NON-PARALLEL VOICE CONVERSION USING VARIATIONAL AUTOENCODERS CONDITIONED BY PHONETIC POSTERIORGRAMS AND D-VECTORS

Yuki Saito [†] [‡], Yusuke Ijima [‡], Kyosuke Nishida [‡], and Shinnosuke Takamichi [†]

[†] The University of Tokyo, Japan [‡] NTT Corporation, Japan

1. SYNOPSIS

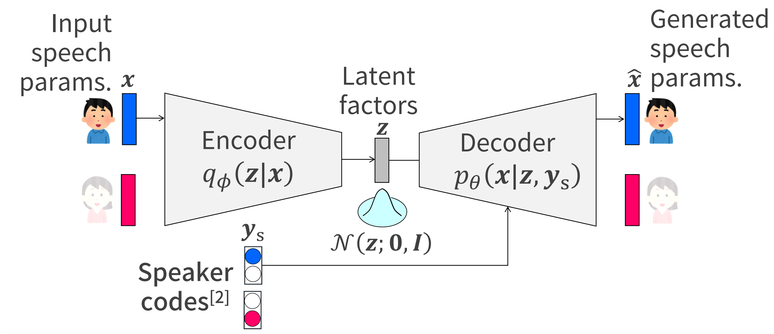
Our approaches for non-parallel Voice Conversion (VC) using Variational AutoEncoders (VAEs):

(1) Introduce Phonetic PosteriorGrams (PPGs) for dealing with speech quality degradation.

(2) Extend conventional one-to-one VC to many-to-many VC (any speakers to any other speakers).

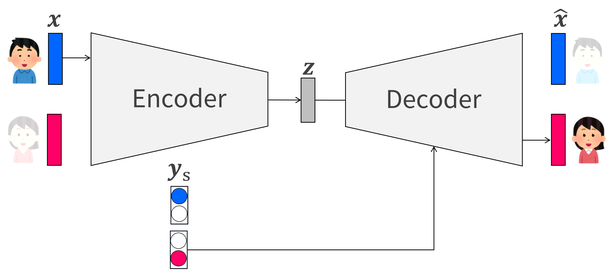
2. CONVENTIONAL VAE-BASED VC[1]

Training VAEs for VC



$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x},\boldsymbol{y}_s) = \underbrace{-D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, \mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})\right)}_{\text{Regularization term of } \boldsymbol{z}} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y}_s)]}_{\text{Reconstruction error of } \boldsymbol{x}}$$

Conversion using trained VAEs



Assumption: $\boldsymbol{z}$ is independent of $\boldsymbol{y}_s$

⇒ Expected to represent phonetic contents.

Problems

(1) Converted speech quality is degraded due to vanishing phonetic properties.

(2) It supports only one-to-one VC.

3. PROPOSED VAE-BASED VC

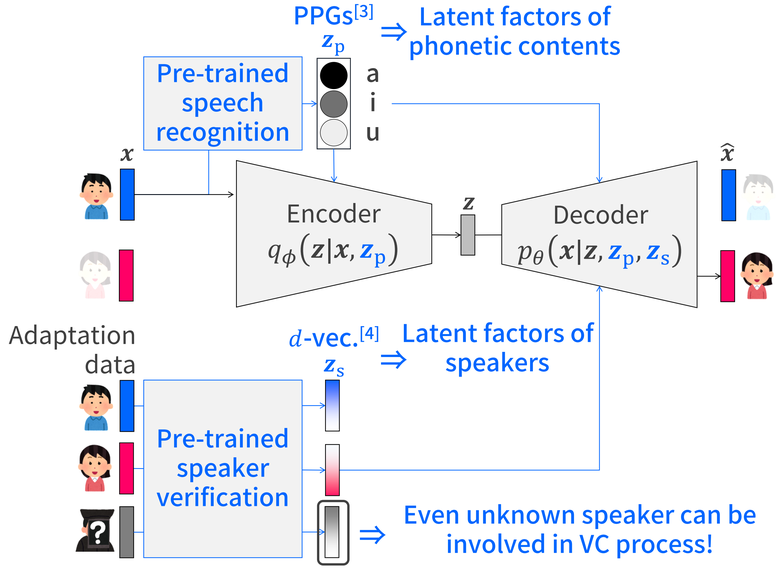
VC using VAEs w/ PPGs and d-vectors

(1) Phonetic contents are **given as PPGs**[4].

