INVESTIGATION OF USING DISENTANGLED AND INTERPRETABLE REPRESENTATIONS FOR ONE-SHOT CROSS-LINGUAL VOICE CONVERSION

Seyed Hamidreza Mohammadi, Taehwan Kim

ObEN Inc.

INTRODUCTION

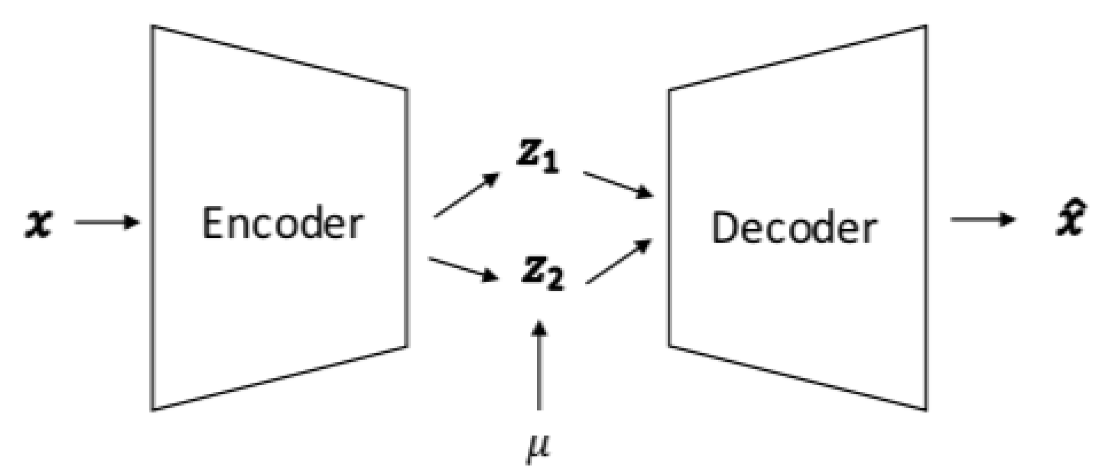
- **Voice Conversion** (VC): convert a source speaker's speech to sound like a target speaker's voice.
- VC preserves target speaker's identity and source phonetic context.
- **Challenges**: requires parallel spoken corpus and enough amount of data; needs to know and include target speaker in training
- We present a one-shot voice conversion technique using factorized hierarchical variational autoencoder (FHVAE)[1] to disentangle speaker identity and linguistic content factors from speech.
- We investigate Mel-cepstrum (MCEP) speech representation and achieve better results compared to baselines.
- We show that modification of these factors allow transformation of voice, even in challenging *cross-lingual* scenario.

FACTORIZED HIERARCHICAL VARIATIONAL AUTOENCODER

- Variational autoencoder (VAE) is a powerful model to uncover hidden representation and generate new data samples, but considers no structure for latent variable Z.
- We use a newly proposed Factorized Hierarchical VAEs (FHVAEs), which have disentagled latent variable Z_1 for linguistic context and Z_2 for speaker identity.
- Joint probability:

$$p_\Phi(X^i, Z_1^i, Z_2^i, \mu^i) = p_\Phi(\mu^i)\prod_{j=1}^{N_{seg}^i} p_\Phi(X^{i,j}|Z_1^{i,j}, Z_2^{i,j})p_\Phi(Z_1^{i,j})p_\Phi(Z_2^{i,j}|\mu^i)$$

- We use LSTM for encoder and decoder.



- To perform VC:
  - Compute Z_1 and Z_2 of the input utterance and target utterances.
  - Compute average diff vector $Z_2^{diff} = Z_2^{trg} - Z_2^{src}$
  - Add average diff vector to source Z_2: $Z_2^{converted} = Z_2 + Z_2^{diff}$

[1] *W. Hsu et al, Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data, NIPS 2017.*

VISUALIZATIONS

- We trained two FHVAEs on:
  1) one on TIMIT English speech corpus with 462 speakers and
  2) a proprietary Chinese corpus with 5200 speakers.
- We test on:
  1) Four CMU-arctic speakers and
  2) Four speakers from THCHS-30 Chinese corpus.

