures [12,13,38]. Therefore, POS and word length are the most frequently used and important features for predicting prosody structures from texts. In this study, we adopt a 47-POS tag set [45] which is used by the NCTU Speech Lab Traditional Chinese Parser. Similar to the usage of word length, the analysis window size for POS is set to at most seven words, i.e., the current word, three words to the left and the right.

5.1.4. *G2P*

G2P set comprises important features characterizing properties of Mandarin prosody: tone, and base-syllable type, or initial-final type. There are five tones in Mandarin Chinese. To account for more prosodic variation that resulted from contextual tones, the tones of the current, following and

 doi:10.20944/preprints201802.0108.v1

previous syllables are considered for prosody generation. There are around 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 numbers of features, we take initial and final types as features to account for information of base-syllable type. In this study, we define 23 initial types and 40 final types. Besides 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.

5.1.5 Advanced Feature Set – *PCs* and *QCs*

The set comprises *PC* and *QC* generated correspondingly by the proposed CRF-based MPM generator and the proposed CRF-based quotation generator. The subset *PC* consists of the predicted punctuation sequence by Eq. (3), i.e. $Y_1^*, Y_2^*, \cdots, Y_T^*$, and the PC by Eq. (4), i.e. $\varphi_{t,k}(\mathbf{X})$, with target settings of bPC, iPCst, and iPCef. The subset *QC* consists of the predicted quotation label sequence, i.e. $Y_1^*, Y_2^*, \cdots, Y_T^*$, and the QC, i.e. $\varphi_{t,k}(\mathbf{X})$, with target settings of bQC and sQC.

5.2. MLP-based Prosody Generation

The prosody generation experiments were conducted by four independent MLPs to train prediction models for syllable logF0 contour (lf0) represented by 4-dimensional discrete orthogonal expansion coefficients [47], syllable duration (Dur) in sec, syllable energy level (Eng) in dB, and inter-syllable pause duration (Pau) in second. The feature vectors for the input layer of the MLPs can be categorized into three main categories for comparison: (1) baseline (*BSL*), (2) the proposed bPC, iPCst and iPCef (*PCset*), and (3) the proposed bQC and sQC (*QCset*). The *BSL* contains the most basic linguistic feature sets: *Raw*, *G2P*, *WordSeg* and *WordPos*. There are 28 and 67 features in the set *Raw* and *G2P*, respectively. The feature sets bPC, iPCst, iPCef, bQC, and sQC respectively are composed of 4, 22, 44, 16, and 38 numerical features representing the marginal probabilities $\varphi_{t,k}(\mathbf{X})$ 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*/*WordPos* were tuned by the development set.

5.3. Objective Tests

Table 15 shows RMSEs for the prosodic-acoustic features by various linguistic feature sets. Generally, the proposed *PCSet* and *QCset* can generally improve the RMSEs w.r.t. *BSL*. For the lf0 prediction, the feature sets with the proposed PCs or QCs generally performed better than the ones without the PCs/QCs. The best RMSE for lf0 was achieved by using the set QC2=*BSL3*+*sQC*. This result may be contributed from the properties of the sQC that models syntactic structures of base phrases or word chunks that are highly correlated with structures of prosodic words (PWs). It is also found that the feature sets with sQC could improve more RMSE than the ones with bQC did because sQC not only describe structures of QPs but also structures of their contexts. The proposed iPCst and iPCef can generally outperform the proposed bPC because they could model structures of sentences that are highly correlated with structures of PPhs or intonation phrases (IPs).

For the predictions of Dur and Pau, the feature sets with *WordPos* could generally outperform the ones without *WordPos*. This partially confirms that the POS combination features are essential for the predictions of the structures of PWs, PPh, and IPs. When adding the proposed QCs and PCs, further improvements were achieved because the QCs and the PCs may provide information that may correlate with structures of PWs, PPh, and IPs. The iPCef could slightly perform better than the iPCst, bQC, and sQC in the prediction of Dur. This is maybe because the iPCef models a forced insertion of an MPM in a sentence to provide more information for pre-boundary syllable duration lengthening. Besides, it is reasonable to see that iPCst gave the best performance in the prediction of Pau since iPCst models structures of sentences which highly correlates with PPhs or IPs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2018 doi:10.20944/preprints201802.0108.v1

