ONE-SHOT VOICE CONVERSION WITH DISENTANGLED REPRESENTATIONS BY LEVERAGING PHONETIC POSTERIORGRAMS

Seyed Hamidreza Mohammadi, Taehwan Kim

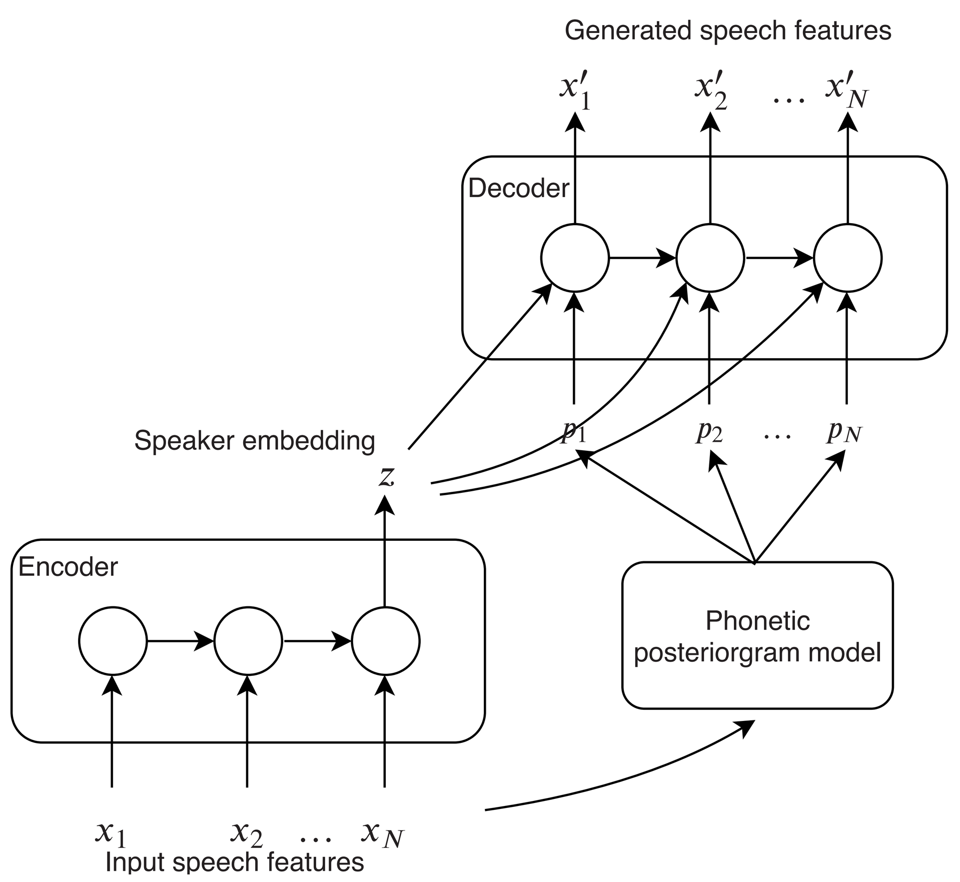
ObEN Inc.

INTRODUCTION

- **Voice Conversion** (VC): converts a source speaker's speech to sound like a target speaker's voice.
  - VC preserves target's speaker identity and source's phonetic context.
  - One-shot VC methods typically disentangle speaker identity and phonetic context. Then speaker identity representation is modified while keeping phonetic context constant.
- **Challenges**: The models cannot fully disentangle these factors as shown in a previous study.
- **Proposal**: We posit that the senone posteriorgrams (PPG) from an already-trained ASR model can be used in lieu of learned phonetic context representations.
  - We focus on learning only the speaker representation.
  - We present a one-shot voice conversion technique by modifying the learned speaker identity representation.
  - Through experiments, we show that modification of these factors allows better disentanglement and hence transformation of voice.

