Improvements on Punctuation Generation Inspired Linguistic Features for Mandarin Prosody Generation

Chen-Yu Chiang[1], Yu-Ping Hung[1], Guan-Ting Liou[2], Yih-Ru Wang[2]

江振宇 洪宇平 劉冠廷 王逸如

[1]Dept. of Communication Engineering, National Taipei University, Taiwan

[2]Dept. of Electrical Engineering, National Chiao Tung University, Taiwan

Introduction

- Prosody generation serves as a function to map from linguistic features to prosodic-acoustic features
- Its performance generally depends on two factors: ability of the prosody prediction model and use of linguistic features
- This paper focuses on the use of linguistic features

Linguistic Features for Mandarin Prosody Generation (1/2)

1. ***Raw***: simply extracted from raw texts: PM, syllable position in a sentence can also be extracted
2. ***WordSeg***: extracted after the word segmentation: lexical word (LW) length, syllable position in a LW, and LW position in a sentence
3. ***WordPos***: part-of-speech (POS) of each LW
4. ***G2P***: generated by a grapheme to phone (G2P) process: important features characterizing properties of Mandarin prosody: tone and base-syllable type
5. ***BasePh***: generated by a base phrase chunker [1,2,15], including type of base phrase, length of base phrase in syllable/LW, and syllable/LW position in a base phrase
6. ***SynTree***: tree representation of grammar made from full syntactic parsing

Linguistic Features for Mandarin Prosody Generation (2/2)

- The sets *BasePh* and *SynTree* comprise higher level of syntactic features than shallow syntactic features (e.g. POS)
  - They generally could improve the performance of prosody generation
- The training/performance of the models for the feature sets *BasePh* and *SynTree* is usually confined by the size of available text corpora parsed with syntactic tree
  - labeling of syntactic tree and base phrase involves time-consuming human labors with linguistic expertise

The Previous Proposed Feature – Punctuation Confidence (PC)

- The PC measures likelihood of inserting major PMs at LW junctures into texts
- PC is produced by a conditional random field (CRF)-based automatic PM generation model given with PM-removed word/POS sequences
- The CRF model can be trained given with large text corpora without human labeling
- Generally, word junctures with higher PC are more likely to be inserted with pause breaks
- The effectiveness of the proposed approach was confirmed by the experiments on a 50K-syllable Mandarin speech corpus [1,6]