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

5.4 Subjective Tests

Mean opinion score (MOS) test and preference test were performed simultaneously by 15 subjects given with 15 synthesized long utterances with lengths from 64 to 125 syllables (99 in average) for each prosody generation method. The feature combinations resulting in the smallest RMSEs for *BSL*/*QCset*/*PCset* in Table 5 were chosen to generate prosodic-acoustic features for speech synthesis by an HMM-based synthesizer [7-10]. There are three types of the proposed feature sets to be compared with the baseline (*BSL*): *QCset*, *PCset*, and *QCset+PCset*. As shown in Table 15, the best feature combination for the *BSL* is the combination of *BSL2* for lf0, *BSL3* for Dur, Eng, and Pau. The best combination for *QCset* is the one of *QC2* for lf0 and Pau, *QC1* for Dur, and *QC3* for Eng while the best combination for *PCset* is the one of *PC4* for lf0, *PC3* for Dur, and *PC2* for Eng and Pau. The feature sets for *QCset+PCset* are *QC2* for lf0, *PC3* for Dur, and *PC2* for Eng and Pau. Before listening to the synthesized utterances by *BSL* and the ones by the proposed method, subjects 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. It is found from Table 16 that proposed *QCset*, *PCset*, and *QCset+PCset* generally could yield slightly more natural speech than *BSL*. The synthesized utterances with prosody generated by *QCset+PCset* achieved the most significant MOS difference to *BSL*. These results again confirm the usefulness of the proposed PC and QC features.

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%

6. Conclusions and Future Works

This paper proposes two fully-automatic machine-extracted linguistic features from an unlimited text input for Mandarin prosody generation. One is the PC which measures the likelihood of inserting an MPM at a word boundary. Another is the QC which measures the likelihood of a word string to be 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 between the prosodic structures and PM types, and structures of QPs. The usefulness of the proposed PC and QC features in Mandarin prosody generation was proved by both objective and subjective tests. It is encouraging to see that the proposed features could improve the performances of Mandarin prosody generation. With the fast growth of deep learning technologies, in the near future, it is worthwhile to transplant CRF-based punctuation generation models to neural network-based models, e.g., long short-term memory recurrent neural network (LSTM-RNN) [48]. The neural network-based punctuation models can be easily integrated with the followed neural network-based prosody generator or speech synthesizer in the training phase. Under this integrated framework, it is also interesting to apply the transfer learning technique [49] to make a neural network learn prosody generation based on a neural network that generates punctuations.

Acknowledgments: This work is primarily supported by a grant from Chunghwa Telecom under the 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-. The authors deeply thank Prof. Yih-Ru Wang of NCTU, Hsinchu, Taiwan, for providing the NCTU Speech Lab Traditional Chinese Parser. The authors also want to thank Academia Sinica, Taiwan for providing the Treebank Corpus, the Academia Sinica Balanced Corpus of Modern Chinese V.4.0, and the online CKIP Parser.

Author Contributions: C.-Y. C wrote the paper, conceived and designed the experiments; Y.-P. H. and H.-Y. Y performed the experiments; Y.-P. H. analyzed the data; I.-B. L. and C.-M. P. contributed reagents/materials/analysis tools.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, A.-J.; Zu, Y.-Q.; Li, Z.-Q. A national database design and prosodic labeling for speech synthesis. In Proceedings of the Oriental Co-ordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA) Workshop, Taipei, Taiwan, 13-14 May 1999; pp. 13–16.
2. Li, A.-J.; Lin, M.-C. Speech corpus of Chinese discourse and the phonetic research. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Beijing, China, 16-20 October 2000; Vol. 4, pp. 13–18.
3. Cao, J.-F. Rhythm of spoken Chinese—Linguistic and paralinguistic evidences. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Beijing, China, 16-20 October 2000; Vol. 2, pp. 357–360.
4. Chen, S.-H.; Hwang, S.-H.; Wang, Y.-R. An RNN-based prosodic information synthesizer for Mandarin text-to-speech. *IEEE Trans. Speech Audio Process*. **1998**, 6, 226–239.
5. Chen, S.-H.; Lai, W.-H.; Wang, Y.-R. A statistics-based pitch contour model for Mandarin speech. *J. Acoust. Soc. Am.* **2005**, 117, 908–925.
6. Chen, S.-H.; Lai, W.-H.; Wang, Y.-R. A new duration modeling approach for Mandarin speech. *IEEE Trans. Speech Audio Process*. **2003**, 11, 308–320.
7. Tokuda, K.; Yoshimur, T.; Masuko, T.; Kobayashi, T.; Kitamura, T. 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, 5-9 June 2000; pp. 1315-1318.
8. Yoshimura, T. Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems. Ph.D. thesis, Nagoya Institute of Technology, Nagoya, Aichi, Japan, January 2002.
9. Zen, H.; Nose, T.; Yamagishi, J.; Sako, S.; Masuko, T.; Black, A.W.; Tokuda, K. The HMM-based speech synthesis system version 2.0. In Proceedings of the Sixth ISCA Workshop on Speech Synthesis (SSW6), Bonn, Germany, 22-24 August 2007; pp. 294-299.
10. The HTS working group, HTS-2.3 source code, and demonstrations. Available online: http://hts.sp.nitech.ac.jp/?Download (accessed on 26 January 2018).
11. Ostendorf, M.; Veilleux, N. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Comput. Linguist*. **1994**, 20, 27–52.
12. Peng, H.-J.; Chen C.-C.; Tseng, C.-Y.; Chen, K.-J. 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, 15-18 December 2004; pp. 173–176.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2018 doi:10.20944/preprints201802.0108.v1