Figure: Speaker embeddings (2D PCA). Each point represents single speaker embedding. Blueish dots are English females and light blueish are Chinese females; and reddish dots are English males and orange dots are Chinese males.

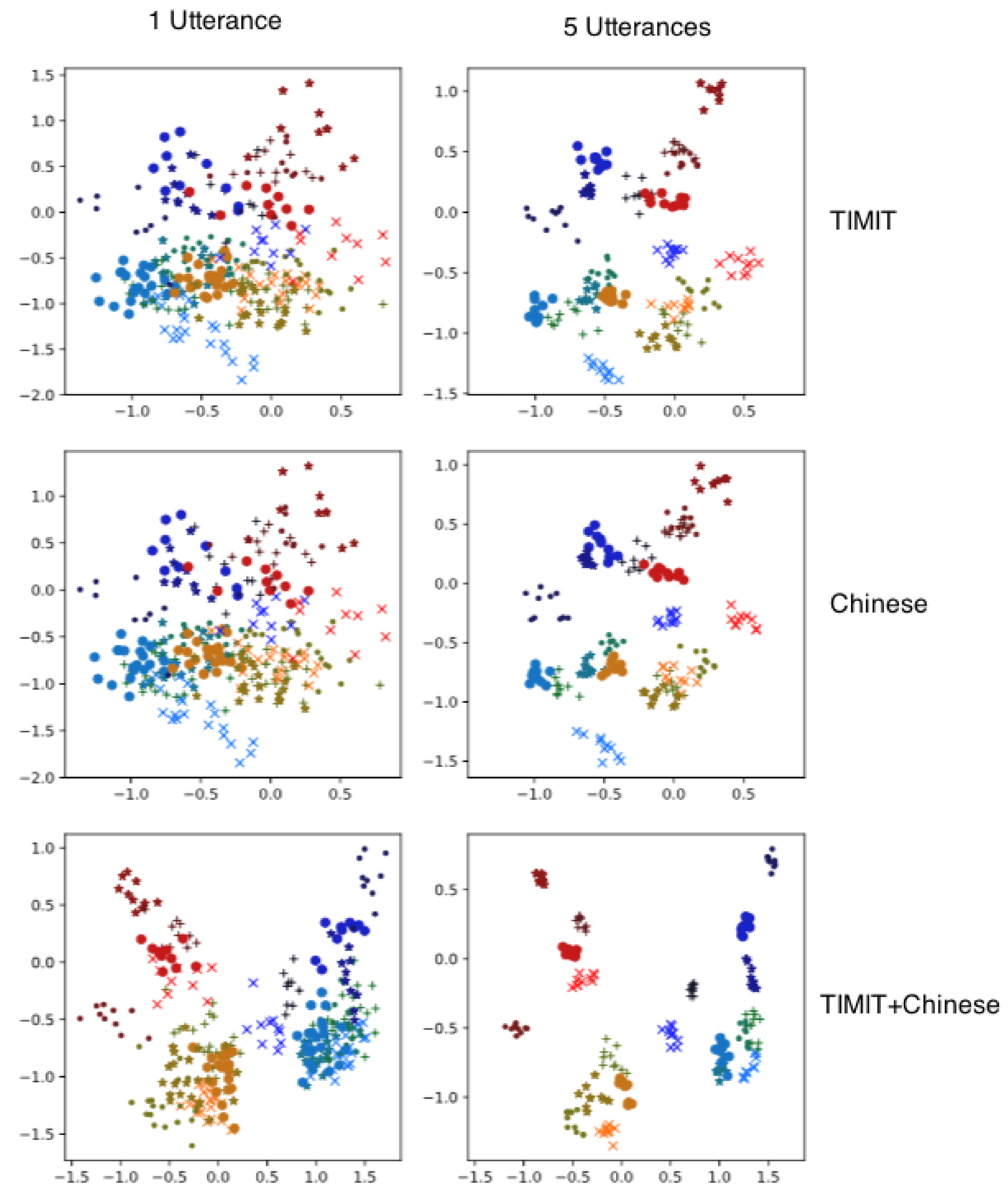
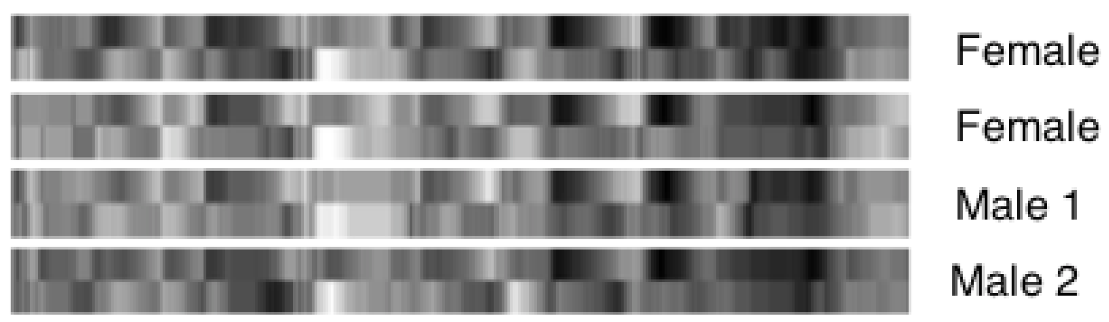


