perception
sensing
instrumentation
lab

TEXAS A&M
UNIVERSITY

Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams

Guanlong Zhao, Shaojin Ding, and Ricardo Gutierrez-Osuna
Presented by Guanlong Zhao (gzhao@tamu.edu)

Department of Computer Science and Engineering
Texas A&M University, U.S.

*Work funded by NSF

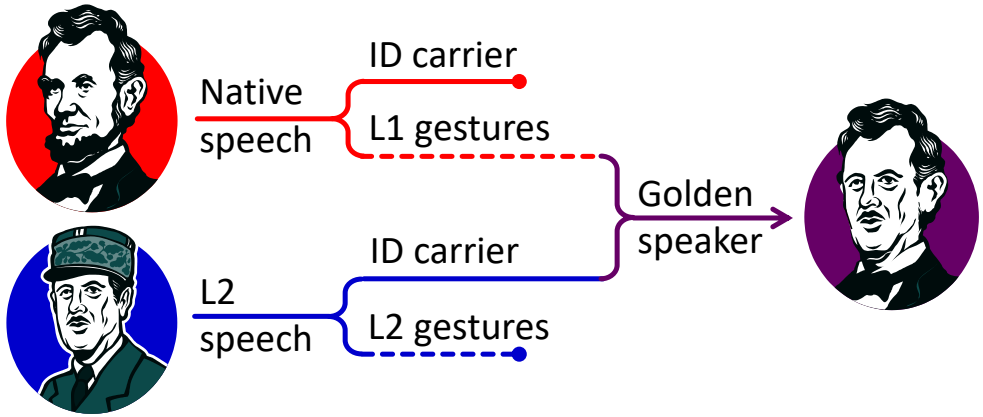
Introduction

Foreign accent conversion

– Create a new voice that has the voice quality of a given non-native speaker and the pronunciation patterns of a native speaker

Challenge

– Divide the speech signal into accent-related cues and voice quality



Related work

Prior approaches on accent conversion

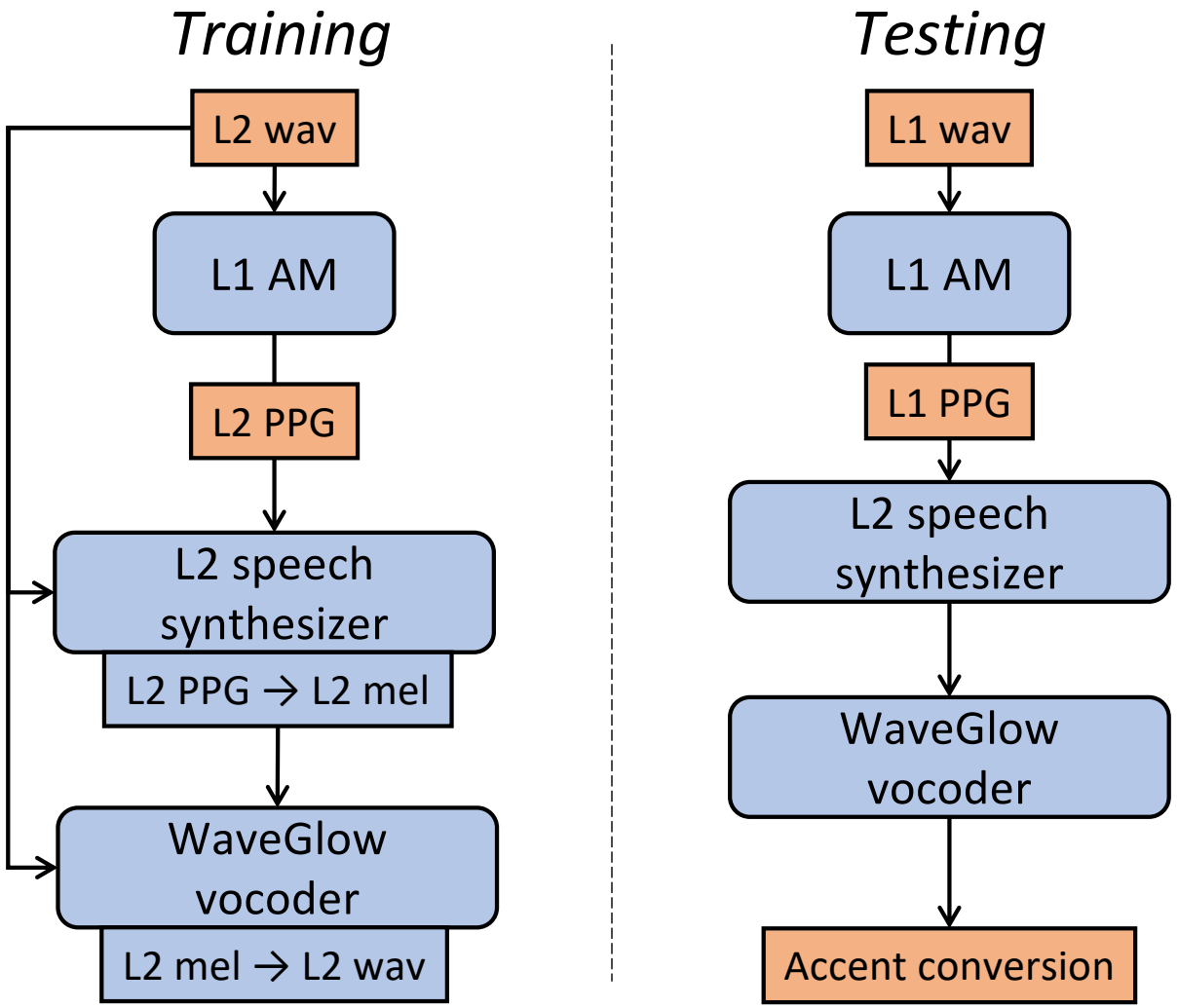
- Acoustics [Aryal 2015, Zhao 2018]
- Articulatory [Aryal 2016]