13. Chu, M; Qian, Y. Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts. *Computat. Linguist. and Chinese Language Process.* **2001**, 6, 61-82.
14. Xu, D.-W.; Wang, H.-F.; Li, G.-H.; Kagoshima, T. Parsing hierarchical prosodic structure for Mandarin speech synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toulouse, France, 14-19 May 2006; Vol. 1, pp. 14–19.
15. Black, A. W.; Taylor, P. Assigning phrase breaks from part-of-speech sequences. In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Rhodes, Greece, 22-25 September 1997; pp. 995–998.
16. Sheng, Z.; Tao, J.-H.; Jiang, D.-L. Chinese prosodic phrasing with extended features. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong, 6-10 April 2003; Vol. 1, pp. 492–495.
17. Li, J.-F.; Hu, G.-P.; Wang, R.-H. Chinese prosody phrase break prediction based on maximum entropy model. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Jeju Island, Korea, 4-8 October 2004; pp. 729–732.
18. Riedi, M. 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, 18-21 September 1995; pp. 599 -602.
19. Sagisaka, Y. 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, 3-6 April 1990; pp. 325 -328.
20. Traber, C. F0 generation with a database of natural F0 patterns and with a neural network. In *Talking Machines: Theories, Models and Designs*; Bailly, G.; Benoit, C., Eds.; Elsevier: Amsterdam, 1992; pp. 287-304.
21. Scordilis, M. S.; Gowdy, J. N. Neural network based generation of fundamental frequency contours. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Glasgow, Scotland, 23-26 May 1989; pp. 219 -222.
22. Chen, G. P.; Bailly, G.; Liu, Q. F; Wang, R. H. A superposed prosodic model for Chinese text-to-speech synthesis. In Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, 15-18 December 2004; pp. 117-120.
23. Bailly, G.; Holm, B. SFC: A trainable prosodic model. *Speech Communication*. 2006, 46, 348-364.
24. Wen, M.-M.; Wang, M.-M.; Hirose, K.; Minematsu, N. Improved Mandarin segmental duration prediction with automatically extracted syntax features. In Proceedings of the 10th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 24-28 October 2010; pp. 621-624.
25. Hsia, C. C.; Wu, C. H.; Wu, J. Y. Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis. *IEEE Trans. Audio, Speech, and Language Process*. **2010**, 18(8), 1994-2003.
26. Wang, M.-M.; Wen, M.-M.; Hirose, K.; Minematsu, N. Improved Generation of Prosodic Features in HMM-based Speech Synthesis. In Proceedings of the Seventh ISCA Workshop on Speech Synthesis (SSW7), Kyoto, Japan, 22-24 September 2010; pp. 359-36.
27. Wang, M.-M.; Wen, M.-M.; Hirose, K.; Minematsu, N. Improved generation of fundamental frequency in HMM-based speech synthesis using generation process model. In Proceedings of the the Annual Conference of the International Speech Communication Association (INTERSPEECH), Makuhari, Chiba, Japan, 26-30 September 2010; pp. 2166-2169.
28. Chiang, C.-Y.; Wang, Y.-R.; Chen, S.-H. 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, 25-30 May 2012; pp. 4597-4600.
29. Hung, Y.-P.; Yeh, H.-Y.; Liao, I.-B.; Pan, C.-M.; Chiang, C.-Y. 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, 20-12 September 2014; pp. 1-5.
30. Chiang, C.-Y.; Hung, Y.-P.; Liou, G.-T.; Wang, Y.-R. 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, 17-20 October 2016; pp. 1–5.
31. Hung, Y.-P. Punctuation Generation Inspired Linguistic Features for Mandarin Prosody Generation. Master, National Taipei University, New Taipei City, Taiwan, 24 July 2015.