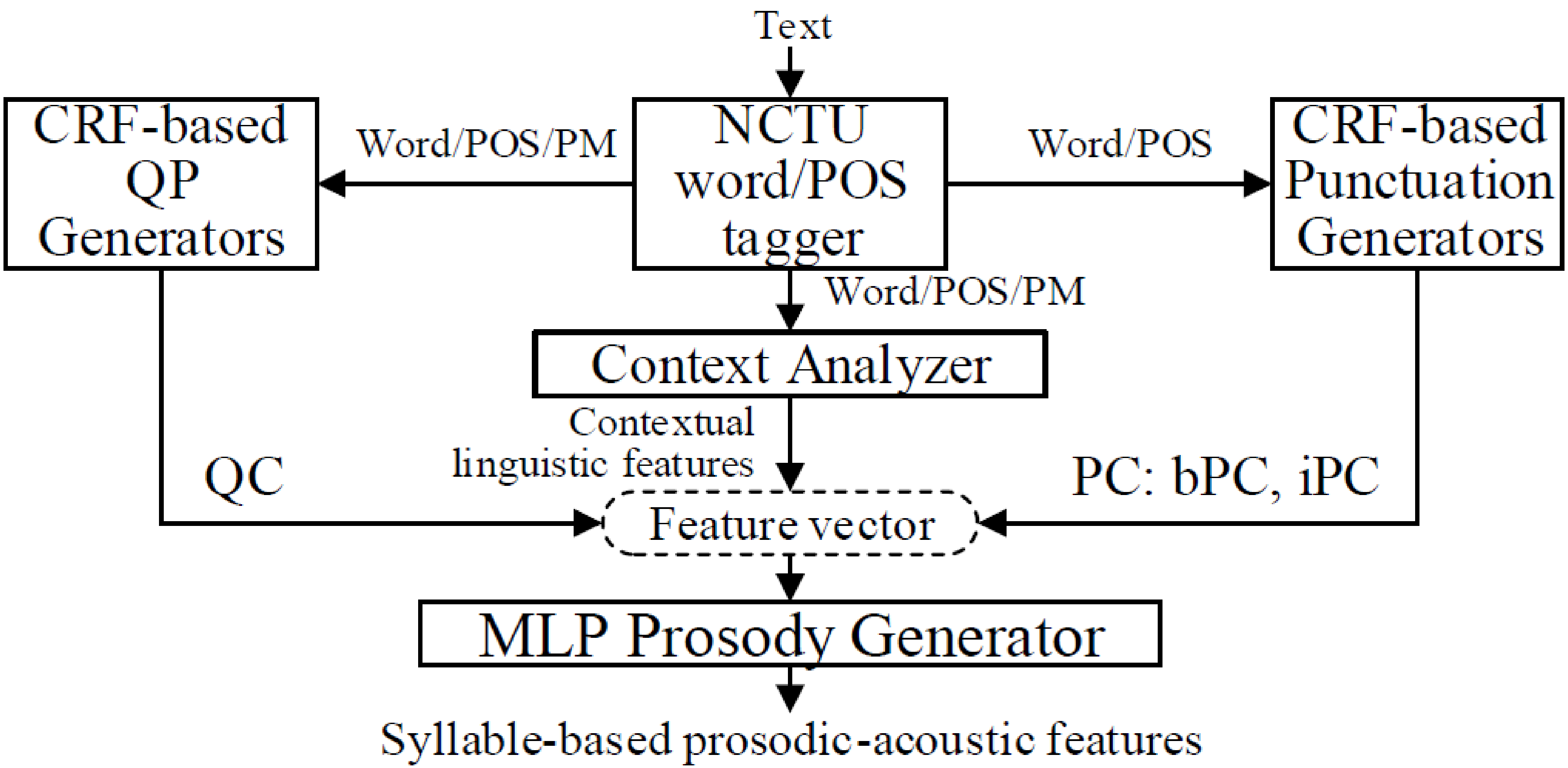
The Proposed Feature in this Study – the Improved PC (iPC)

- The iPC is a modified version the PC [6] (referred to as the basic PC, bPC, thereafter)
  - considers both insertion of major PM and structures of sentences
  - sentence structures are highly correlated with prosodic phrase (PPh) structures → the iPC may give a better prediction of prosodic phrase structures

The Proposed Feature in this Study – the Quotation Confidence (QC)

- The QC is generated by a CRF model that predicts structures of quoted word strings (i.e. quoted phrase, QP) from word/POS sequences
- The QC can be regarded as a statistical linguistic feature to measure likelihood of word strings being quoted by a left bracket and a right bracket
  - Words in the brackets are closely related to constitute a larger unit with complex or more specific meanings for human language understanding
  - less prosodic breaks are inserted within a quoted word string →emphasized with some variations in prosodic-acoustic features

The Design of the Experiment



The Generation of PC

CRF-based punctuation generator:

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

PM sequence ($\mathbf{Y}$); linguistic feature sequence: word/POS ($\mathbf{X}$)

Template function:

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

Y_t: type of PM between t-th and (t+1)-th LWs

y_k: k-th possible tag (i.e. PM type)

PM sequence can be predicted by

$$Y_1^*,\ldots,Y_T^*=\arg\max_{\mathbf{Y}} P(\mathbf{Y}\mid\mathbf{X}) \qquad (3)$$

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

$\varphi_{t,k}$: The PC of k-th PM type for each t-th LW juncture

The Design of Prediction Targets

Table 1. Targets for iPCs

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

(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/E2y_0 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 2: *(a) original word/PM sequence. The tag labelings for the training of bPC (b), iPCs (c), and iPCf (d).*

The Experiment of PC Generation and Evidence (1/2)

