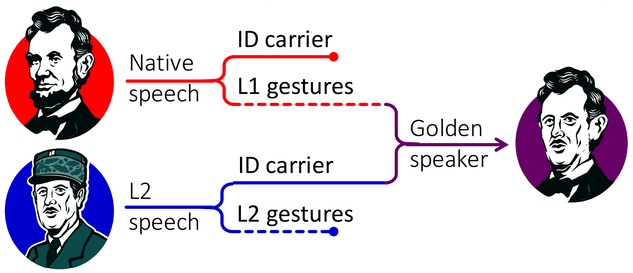
ACCENT CONVERSION USING PHONETIC POSTERIORGRAMS

Guanlong Zhao, Sinem Sonsaat, John Levis, Evgeny Chukharev-Hudilainen, and Ricardo Gutierrez-Osuna

TEXAS A&M UNIVERSITY · IOWA STATE UNIVERSITY

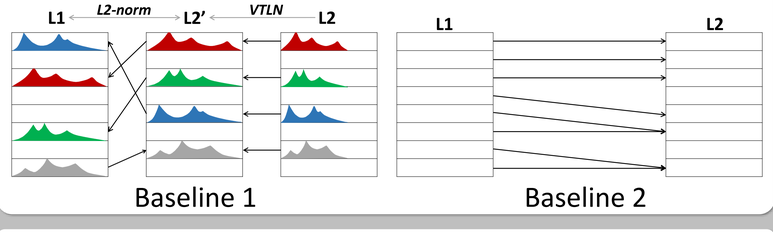
Introduction

- **Objective:** Create speech with a nonnative speaker's voice but the content and pronunciation of a native speaker [1]
- **Idea:** Use voice conversion to capture the nonnative speaker's identity; use careful frame pairing to preserve the native speaker's pronunciation patterns
- **Problem:** The frame alignment method needs to be able to avoid pairing native speech frames with nonnative frames that contain mispronunciations/hesitations/pauses



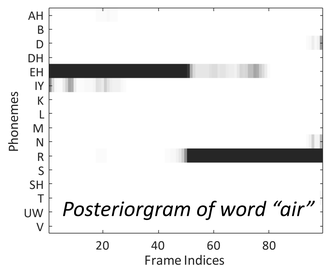
Previous methods for frame pairing

- **Baseline 1: Acoustic similarity [2]**
  - Learn a VTLN transform to reduce physiological differences in vocal tract between the two speakers $T^* = \underset{T}{\operatorname{argmin}} \|\boldsymbol{x} - T\boldsymbol{y}\|^2$
  - For each native vector $\boldsymbol{x_i}$, we find its closest L2 vector $\boldsymbol{y_i^*}$ as $\boldsymbol{y_i^*} = \underset{\forall \boldsymbol{y}}{\operatorname{argmin}} \|\boldsymbol{x_i} - T^*\boldsymbol{y}\|^2$; repeat the same for each L2 vectors $\boldsymbol{y_i}$, $\boldsymbol{x_i^*} = \underset{\forall \boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{x} - T^*\boldsymbol{y_i}\|^2$
- **Baseline 2: Time-alignment (DTW)**



Why do we need a new frame pairing method?

- VTL is just one of potentially many differences between two speakers
- DTW is problematic when the target speaker is nonnative
- **Better solution:** pair frames in a speaker independent space, e.g., the posteriorgram space



Proposed: Use posteriorgrams for frame pairing

- **Posteriorgram:** Compute a feature vector of phonetic posteriors for each speech frame $\boldsymbol{x_i}$

$$\boldsymbol{\mathcal{L}_{x_i}} = [P(l_1|\boldsymbol{x_i}), P(l_2|\boldsymbol{x_i}), \dots, P(l_V|\boldsymbol{x_i})]$$

$V = \{l_1, l_2, \dots, l_V\}$ is the predefined senone set

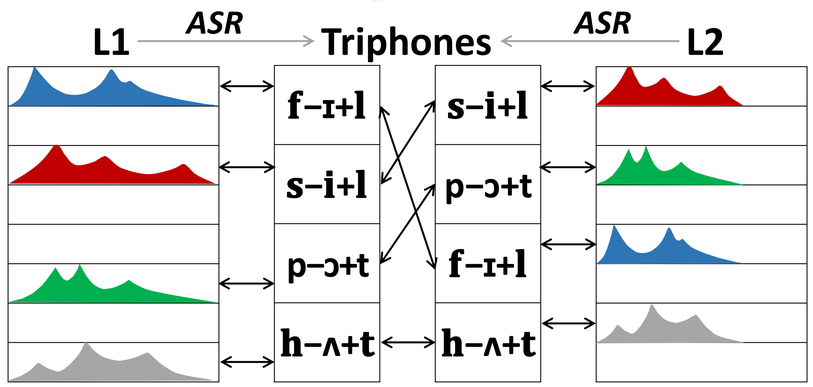
- **Similarity metric:** Symmetric Kullback-Leibler (KL) divergence

$$D\left(\boldsymbol{\mathcal{L}_{x_i}}, \boldsymbol{\mathcal{L}_{x_j}}\right) = \left(\boldsymbol{\mathcal{L}_{x_i}} - \boldsymbol{\mathcal{L}_{x_j}}\right) \cdot \left(\log \boldsymbol{\mathcal{L}_{x_i}} - \log \boldsymbol{\mathcal{L}_{x_j}}\right)$$