⇒ **Phonetic contents can be restored!**

(2) Discrete speaker codes are replaced with **continuous d-vectors**[5].

⇒ Any speakers' characteristics can be converted into any others!



Estimating unknown speaker representation

(1) Speaker code adaptation[6]

Using backprop. to adapt speaker codes

(2) d-vector averaging

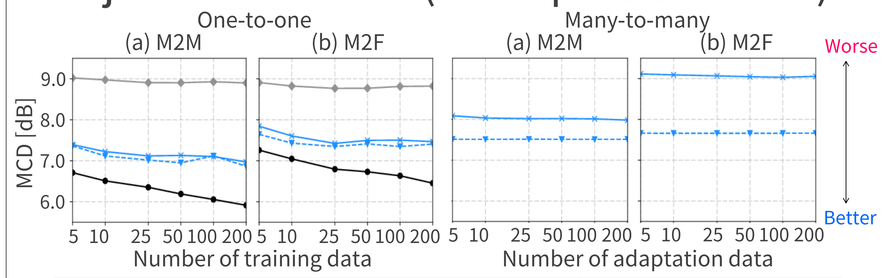
Using averaged d-vec. in **voiced** region of the speaker's adaptation data

4. EXPERIMENTAL EVALUATION

Experimental conditions

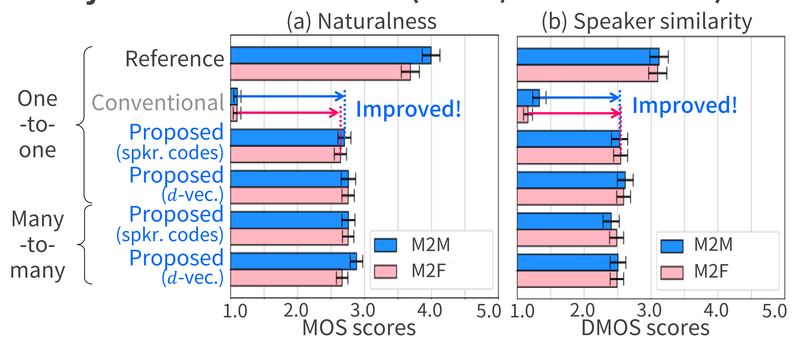
Speech corpora	(a) **For recognition/verification models:** including 260 speakers (130 male and 130 female speakers) (b) **For VC models:** divided parallel 425 utterances (2 male and 1 female speakers)
Speech params. (including Δ & ΔΔ)	40-dim. mel-cepstral coefficients, Log F0, 10-dim. bap
DNN architecture	Feed-Forward (see our paper)
Dim. of PPGs/d-vec./$\boldsymbol{z}$	56 (time variant)/16 (time invariant)/64
VC settings	One-to-one: VC models are trained with corpora (b). Many-to-many: VC models are trained with corpora (a).

Objective evaluation (Mel-Cepstral Distortion)



*DNN trained w/ **parallel** data (M2M: Male-to-Male, M2F: Male-to-Female)

Subjective evaluation (MOS/DMOS tests)



[References]

[1] Hsu et al., *Proc. APSIPA ASC*, 2016. [2] Hojo et al., *IEICE Trans. on Info. And Syst.*, 2017. [3] Dilokthanakul et al., *arXiv:1611.02648,* 2016. [4] Sun et al., *Proc. ICME,* 2016. [5] Variani et al., *Proc. ICASSP,* 2014. [6] Luong et al., *Proc. ICASSP,* 2017.