FEATURE-BASED ADAPTATION FOR SPEAKING STYLE SYNTHESIS

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

{wuxx, lfsun, 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)

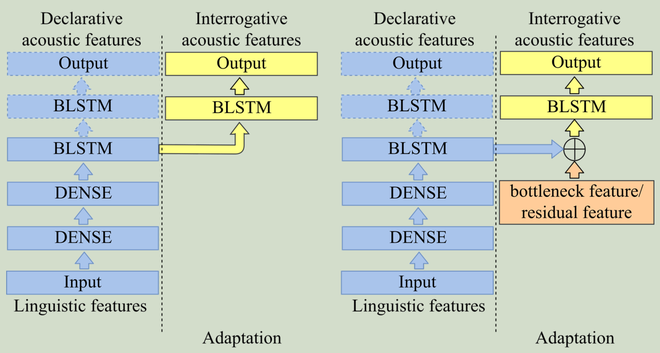
2. Style Adaptation Frameworks

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



(a) Basic framework (b) Proposed framework

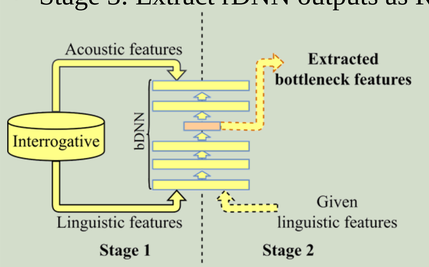
3. Style Features

Interrogative Style Bottleneck Features (BNFs)

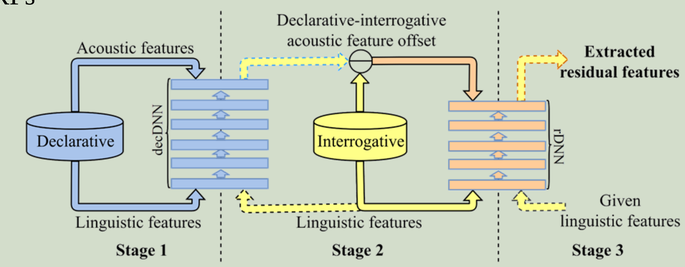
- Stage 1: Train interrogative bottleneck DNN (bDNN)
- Stage 2: Extract bottleneck layer outputs as BNFs

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



(c) BNFs extraction procedure



(d) RFs extraction procedure

4. Experiments (I)

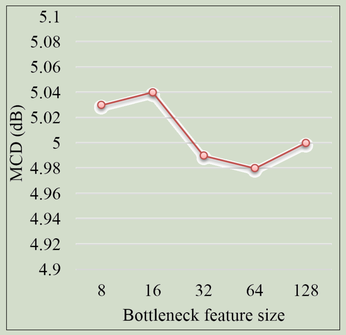
Corpus	***One female Mandarin native speaker***
Training Data	***5000 utterances with declarative style (~5 hrs)***
Adaptation Data	***400 utterances with interrogative style (~20 mins)***
Testing Data	***84 utterances with interrogative style (~5mins)***

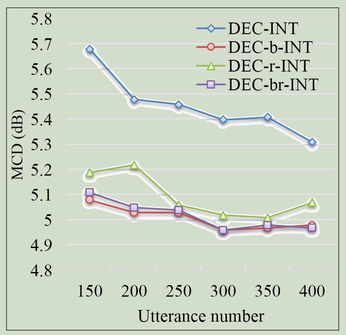
Six Systems

- INT: Trained with interrogative data
- DEC: Trained with declarative data
- DEC-INT: Adapt top layers of DEC with interrogative data
- DEC-b-INT: Adaptation with BNFs
- DEC-r-INT: Adaptation with RFs
- DEC-br-INT: Adaptation with concatenation of BNFs and RFs

Table 1. Objective evaluation of 6 systems.

Systems	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	V/UV Error Rate (%)
INT	6.02	4.81	30.01	12.11
DEC	5.63	4.49	30.21	7.74
DEC-INT	5.31	4.42	28.19	7.15
DEC-b-INT	4.98	**4.29**	27.41	**6.83**
DEC-r-INT	5.07	4.32	27.56	7.07
DEC-br-INT	**4.97**	**4.29**	**27.08**	6.84

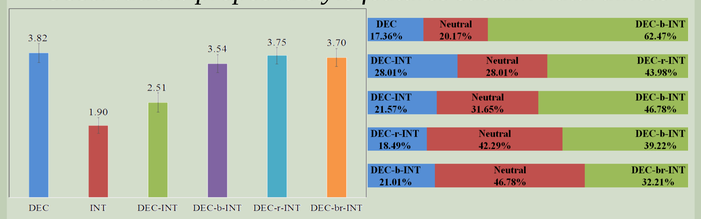




5. Experiments (II)

Mean Opinion Score (MOS) on Naturalness

- 20 listeners, 17 samples/system, 5-point scale
- *Observation: proposed style features sustain naturalness*

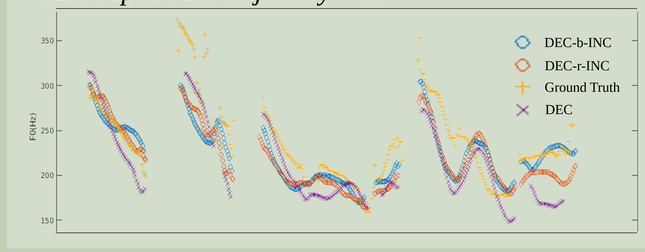


AB Preference Test on Similarity

- 20 listeners, 17 samples/system
- Provides an interrogative style preference choice: A, B or Neutral
- *Style features significantly enhance adaptation*
- *BNFs and RFs combination is best*

F0 Analysis

- *RFs capture F0 trajectory well*



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 (61531166002, N_CUHK404/15).

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