- **Pair frames:** Find the closest pairing for each native ($\boldsymbol{x_i}$) and nonnative ($\boldsymbol{y_i}$) frame

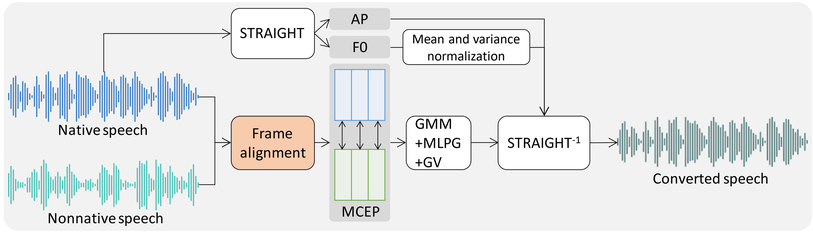
$$\boldsymbol{y_i^*} = \underset{\forall \boldsymbol{y}}{\operatorname{argmin}}\, D\left(\boldsymbol{\mathcal{L}_{x_i}}, \boldsymbol{\mathcal{L}_{y}}\right)$$

$$\boldsymbol{x_i^*} = \underset{\forall \boldsymbol{x}}{\operatorname{argmin}}\, D\left(\boldsymbol{\mathcal{L}_{x}}, \boldsymbol{\mathcal{L}_{y_i}}\right)$$

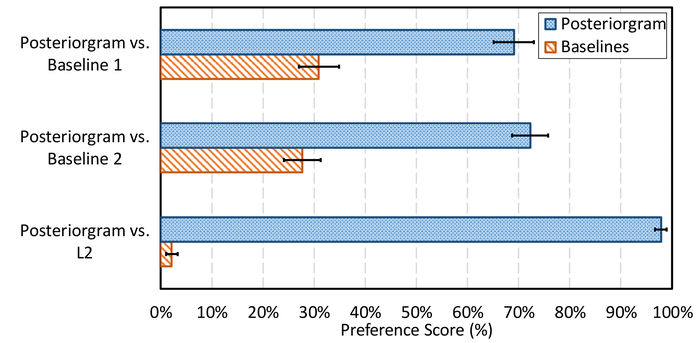


Experimental setup

- **Acoustic model:** A p-norm DNN with 18 hidden layers, trained on Librispeech (960h), 5816 senones
- **Dataset:** Native speakers from CMU ARCTIC: BDL (m), CLB (f). L2 English speakers from L2-ARCTIC: TNI (Hindi, f), RRBI (Hindi, m), HKK and YKWK (Korean, m), ABA (Arabic, m). Each speaker has 100 and 50 utts for training and testing, respectively
- **Systems:** use *posteriorgram*/*Baseline 1*/*Baseline 2* for frame pairing; fix the spectral and prosody conversion components
- **AC pairs:** BDL to RRBI, BDL to HKK, BDL to YKWK, BDL to ABA, and CLB to TNI

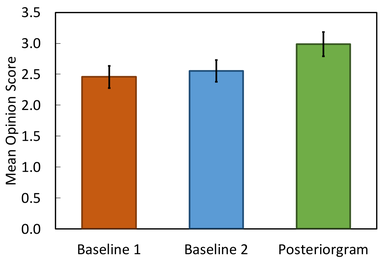


Subjective evaluation



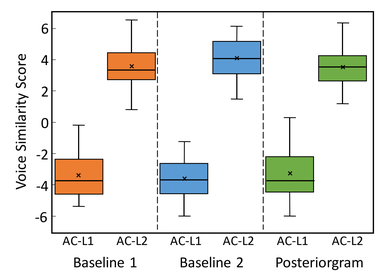
Accentedness (preference test)

- The Posteriorgram conversions were more native-like than the original L2 utterances (mean: 98%, STD: 3%)
- The Posteriorgram method outperformed both Baseline 1 (mean: 69%, STD: 11%) and 2 (mean: 72%, STD: 10%)



Acoustic quality (MOS)

- No statistical differences between the two baselines (2.5 vs. 2.6; p=0.43)
- The posteriorgram system (3.0) was statistically higher than the baselines



Speaker identity

- Three systems have similar voice similarity scores (VSS)
- No statistically-significant differences in VSS between the posteriorgram and the baselines

Discussion

- We proposed a new frame-pairing method based on the phonetic similarity between acoustic frames
- Merely changing the frame pairing method can lead to significant improvement in acoustic quality and "nativeness" while keeping the voice quality of the nonnative speaker
- **Future work:** Apply this technique to pronunciation training in classroom settings

References

[1] D. Felps, *et al.*, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920-932, 2009.

[2] S. Aryal and R. Gutierrez-Osuna, "Can Voice Conversion Be Used to Reduce Non-Native Accents?," in *ICASSP*, 2014, pp. 7879-7883.

Samples and data