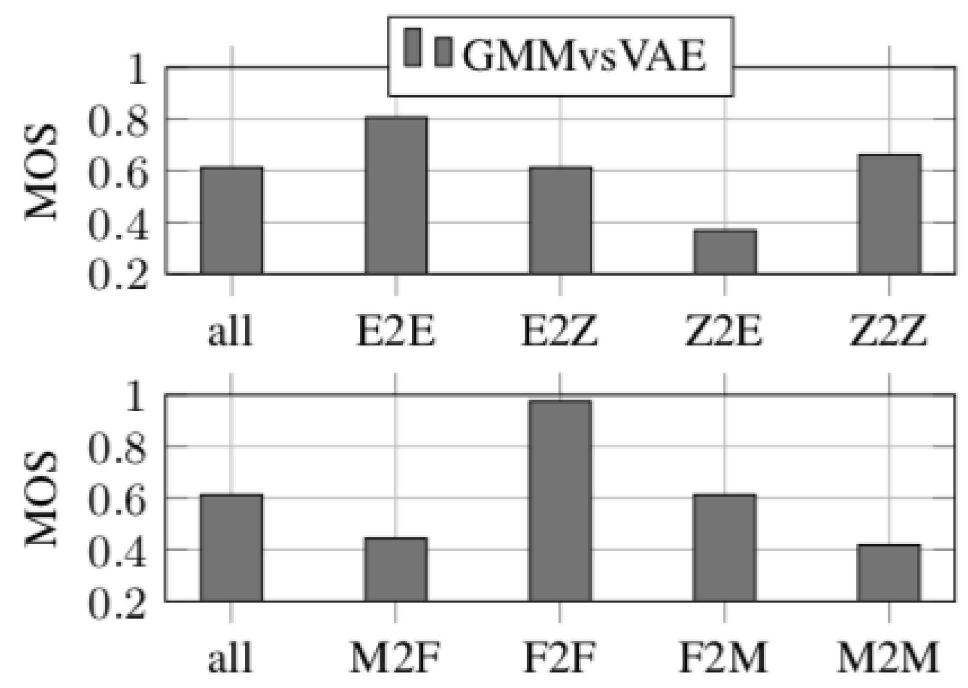
Figure: Linguistic embeddings (2D PCA) of sentence "She had your dark suit in greasy wash water all year." for two females (top) and males (bottom).



