FEATURE-BASED ADAPTATION FOR SPEAKING STYLE SYNTHESIS

Xixin Wu, Lifa Sun, Shiyin Kang, Songxiang Liu, Zhiyong Wu, Xunying Liu and Helen Meng

[wwxx, Ifsun, sxliu, xyliu, hmmeng]@se.cuhk.edu.hk, zywu@sz.tsinghua.edu.cn, shiyinkang@tencent.com

The Chinese University of Hong Kong, Tsinghua University, Tencent AI Lab

1. Introduction

Objective
- Adapt speech synthesizer trained on declarative style data to synthesize interrogative style
- Interrogative data generally sparse

Motivation
- Previous work mainly consider utterance-level mismatch between source and target voices and lack modeling of local context-level mismatch
- Leverage frame-level features encoding context characteristics for interrogative style synthesis

Approach
- Feature-based adaptation from declarative to interrogative style using:
  (1) Interrogative style bottleneck features (BNFs)
  (2) Style difference residual features (RFs)

2. Style Adaptation Frameworks

Basic Model-based Adaptation Framework
- [Fan 2015, Zheng 2017]
- Train large network in ample declarative style data
- Adapt only top-layer parameters with limited interrogative style data

Proposed Feature-based Adaptation Framework
- Inject style features in the top layers
- Adapt top layers only

3. Style Features

Interrogative Style Bottleneck Features (BNFs)
- Stage 1: Train interrogative bottleneck DNN (bDNN)
- Stage 2: Extract bottleneck layer outputs as BNFs

Style Difference Residual Features (RFs)
- Stage 1: Train DNN with ample declarative data (decDNN)
- Stage 2: Obtain declarative-interrogative acoustic difference features to train rDNN
- Stage 3: Extract rDNN outputs as RFs

4. Experiments (I)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>One female Mandarin native speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data</td>
<td>5000 utterances with declarative style (~5 hrs)</td>
</tr>
<tr>
<td>Adaptation Data</td>
<td>400 utterances with interrogative style (~20 mins)</td>
</tr>
<tr>
<td>Testing Data</td>
<td>84 utterances with interrogative style (~5 mins)</td>
</tr>
</tbody>
</table>

6. Conclusions

- Feature-based adaptation for DNN-based speech synthesis
- Frame-level style features:
  (1) Interrogative style bottleneck features
  (2) Style difference residual features
- Frame-level style features effectively adapts for cross-style synthesis

7. Acknowledgement

This work is partially supported by National Natural Science Foundation of China-Research Grants Council of Hong Kong (NSFC-RGC) joint fund (6151166002, N_CUHK404/15).

Some samples are available in “http://www1.se.cuhk.edu.hk/~wuxx/ICASSP18/styleadapt.html”