Improvements on Punctuation Generation
Inspired Linguistic Features for Mandarin Prosody Generation

Chen-Yu Chiang¹, Yu-Ping Hung¹, Guan-Ting Liou², Yih-Ru Wang²

¹Dept. of Communication Engineering, National Taipei University, Taiwan
²Dept. 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

- CRF-based QP Generators
- NCTU word/POS tagger
- CRF-based Punctuation Generators

- Text
- Word/POS
- Word/POS

- Context Analyzer
  - QC
  - Contextual linguistic features
  - Feature vector
  - PC: bPC, iPC

- MLP Prosody Generator
  - Syllable-based prosodic-acoustic features
The Generation of PC

CRF-based punctuation generator:

\[ P(Y \mid X) = \frac{1}{N(X)} \exp \left( \sum_{i=1}^{T} \sum_{i=1}^{I} \lambda_i f_i (Y_i = y, Y_{i-1}, X) \right) \]  \hspace{1cm} (1)

Template function:

\[ f_i (Y_i = y, Y_{i-1}, X) = \begin{cases} 1, & \text{if } X = h_j \text{ is satisfied and } y = y_k \\ 0, & \text{otherwise} \end{cases} \]  \hspace{1cm} (2)

PM sequence can be predicted by

\[ Y_1^*, ..., Y_T^* = \arg \max_Y P(Y \mid X) \]  \hspace{1cm} (3)

\[ \varphi_{t,k}(X) = P(Y_t = y_k \mid X) \]  \hspace{1cm} (4)

The PC of \( k \)-th PM type for each \( t \)-th LW juncture
The Design of Prediction Targets

Table 1. Targets for iPCs

<table>
<thead>
<tr>
<th>Target Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1: 1\textsuperscript{st} LW</td>
<td>1\textsuperscript{st} LW if sentence length in LW is odd and less than 9</td>
<td>E4: 4\textsuperscript{th} last LW</td>
</tr>
<tr>
<td>B2: 2\textsuperscript{nd} LW</td>
<td>2\textsuperscript{nd} LW if sentence length in LW is equal or more than 9</td>
<td>E3: 3\textsuperscript{rd} last LW</td>
</tr>
<tr>
<td>B3: 3\textsuperscript{rd} LW</td>
<td>intermediate LW if sentence length in LW is odd and less than 9</td>
<td>E2: 2\textsuperscript{nd} last LW</td>
</tr>
<tr>
<td>B4: 4\textsuperscript{th} LW</td>
<td>intermediate LW if sentence length in LW is equal or more than 9</td>
<td>E1: 1\textsuperscript{st} last LW</td>
</tr>
<tr>
<td>I: Intermediate LW</td>
<td></td>
<td>S: single LW</td>
</tr>
</tbody>
</table>

(a) 望遠鏡可以用來看天上明亮閃爍的星星，或是水際的野鳥，也可以用來看人。

(b) 望遠鏡\textsubscript{y0} 可以/\textsubscript{y0} 用來/\textsubscript{y0} 看/\textsubscript{y0} 天/\textsubscript{y0} 上/\textsubscript{y0} 明亮/\textsubscript{y0} 閃爍/\textsubscript{y0} 的/\textsubscript{y0} 星星/\textsubscript{y1} 或是/\textsubscript{y0} 水際/\textsubscript{y0} 的/\textsubscript{y0} 野鳥/\textsubscript{y1} 也/\textsubscript{y0} 可以/\textsubscript{y0} 用來/\textsubscript{y0} 看/\textsubscript{y0} 人/\textsubscript{y1}

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

(d) Instance 1: 望遠鏡/\textsubscript{E1} 可以/\textsubscript{E2} 用來/\textsubscript{E3} 看/\textsubscript{E4} 天/\textsubscript{M} 上/\textsubscript{M} 明亮/\textsubscript{E4} 閃爍/\textsubscript{E3} 的/\textsubscript{E2} 星星/\textsubscript{E1} 或是/\textsubscript{B1} 水際/\textsubscript{B2} 的/\textsubscript{E2} 野鳥/\textsubscript{E1}

Instance 2: 或是/\textsubscript{B1} 水際/\textsubscript{B2} 的/\textsubscript{E2} 野鳥/\textsubscript{E1} 也/\textsubscript{B1} 可以/\textsubscript{B2} 用來/\textsubscript{I} 看/\textsubscript{E2} 人/\textsubscript{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 Acadamia 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

<table>
<thead>
<tr>
<th>type</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>quote</td>
<td>( )</td>
<td>{ }</td>
<td>{ }</td>
<td>「」</td>
</tr>
<tr>
<td>type</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>quote</td>
<td>『』</td>
<td>〈 〉</td>
<td>【 】</td>
<td>《 》</td>
</tr>
</tbody>
</table>

• **Type 1** - ( ) : as enumeration.