Phonetic Posteriorgrams (PPGs)

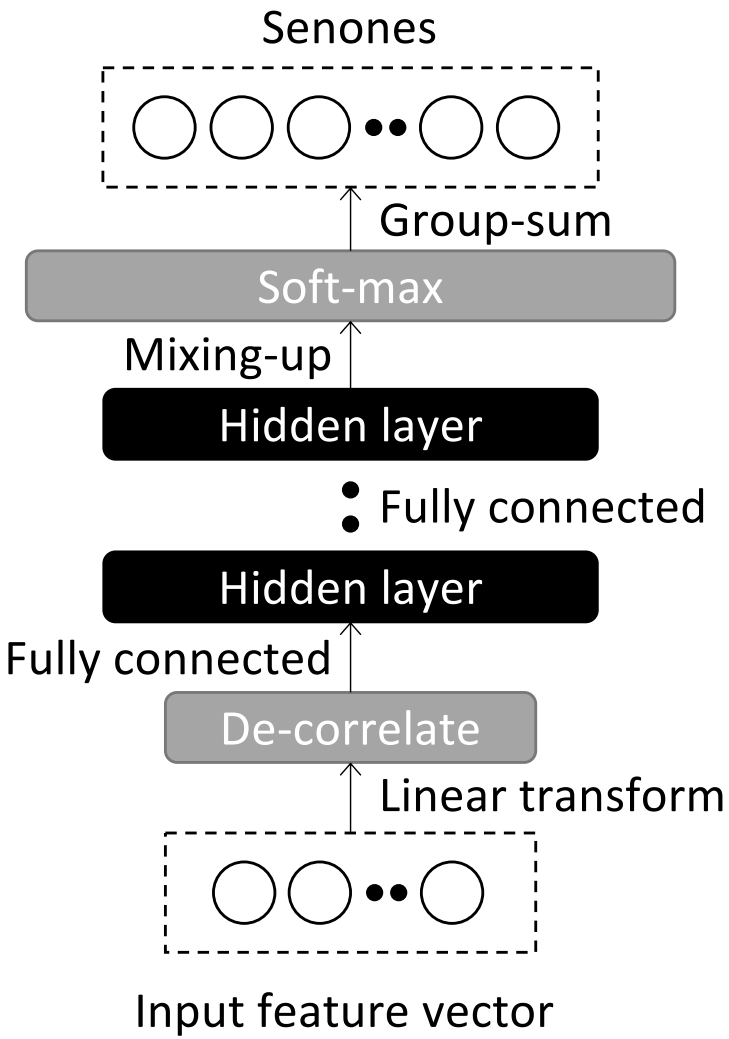
- Spoken term detection [Hazen 2009]
- Mispronunciation detection [Lee 2013]
- Personalized TTS [Sun 2016]
- Voice conversion [Xie 2016, Zhou 2016]