- The voice conversion samples are available at: https://shamidreza.github.io/is18samples

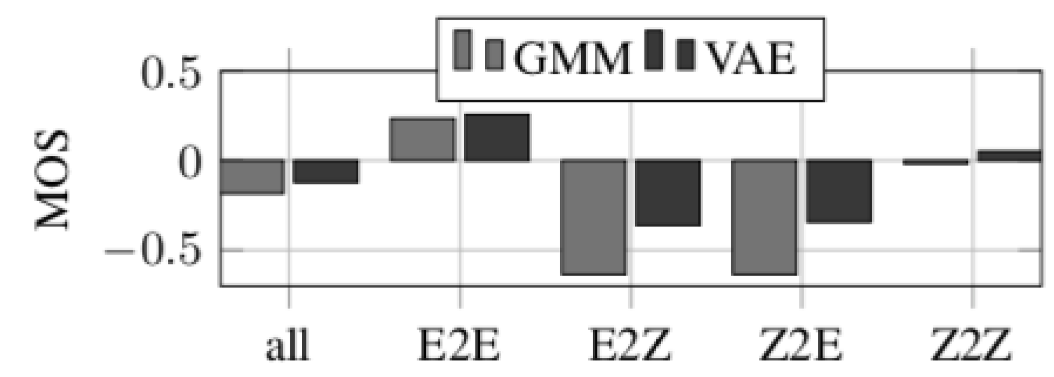
EXPERIMENT: SPEECH QUALITY

- MCEP is used as speech representation and TIMIT for training in experiments below.
- We use VAE (FHVAE) with STFT and GMM as baselines.
- We show the speech quality Comparative Mean Opinion Score (CMOS) in which 40 listeners score which sample quality is better by using +2 (much better) to -2 (much worse) score.
- VAE vs. VAE-STFT: +1.25±0.12 mean score towards VAE.
- VAE vs. GMM: +0.61±0.14 mean score towards VAE

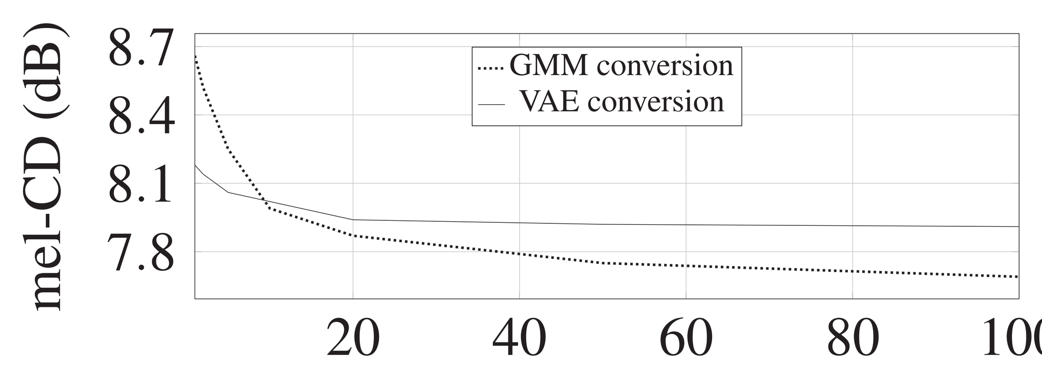


EXPERIMENT: SPEAKER SIMILARITY

- 40 listeners listen and rate A and B with score ranging from +2 (definitely same) to -2 score (definitely different).
- The results show GMM and VAE achieving -0.18±0.15 and -0.12±0.16. E2E achieves the best performance.



- Effects of varying number of target sentences from 1 to 100.



ObEN Inc — Pasadena, California USA

email: hamid@oben.com, taehwan@oben.com

www.oben.com