MODEL

- Our proposed model consists of an encoder and a decoder (RNNs)
  - Encoder's input is MCEPs, outputting a speaker embedding vector: $z = E(\mathbf{X})$
  - Decoder takes the generated speaker embedding along with PPG sequence as input, and generates the acoustic features: $X' = D(\mathbf{P}, z)$



- We train the model by optimizing the training loss:

$$\ell(X, X') = \Sigma_{i=1}^{N} ||x_i - x'_i||_2^2$$

- To perform VC:
  - Compute z^{src} and z^{trg} of the source and target utterances.
  - Compute average diff vector $\bar{z}^{diff} = \bar{z}^{trg} - \bar{z}^{src}$
  - Add average diff vector to source $z^{converted} = z^{src} + \bar{z}^{diff}$

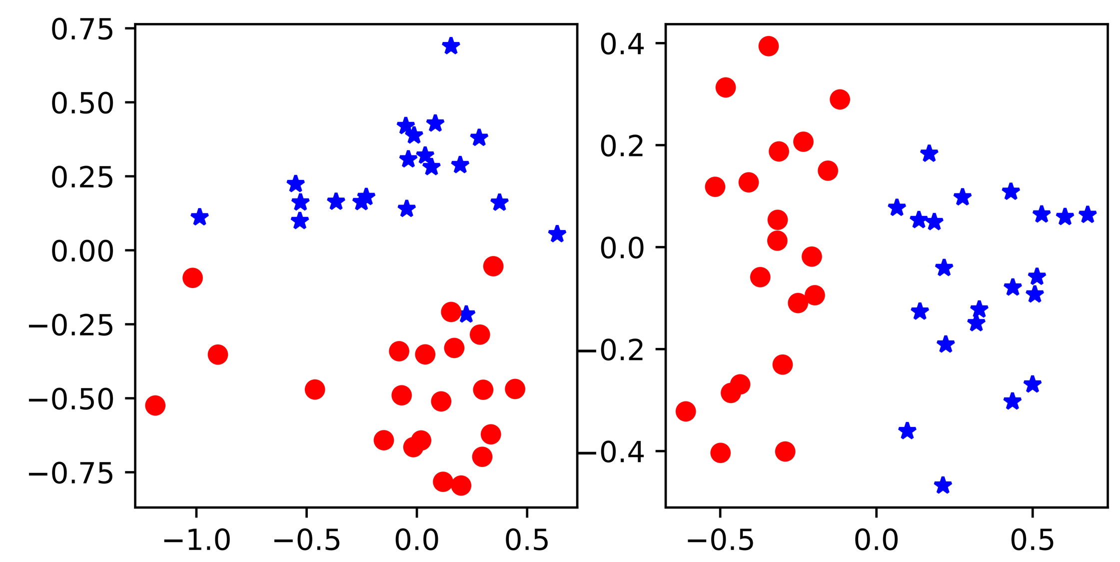
EXPERIMENT

- We used the TIMIT corpus as the training data.
- To compute the phonetic posteriorgrams, we use Kaldi
- We use librispeech as speech corpus to train ASR.

Encoder	
recurrent layer	GRU-1024, Dropout
output layer	FC-D_z, ReLU

Decoder	
dense block	input PPGs, FC-1024, ReLU, Dropout
combine layer	dense output + speaker embedding z
dense block	FC-1024, ReLU, Dropout
recurrent layer	GRU-1024, Dropout
output layer	FC-D_x

The network architectures of our encoder and decoder models.



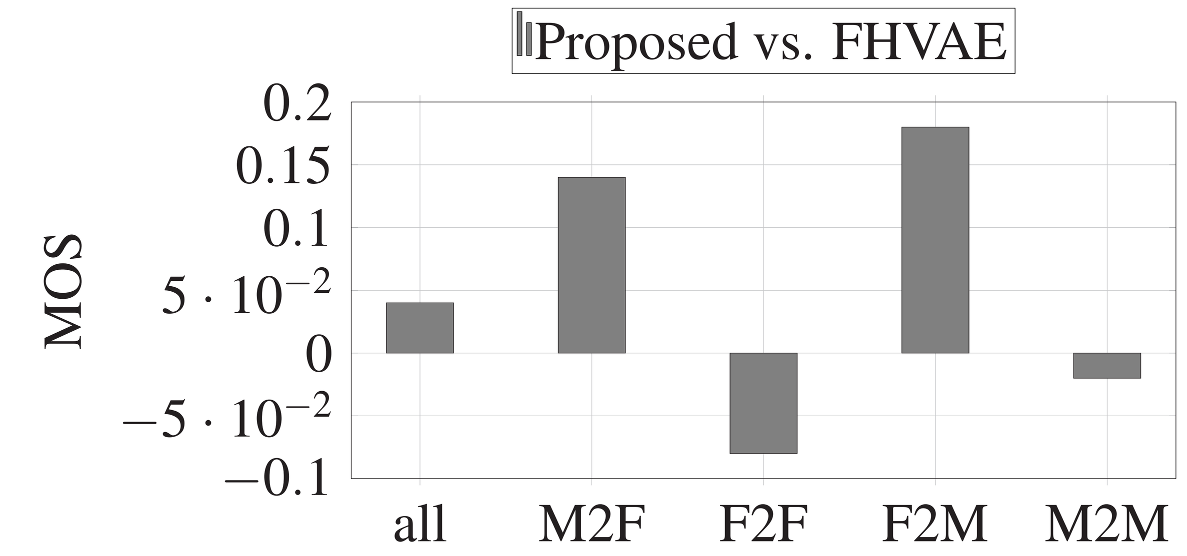
Visualization of speaker embedding: Blue dots are male speakers and red dots are females. FHVAE (left) vs. Proposed (right).