• **Type 2** - { } : titles of books or article

• **Type 3** - [ ] : captions of articles

• **Type 4** - 「」 and 『』 : contributes most samples (66%) for the QP predictions $\rightarrow$ word chunks or base phrases.

• **Types 5, 6 and 7** - 〈 〉【 】《 》: similar to the Type 2

• **Type 8** - “”: proper nouns, popular phrases, or sentence-like unit

• **Type 9** - “” : similar to the type 4
The Design of Prediction Targets

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

<table>
<thead>
<tr>
<th>Length in LW</th>
<th>Tag format</th>
<th>Length</th>
<th>Tag format</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S</td>
<td>4</td>
<td>B B2 M E</td>
</tr>
<tr>
<td>2</td>
<td>B E</td>
<td>5</td>
<td>B B2 M M E</td>
</tr>
<tr>
<td>3</td>
<td>B I E</td>
<td>6</td>
<td>B B2 B3 M M E</td>
</tr>
</tbody>
</table>

(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 ($\varphi_{t,k}(X)$ 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.

<table>
<thead>
<tr>
<th>Feature set combinations</th>
<th>f0(logHz)</th>
<th>Dur(ms)</th>
<th>Eng(dB)</th>
<th>Pau(ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BSL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSL1 = Raw+G2P</td>
<td>.191</td>
<td>43.77</td>
<td>3.72</td>
<td>71.73</td>
</tr>
<tr>
<td>BSL2 = BSL1+WordSeg</td>
<td><strong>.182</strong></td>
<td>39.93</td>
<td>3.53</td>
<td>64.62</td>
</tr>
<tr>
<td>BSL3 = BSL2+WordPos</td>
<td>.186</td>
<td><strong>39.23</strong></td>
<td><strong>3.50</strong></td>
<td><strong>59.56</strong></td>
</tr>
<tr>
<td><strong>QCset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QC1 = BSL3+bQC</td>
<td>.170</td>
<td>37.70</td>
<td>3.52</td>
<td>58.66</td>
</tr>
<tr>
<td>QC2 = BSL3+sQC</td>
<td><strong>.169</strong></td>
<td>37.83</td>
<td>3.52</td>
<td><strong>57.95</strong></td>
</tr>
<tr>
<td>QC3 = BSL2+bQC</td>
<td>.176</td>
<td>39.83</td>
<td><strong>3.44</strong></td>
<td>64.50</td>
</tr>
<tr>
<td>QC4 = BSL2+sQC</td>
<td>.172</td>
<td>39.30</td>
<td>3.54</td>
<td>63.33</td>
</tr>
<tr>
<td><strong>PCset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC1 = BSL3+bPC</td>
<td>.185</td>
<td>38.33</td>
<td>3.48</td>
<td>58.29</td>
</tr>
<tr>
<td>PC2 = BSL3+iPCs</td>
<td>.175</td>
<td>37.82</td>
<td><strong>3.43</strong></td>
<td><strong>57.29</strong></td>
</tr>
<tr>
<td>PC3 = BSL3+iPCf</td>
<td>.174</td>
<td><strong>37.34</strong></td>
<td>3.47</td>
<td>58.72</td>
</tr>
<tr>
<td>PC4 = BSL2+iPCs</td>
<td><strong>.173</strong></td>
<td>38.39</td>
<td>3.46</td>
<td>63.93</td>
</tr>
<tr>
<td>PC5 = BSL2+iPcf</td>
<td>.174</td>
<td>38.05</td>
<td>3.48</td>
<td>62.56</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>the proposed sets</th>
<th>the proposed set vs. BSL</th>
<th>No prefer.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>QCset</strong></td>
<td>34% (3.45 ± 0.42)</td>
<td>25% (3.40 ± 0.45)</td>
</tr>
<tr>
<td><strong>PCset</strong></td>
<td>37% (3.55 ± 0.41)</td>
<td>21% (3.34 ± 0.48)</td>
</tr>
<tr>
<td><strong>QCset+PCset</strong></td>
<td>38% (3.57 ± 0.41)</td>
<td>22% (3.29 ± 0.48)</td>
</tr>
</tbody>
</table>
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


[26] The HTS working group. HTS-2.2 source code and demonstrations, available: http://hts.sp.nitech.ac.jp/?Download