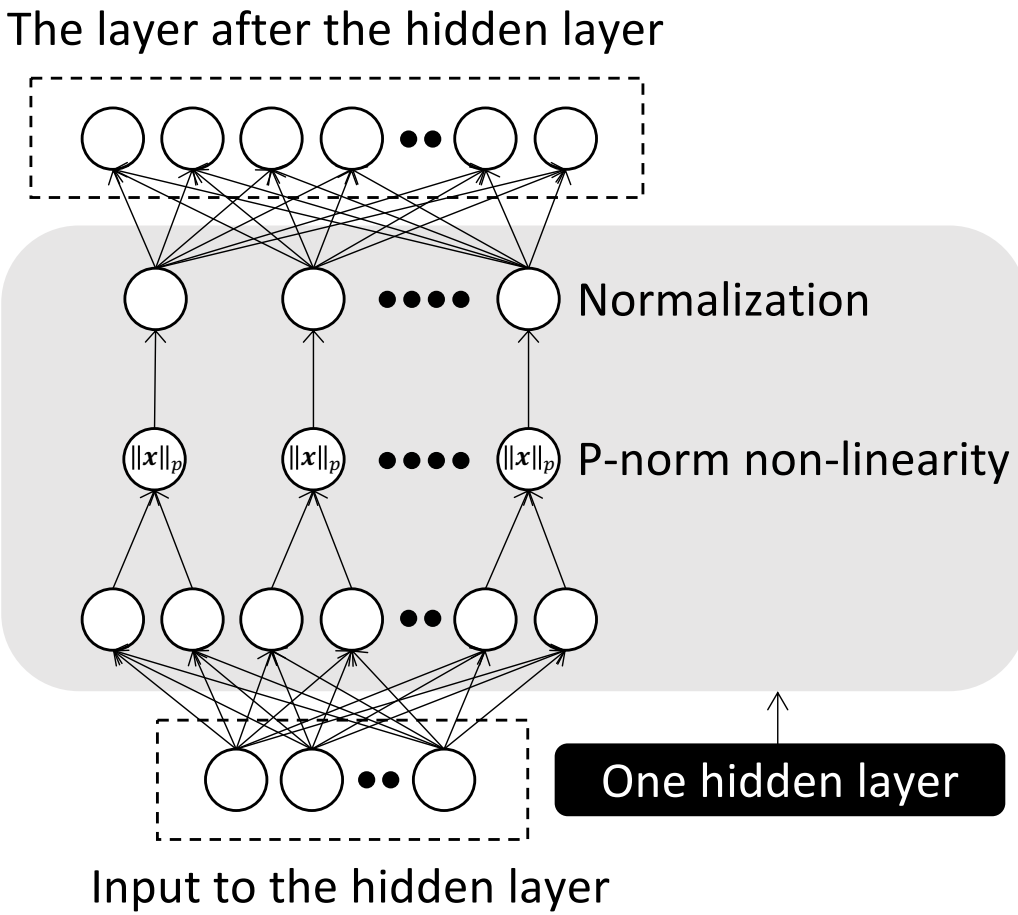
Aryal, Sandesh, and Ricardo Gutierrez-Osuna. "Can voice conversion be used to reduce non-native accents?." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.
Zhao, Guanlong, et al. "Accent conversion using phonetic posteriorgrams." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
Aryal, Sandesh, and Ricardo Gutierrez-Osuna. "Data driven articulatory synthesis with deep neural networks." Computer Speech & Language 36 (2016): 260-273.
Hazen, Timothy J., Wade Shen, and Christopher White. "Query-by-example spoken term detection using phonetic posteriorgram templates." 2009 IEEE Workshop on Automatic Speech Recognition & Understanding. IEEE, 2009.
Lee, Ann, Yaodong Zhang, and James Glass. "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.
Sun, Lifa, et al. "Personalized, Cross-Lingual TTS Using Phonetic Posteriorgrams." INTERSPEECH. 2016.
Xie, Feng-Long, Frank K. Soong, and Haifeng Li. "A KL Divergence and DNN-Based Approach to Voice Conversion without Parallel Training Sentences." Interspeech. 2016.
Zhou, Yi, et al. "Cross-lingual Voice Conversion with Bilingual Phonetic Posteriorgram and Average Modeling." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