- The feature templates: contextual LW, POSs, length of LW, and the combinations of the above features
- The CRF models were trained by the Acdamia Sinica Balanced Corpus (ASBC) [18] training set with 6,625,277 words and the best feature templates were tuned by the results on the test set with 2,817,785 words
- The precision/recall of PM generations on the test set for bPC, iPCf, and iPCs are respectively 94.1%/93.0%, 96.7%/95.9%, and 95.5%/95.3%

The Experiment of PC Generation and Evidence (2/2)

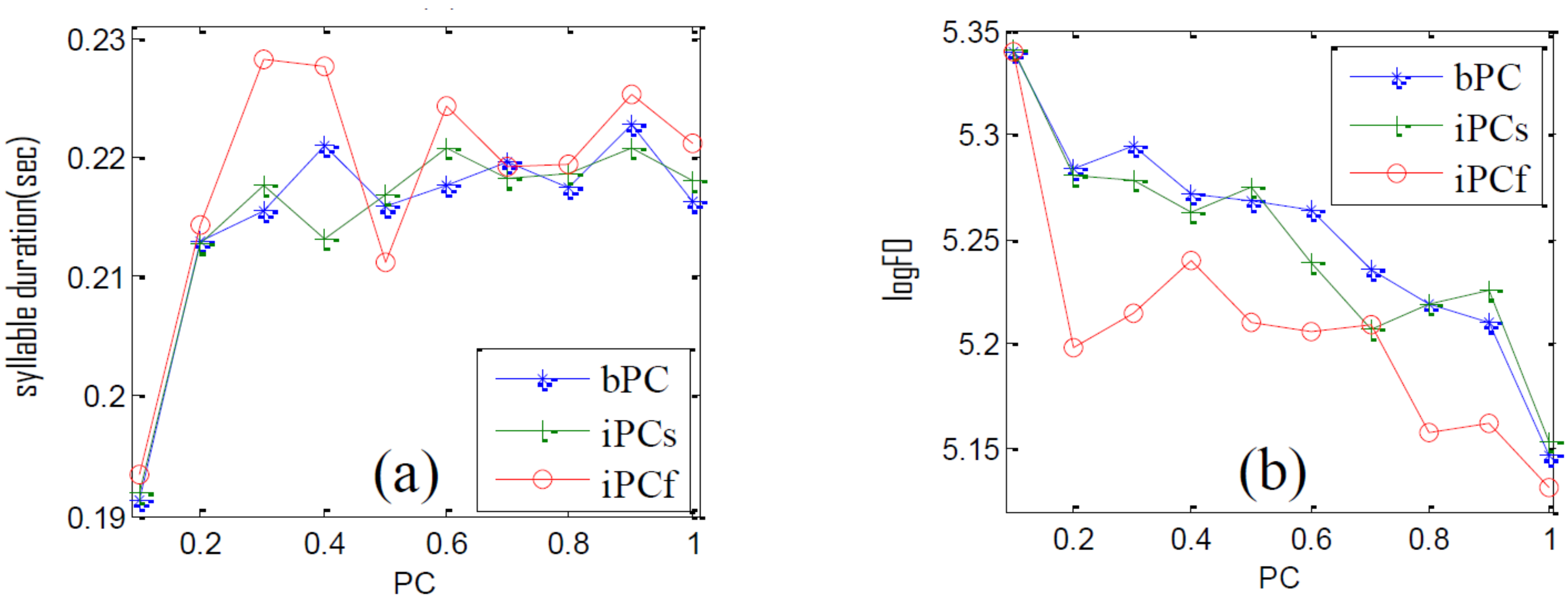


Figure 3: *(a) bPC, iPCs, and iPCf for the tag E1 (predicted sentence boundary) vs. average syllable durations, and (b) bPC, iPCs, and iPCf for E1 vs. average syllable logF0 mean.*

Analysis on Quotation

Table 2. *Categorization of 26 Chinese quotation marks*

type	1				2				3				4	
quote	（	）	(	)	｛	｝	{	}	〔	〕	﹝	﹞	「	」
type	4		5		6		7		8		9			
quote	『	』	〈	〉	【	】	《	》	“	”	〝	〞		