 doi:10.20944/preprints201802.0108.v1

32. Guo, Y.-Q.; Wang, H.-F.; Genabith J. V. A Linguistically Inspired Statistical Model for Chinese Punctuation Generation. *ACM Trans. on Asian Language Processing*. **2010**, 9(2).
33. Tseng, C.-Y. Mandarin speech prosody: issues, pitfalls and directions. In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland, 1-4 September 2003; pp. 2341-2344.
34. Available on http://www.aclclp.org.tw/use_asbc.php
35. Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the the Eighteenth International Conference on Machine Learning (ICML 2001), Williamstown, MA, USA, 28 June-1 July 2001; pp. 282-289.
36. CRF++: Yet Another CRF toolkit. Available online: https://taku910.github.io/crfpp/ (accessed on 26 January 2018)
37. Silverman, K.; Beckman, M.; Pitrelli, J.; Ostendorf, M.; Wightman, C.; Price, P.; Pierrehumbert, J.; Hirschberg, J. ToBI: A standard for labeling English prosody. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Banff, Alberta, Canada, 13-16 October 1992; pp. 867–870.
38. Taylor, P.-A. The tilt intonation model. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Sydney, Australia, 30 November-4 December 1998; pp. 1383–1386.
39. Li, A.-J. Chinese prosody and prosodic labeling of spontaneous speech. In Proceedings of the ISCA International Conference on Speech Prosody (Speech Prosody), Aix-en-Provence, France, 11-13 April 2002; pp. 39–46.
40. Chiang, C.-Y.; Chen, S.-H.; Yu, H.-M.; Wang, Y.-R. Unsupervised joint prosody labeling and modeling for Mandarin speech. *J. Acoust. Soc. Amer.* **2009**, 125(2), 1164–1183.
41. Chiang, C.-Y.; Chen, S.-H.; Wang, Y.-R. Advanced Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech and Its Application to Prosody Generation for TTS. In Proceedings of the the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brighton, UK, 6-10 September 2009; pp. 504-507.
42. Chen, S.-H.; Yang, J.-H. Yang; Chiang, C.-Y. Chiang; Liu, M.-C. Liu; Wang, Y.-R. A New Prosody-Assisted Mandarin ASR System. *IEEE Trans. Audio, Speech, and Language Process.* **2012**, 20(6), pp. 1669-1684.
43. The NCTU Speech Lab Traditional Chinese Parser. Available online: http://parser.speech.cm.nctu.edu.tw/ (accessed on 26 January 2018)
44. Lin, A.-H.; Wang, Y.-R.; Chen, S.-H. Traditional Chinese parser and language modeling for Mandarin ASR. 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, Idia, 25-17 November 2013; pp. 1-5.
45. Chen, K.-J.; Huang, C.-R. 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., 1993.
46. Chiang, C.-Y.; Yu, H.-M.; Wang, Y.-R.; Chen, S.-H. 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, 31 March-4 April 2008; pp. 4381-4384.
47. Chen, S.-H. Chen; Wang, Y.-R. Vector quantization of pitch information in Mandarin speech. *IEEE Trans. Commun*. **1990**, 38(9), 1317-1320
48. Hochreiter, S.; and Schmidhuber, J. Long short-term memory. *Neural Computation*. **1997**, 9(8), 1735–1780, 1997.
49. Bengio, Y. Deep Learning of Representations for Unsupervised and Transfer Learning. In Proceedings of the of ICML Workshop on Unsupervised and Transfer Learning, Bellevue, Washington, USA, 2 July 2012; pp. 17-36.