Method overview



Acoustic modeling

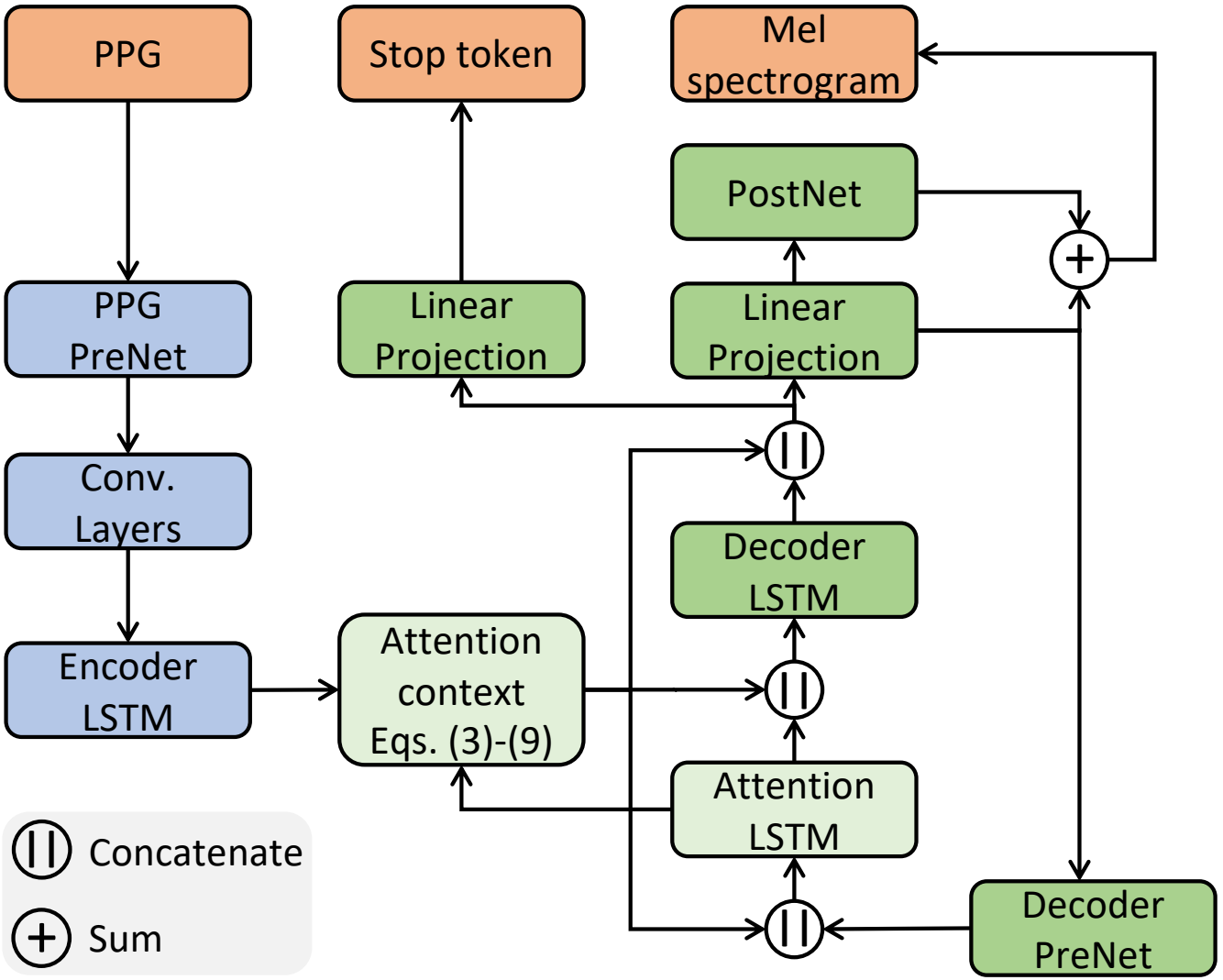




[Zhang, ICASSP'14]

Zhang, Xiaohui, et al. "Improving deep neural network acoustic models using generalized maxout networks." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

PPG->Mel-spectrogram



Long sequence and attention

- Original Tacotron 2 was designed to accept character sequences as input
- Significantly shorter than our PPG sequences
  - Tens of characters vs. a few hundred frames
- Solutions
  - Train the PPG-to-Mel model with shorter PPG sequences ✗
  - Add a locality constraint to the attention mechanism ✓

Locality constraint on attention

Location-sensitive attention $\alpha_i = \mathrm{Attend}(s_{i-1}, \alpha_{i-1}, h) = [\alpha_i^1, \ldots \alpha_i^T]$