- <u>**Type 1**</u> - （ ） : as enumeration.
- <u>**Type 2**</u> - ｛ ｝ : titles of books or article
- <u>**Type 3**</u> - 〔 〕 : captions of articles
- <u>**Type 4**</u> - 「 」 and 『 』 : contributes most samples (66%) for the QP predictions → word chunks or base phrases.
- <u>**Types 5, 6 and 7**</u> - 〈 〉 【 】 《 》 : similar to the Type 2
- <u>**Type 8**</u> - “”: proper nouns, popular phrases, or sentence-like unit
- <u>**Type 9**</u> - 〝〞 : similar to the type 4

The Design of Prediction Targets

Table 3. *Tag format for labeling of target QP for bQC.*

Length in LW	Tag format	Length	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

(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 4: *Original word/PM tokens (a), and exemplar tag labelings for bQC training (b) and the sQC training (c).*

The Experiment of QP Generation and Evidence (1/2)

- Only 0.69% of the ASBC text corpus contributed instances of QPs
- To make the CRF models 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
- The features for the QC training are words and POSs
- The precision and recall for predicting bQC are respectively 60.6% and 39.0%
- The precision and recall for sQC are respectively 55.6% and 52.2%

The Experiment of QP Generation and Evidence (2/2)

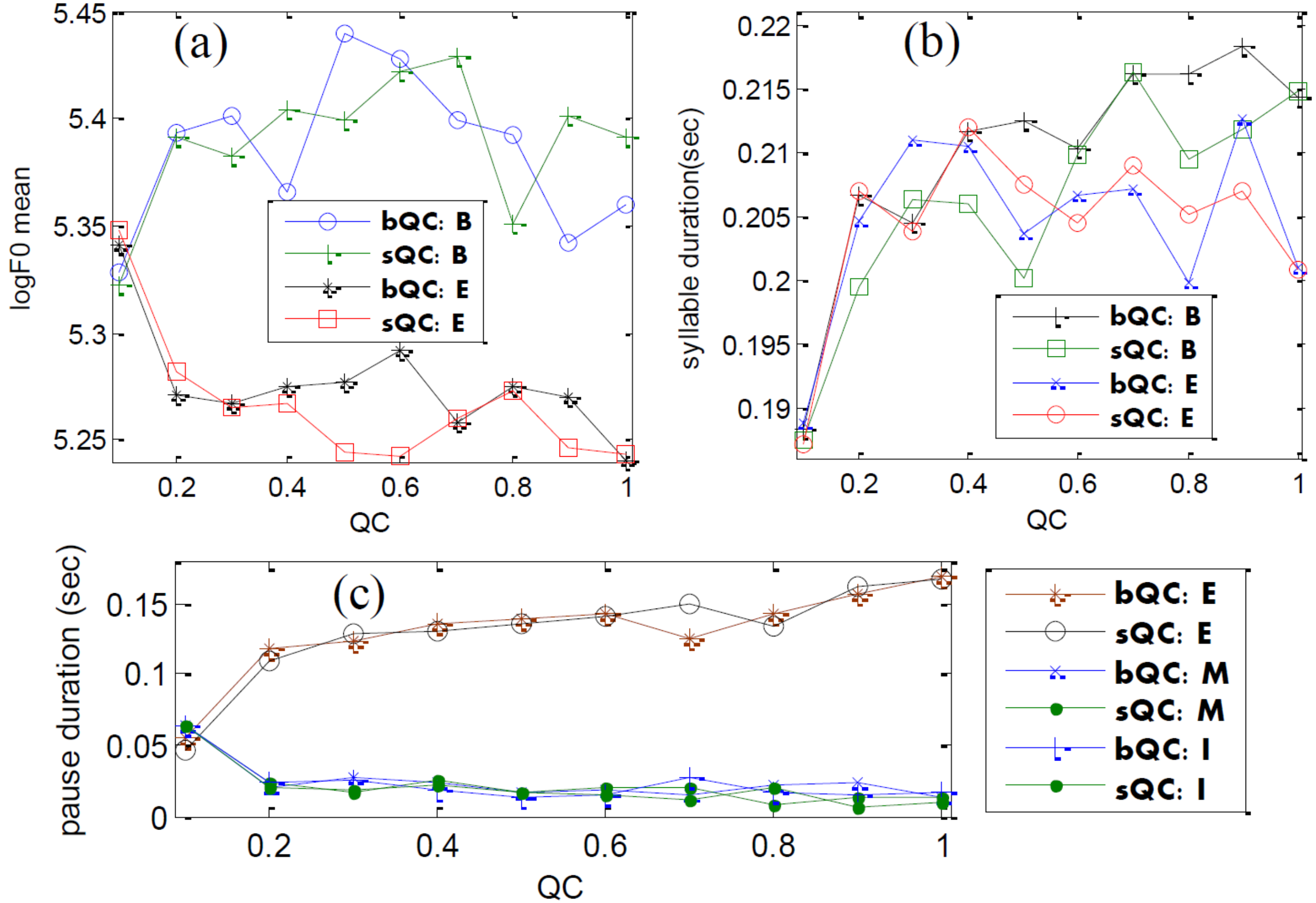


Figure 5: *bQCs and sQCs for the tags B, M, I, and/or E vs. (a) average syllable logF0 means, (b) average syllable duration, and (c) average pause durations.*

Experiments of Prosody Generation (1/2)

- The treebank speech corpus was used to evaluate the usability of the PC and QC in the prosody generation
  - a training set of 301 utterances (41,317 syllables), a development set of 75 utterances (10,551 syllables), and a test set of 44 utterances (3,898 syllables)
- The four independent MLPs were trained to predict syllable logF0 contour (lf0), syllable duration (Dur), syllable energy level (Eng), and inter-syllable pause duration (Pau)