- We use Factorized Hierarchal Variational Autoencoder (FHVAE) [1] as baseline.
- We observe that
  - The proposed model's computed speaker embeddings for different speakers fall further apart compared to FHVAE.
  - Also they are more evenly distributed compared to VAE embeddings which tend to be more densely distributed.
  - The gender clusters have a better separation margin.
- This subjectively depicts a more robust speaker embedding quality.
- The voice conversion samples are available at:
  - https://shamidreza.github.io/is19samples

[1] *Seyed Hamidreza Mohammadi and Taehwan Kim, Investigation of using disentangled and interpretable representations for one-shot cross-lingual voice conversion." Interspeech 2018.*

EXPERIMENT: SPEECH QUALITY

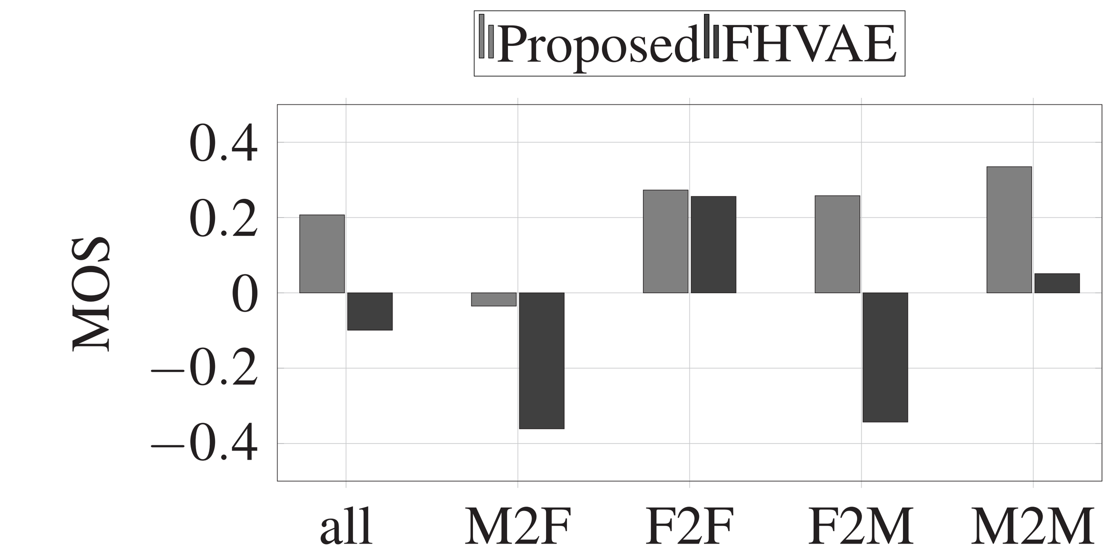
- We show the speech quality Comparative Mean Opinion Score (CMOS) in which 50 listeners score which sample quality is better by using +2 (much better) to -2 (much worse) score.
- We found statistical significant preference scores for F2M condition.
- We did not find statistical significant preference scores for other conditions.



Speech Quality average score with gender break-down. Positive scores favor proposed model. F2M preference score is statistically significant.

EXPERIMENT: SPEAKER SIMILARITY

- 50 listeners listen and rate A and B with score ranging from +2 (definitely same) to -2 score (definitely different)
- The results show proposed and FHVAE achieving 0.20±0.11 and -0.10±0.12
- The proposed model performs statistically significantly better than FHVAE in all comparison pairs



Speaker Similarity average score with gender break-down. Positive scores are desirable. (confidence intervals for all is close to 0.11, and all score-pairs are statistically significant)