Hidden state of attention LSTM

$$s_i = \mathrm{AttentionLSTM}\big(s_{i-1}, g_i, \mathrm{PreNet}(y_i)\big)$$

y_i - predicted acoustic feature

Attention context

$$g_i = \sum_{j=1}^{T} \alpha_i^j h_j$$

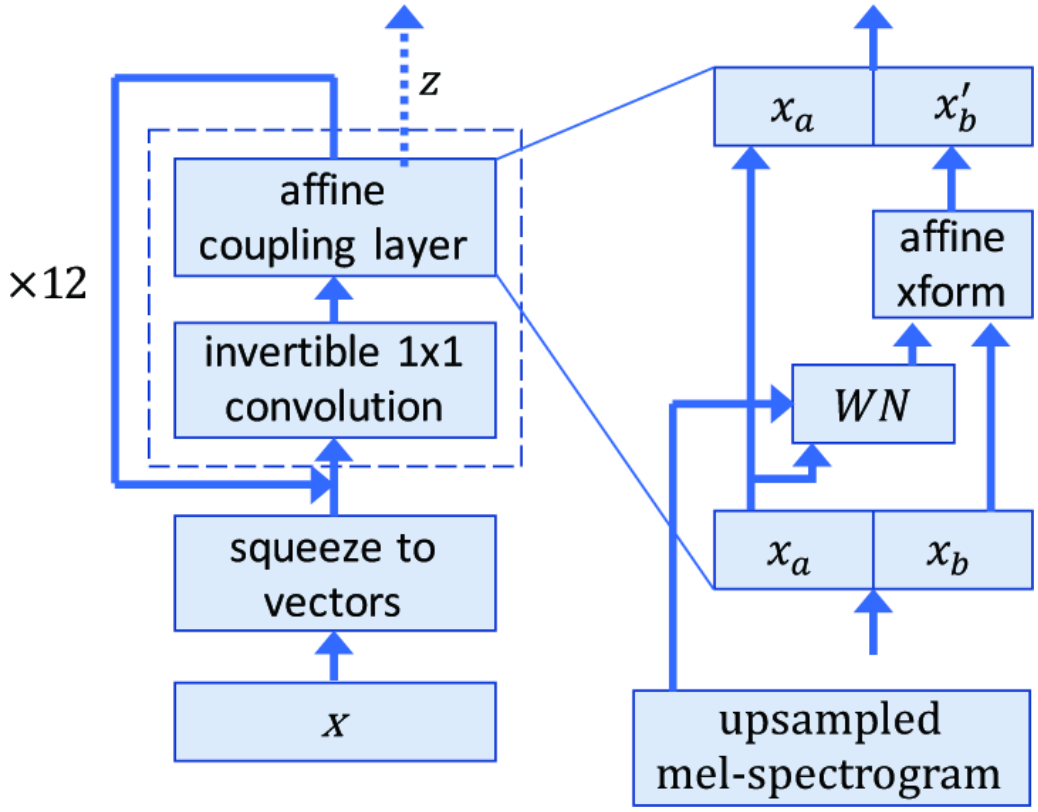
h_j - encoder hidden states

Locality constraint $\tilde{h} = [0, \ldots, 0, h_{i-w}, \ldots, h_{i+w}, 0, \ldots, 0]$

Constrained attention weights $\alpha_i = \mathrm{Attend}\big(s_{i-1}, \alpha_{i-1}, \tilde{h}\big)$

Mel-spectrogram->Speech

WaveGlow



[Prenger, Valle, and Catanzaro, ICASSP, 2019]

Prenger, Ryan, Rafael Valle, and Bryan Catanzaro. "Waveglow: A flow-based generative network for speech synthesis." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

Experimental setup

Acoustic model

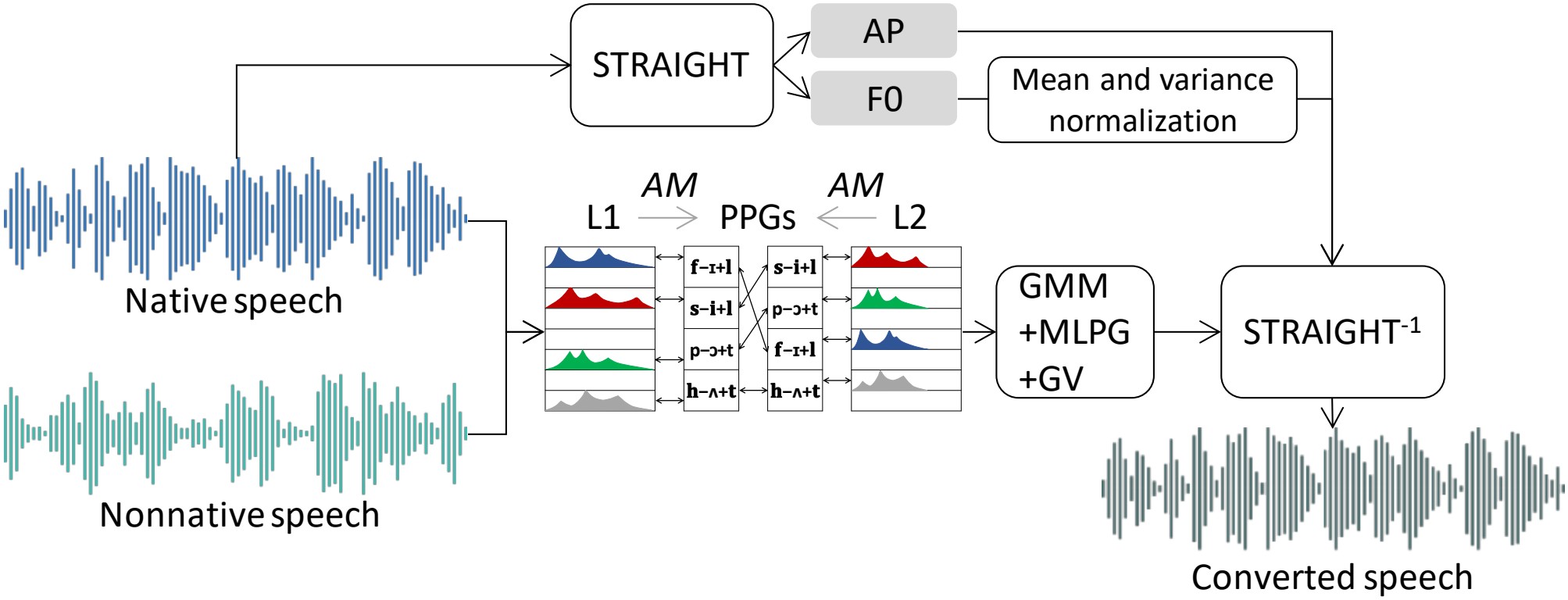
- Trained on Librispeech (960h) with Kaldi
- Five hidden layers and an output layer with 5816 senones