Experiments of Prosody Generation (2/2)

- 28 and 67 features in the set *Raw* and *G2P*, respectively
- 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
- The feature sets bPC, iPCs, iPCf, bQC, and sQC respectively are composed of 2, 11, 22, 8, and 19 numerical features ($\varphi_{t,k}(\mathbf{X})$ for some *k*-th target tags of PC or QC at *t*-th word)

Objective/Subjective Tests

Table 4. *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***
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
PCset	*PC1= BSL3+bPC*	.185	38.33	3.48	58.29
	PC2= BSL3+iPCs	.175	37.82	***3.43***	***57.29***
	PC3= BSL3+iPCf	.174	***37.34***	3.47	58.72
	PC4= BSL2+iPCs	***.173***	38.39	3.46	63.93
	PC5= BSL2+iPcf	.174	38.05	3.48	62.56

Table 5. *Preferences (%) and MOSs (numbers in brackets ± standard deviation) for the two subjective tests.*

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

Conclusions and Feature Works

- The effectiveness of the proposed iPC and QC features were proved to improve the performances of Mandarin prosody generation by both the objective and subjective tests
- In the future, we will investigate the usability of the iPC and QCs in construction of an HMM-based speech synthesizer. The prediction capability by combining features of the iPC and QCs will also be explored

Thank you for your attention

Contact:

江振宇 (Chen-Yu CHIANG)

cychiang@mail.ntpu.edu.tw

Homepage:

http://cychiang.tw/

Acknowledgements

This work was supported the MOST of Taiwan under Contract No. NSC-102-2221-E-305-005-MY3.

References

[1] Yu-Ping Hung, Han-Yun Yeh, I-Bin Liao, Chen-Ming Pan, and Chen-Yu Chiang, "An investigation on linguistic features for Mandarin prosody generation," in Proc. OCOCOSDA'2014, pp.1-5, 10-12 Sept. 2014

[2] Z. Sheng, J.-H. Tao, and D.-L. Jiang, "Chinese prosodic phrasing with extended features," ICASSP'2003, vol.1, pp.492–495.

[3] Miaomiao Wen; Miaomiao Wang; Hirose, K.; Minematsu, N., "Improved Mandarin segmental duration prediction with automatically extracted syntax features," Signal Processing (ICSP), 2010 IEEE 10th International Conference on, vol., no., pp.621,624, 24-28 Oct. 2010.

[4] Miaomiao Wang, Miaomiao Wen, Keikichi Hirose, Nobuaki Minematsu, "Improved Generation of Prosodic Features in HMM-based Speech Synthesis," In SSW7, pages 359-362

[5] M.Wang, M. Wen, K. Hirose, and N. Minematsu, "Improved generation of fundamental frequency in HMM-based speech synthesis using generation process model", in Proc. INTERSPEECH, 2010, pp.2166-2169.

[6] Chen-Yu Chiang, Yih-Ru Wang, Sin-Horng Chen, "Punctuation generation inspired linguistic features for Mandarin prosodic boundary prediction," in Proc. ICASSP-2012, pp. 4597- 4600, 25-30 March 2012

[7] H.-J. Peng, C.-C. Chen, C.-Y. Tseng, and K.-J. Chen, "Predicting prosodic words from lexical words—A first step towards predicting prosody from text," ISCSLP 2004, pp.173–176.

[8] M. Chu and Y. Qian, "Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts," Computat. Linguist. and Chinese Language Processing, 6, pp.61-82 (2001)

[9] D.-W. Xu, H.-F. Wang, G.-H. Li, and T. Kagoshima, "Parsing hierarchical prosodic structure for Mandarin speech synthesis," ICASSP'2006, vol.1, pp.14–19.

[10] J.-F. Li, G.-P. Hu, and R.-H. Wang, "Chinese prosody phrase break prediction based on maximum entropy model," Interspeech 2004, pp. 729–732.

[11] G. P. Chen, G. Bailly, Q. F. Liu and R. H. Wang, "A superposed prosodic model for Chinese text-to-speech synthesis," in Proc. of ISCSLP 2004, pp. 117-120, Dec. 2004.

[12] Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang, "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. 6, pp. 226-239, May 1998

[13] C. C. Hsia, C. H. Wu, and J. Y. Wu, "Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis," IEEE Trans. Audio, Speech, and Language Processing, vol. 18, no. 8, pp.1994-2003, August 2010.

[14] Chen-Yu Chiang, Sin-Horng Chen and Yih-Ru Wang, "Advanced Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech and Its Application to Prosody Generation for TTS," in Proc. Interspeech 2009, Brighton, UK, Sept. 2009, pp. 504-507.

[15] The NCTU Speech Lab Traditional Chinese Parser, available at http://parser.speech.cm.nctu.edu.tw/
[16] A.-H. Lin, Y.-R. Wang, and S.-H. Chen, "Traditional Chinese parser and language modeling for Mandarin ASR," In Proc. O'COCOSDA'13, Gurgaon, Idia, 25-27 Nov. 2013, pp. 1-5.
[17] S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," IEEE Trans. Commun., vol. 38, no. 9, pp. 1317-1320,
[18] Available on http://www.aclclp.org.tw/use_asbc.php
[19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Proc. of ICML, pp.282-289, 2001
[20] CRF++: Yet Another CRF toolkit, available at http://crfpp.googlecode.com/svn/trunk/doc/index.html
[21] C.-R. Huang, K.-J. Chen, F.-Y. Chen, Z.-M. Gao, and K.-Y. Chen, "Sinica Treebank: Design criteria, annotation guidelines, and pn-line interface," Proceedings of the Second Chinese Language Processing Workshop 2000, pp.29–37.
[22] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech," J. Acoust. Soc. Am. 125, No. 2, pp. 1164-1183 (2009).
[23] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", in Proc. ICASSP, Jun. 2000, pp.1315-1318.
[24] T. Yoshimura, Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems, Ph.D. thesis, Nagoya Institute of Technology, Jan. 2002.
[25] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, "The HMM-based speech synthesis system version 2.0," Proc. of ISCA SSW6, Bonn, Germany, Aug. 2007.
[26] The HTS working group, HTS-2.2 source code and demonstrations, available: http://hts.sp.nitech.ac.jp/?Download