Data

- Native English speakers: General American; BDL (M) & CLB (F); CMU-ARCTIC corpus [Kominek and Black, 2004]
- Non-native English speakers: YKWK (Korean, M) & ZHAA (Arabic, F); L2-ARCTIC corpus [Zhao *et al*., 2018]
- One hour per speaker for training; 50 utts for testing
- Conversion pairs: BDL-YKWK, CLB-ZHAA

Kominek, John, and Alan W. Black. "The CMU Arctic speech databases." Fifth ISCA workshop on speech synthesis. 2004.
Zhao, Guanlong, et al. "L2-ARCTIC: A non-native English speech corpus." Interspeech, 2018.

Baseline

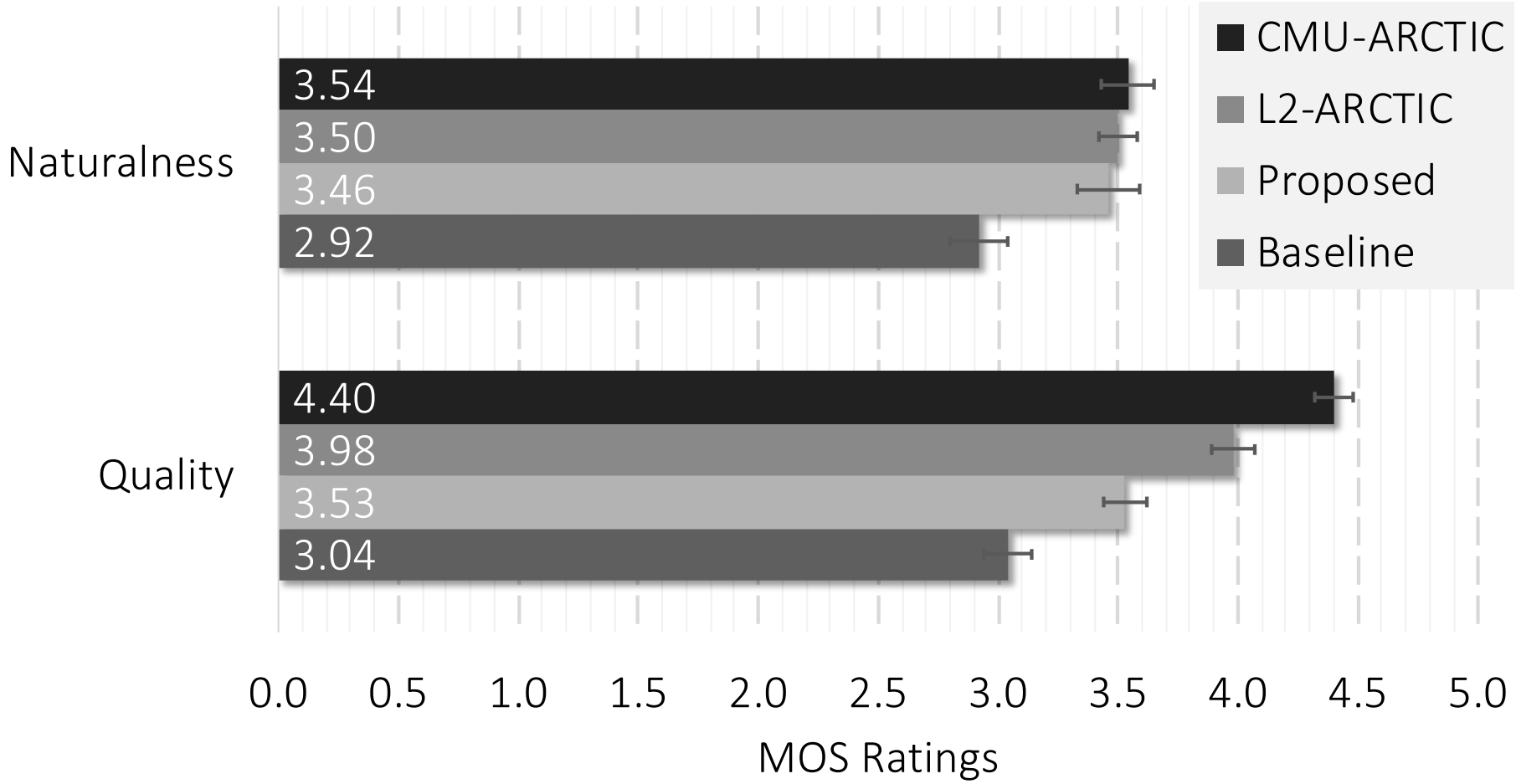


[Zhao & Gutierrez, TASLP, 2019]

https://github.com/guanlongzhao/ppg-gmm

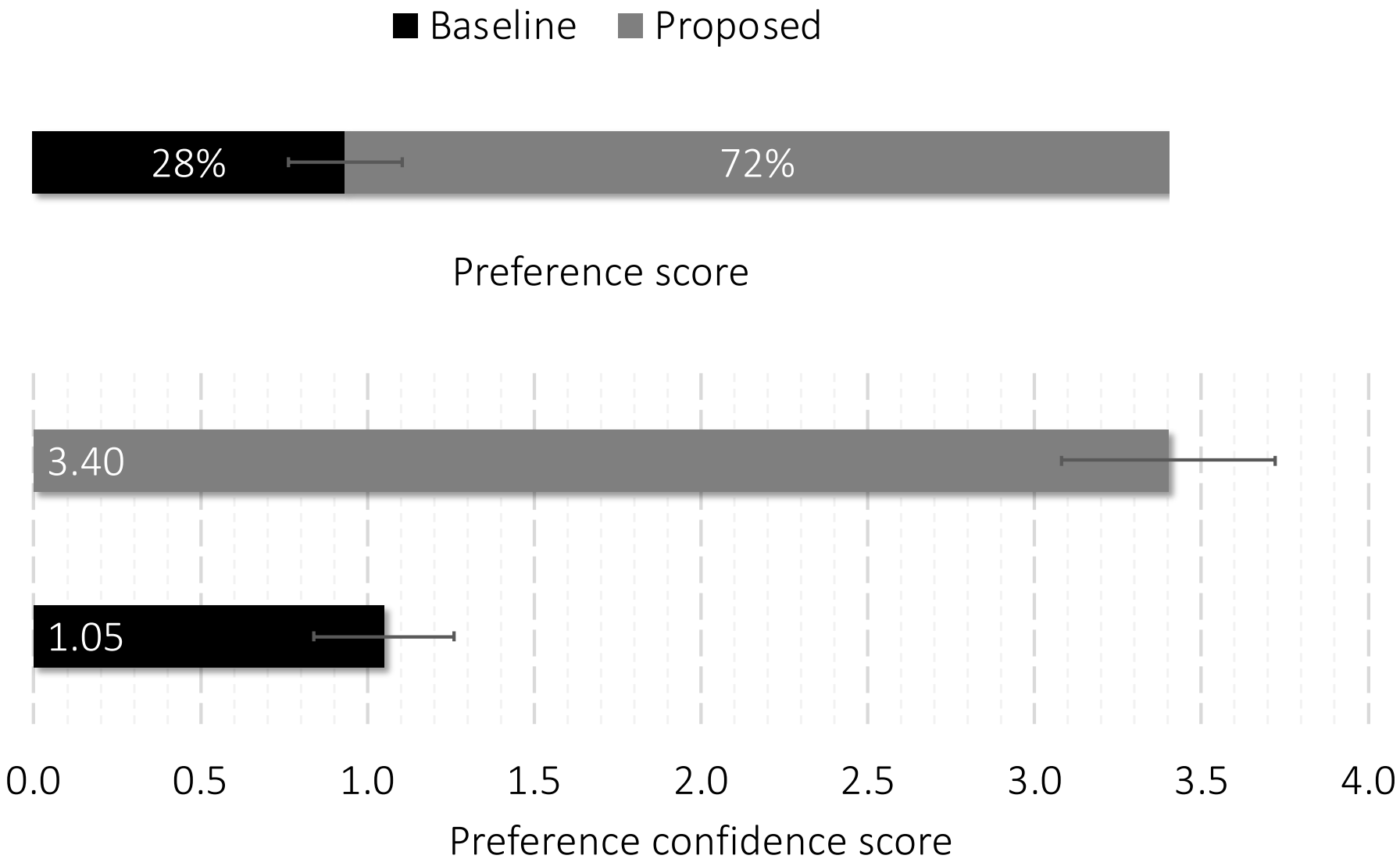
Zhao, Guanlong, and Ricardo Gutierrez-Osuna. "Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion." IEEE/ACM Transactions on Audio, Speech, and Language Processing 27.10 (2019): 1649-1660.

Results: MOS



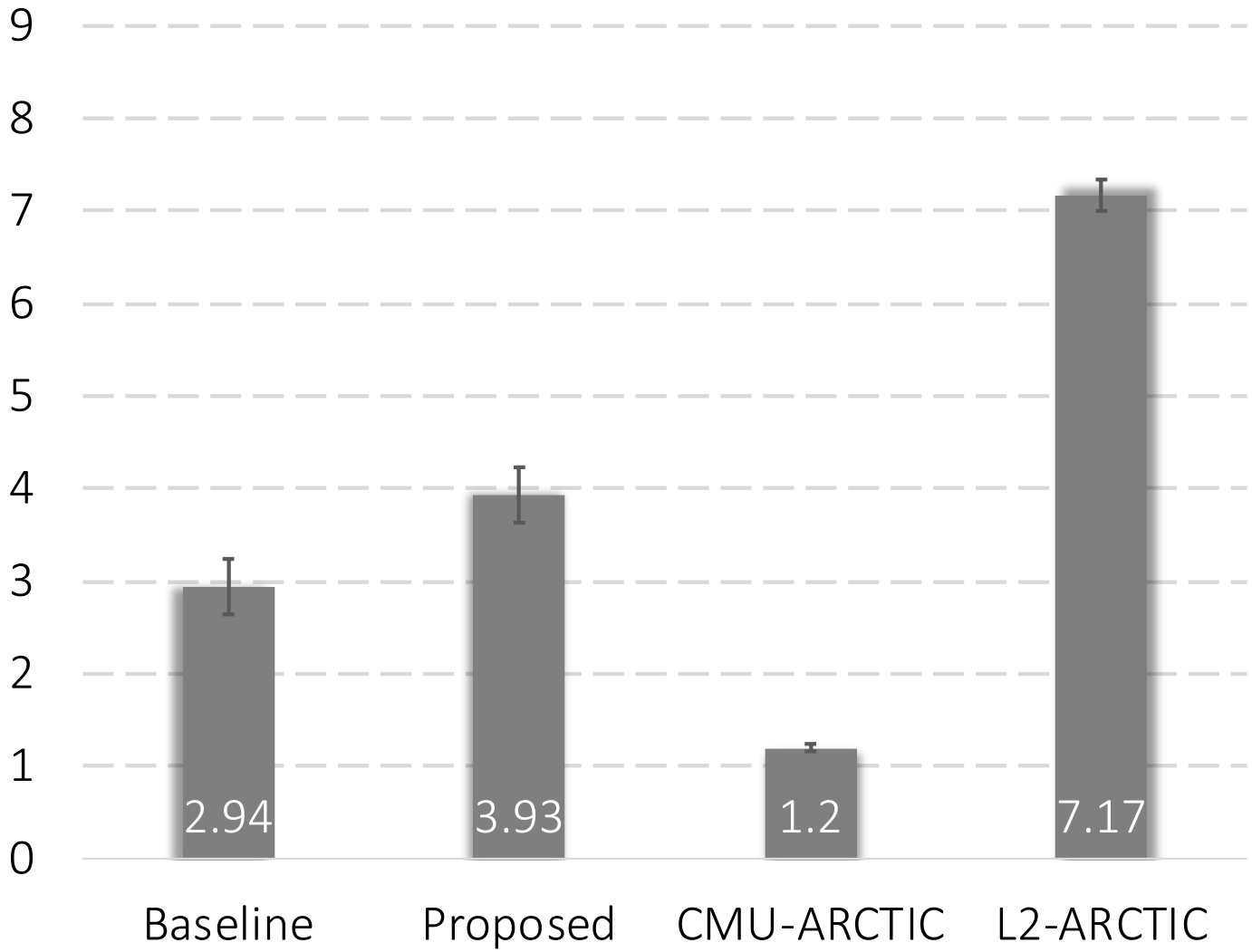
- Rated 50 utts per system, $\geq$ 17 ratings per utt; Amazon MTurk
- *No* statistically significant difference b/w "Proposed" and either CMU-ARCTIC ($p = 0.35$) or L2-ARCTIC ($p = 0.54$) on the naturalness MOS
- Other differences are significant

Results: voice similarity preference



- XAB style preference test with a 7-point confidence rating in the end
- 17 raters rated 50 random L2-ARCTIC:Proposed:Baseline pairs; audios played in reverse

Results: accentedness



- Rated in a 9-point scale, 9 being the most accented
- Each utt rated by 18 listeners who live in the U.S.

Discussion

Better MOS

- Proposed an easy-to-implement locality constraint on the attention mechanism to make the PPG-to-Mel model trainable on utterance-level samples
- MOS ratings are lower than those in the original Tacotron 2 and WaveGlow paper; largely because their systems were trained with 24× more data

Better voice similarity

- Generated the non-native speaker's excitation directly from the synthesized mel-spectrogram

Accentedness

- Significantly less accented than the non-native speech
- Slightly increased accentedness ratings compared with baseline
  - AM inevitably produces recognition errors when extracting the PPG
  - The proposed model does not explicitly model stress and intonation patterns

Conclusion

Future work

- Training the PPG-to-Mel and WaveGlow models jointly
- Incorporate intonation information into the modeling process
- Reduce training data size – transfer learning
- Eliminate the need for a reference utterance at synthesis time

Open-source

- Data: https://psi.engr.tamu.edu/l2-arctic-corpus/
- Code and pre-trained models
  - Proposed: https://github.com/guanlongzhao/fac-via-ppg
  - Baseline: https://github.com/guanlongzhao/ppg-gmm
- Demo: https://guanlongzhao.github.io/demo/fac-via-ppg/

Thanks
Q & A