19 Face-to-voice conversion and voice-to-face conversion

- Crossmodal voice conversion with deep generative models -

Abstract

Humans are able to imagine a person's voice from the person's appearance and imagine the person's appearance from his/her voice. In this work, we take an information-theoretic approach using deep generative models to develop a method that can convert speech into a voice that matches an input face image and generate a face image that matches the voice of the input speech by leveraging the correlation between faces and voices. We propose a model, consisting of a speech encoder/decoder, a face encoder/decoder and a voice encoder. We use the latent code of an input face image encoded by the face encoder as the auxiliary input into the speech decoder and train the speech encoder/decoder so that the original latent code can be recovered from the generated speech by the voice encoder. We also train the face decoder along with the face encoder to ensure that the latent code will contain sufficient information to reconstruct the input face image.

Crossmodal Voice Conversion/Face Image Generation

Leverage underlying correlation between voices and appearances to

→
- convert speech into a voice that matches an input face image, and
- generate face image that matches input speech.

Information-theoretic approach using deep generative models

Voice Convertor G: Neural network (NN) that converts input speech $\mathbf{x}$ into $\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{y})$ by using face image $\mathbf{y}$ as auxiliary input

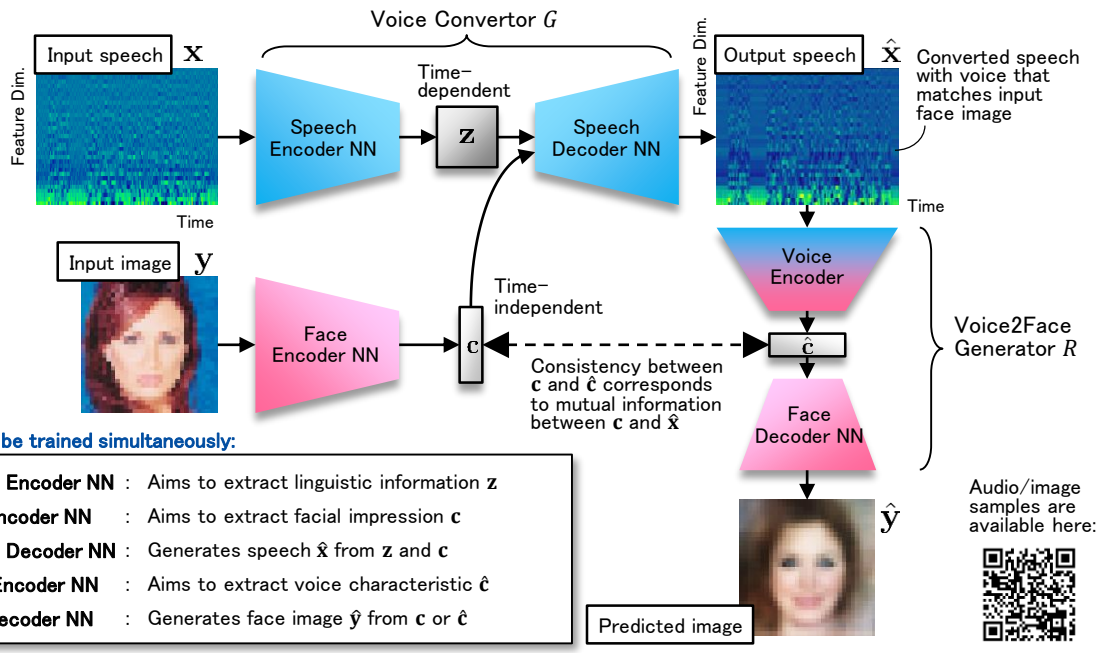
Training objective: Train G so that mutual information between $\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{y})$ and $\mathbf{y}$ is maximized

$$I[G(\mathbf{x}, \mathbf{y}) \| \mathbf{y}] \geq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p(\mathbf{x},\mathbf{y})}[\log R(\mathbf{y}|G(\mathbf{x}, \mathbf{y}))]$$

→ Maximize lower bound w.r.t. G and R

Speech and face image pair

NN that approximates posterior $p(\mathbf{y}|\mathbf{x})$



Five NNs to be trained simultaneously:

Speech Encoder NN	:	Aims to extract linguistic information $\mathbf{z}$
Face Encoder NN	:	Aims to extract facial impression $\mathbf{c}$
Speech Decoder NN	:	Generates speech $\hat{\mathbf{x}}$ from $\mathbf{z}$ and $\mathbf{c}$
Voice Encoder NN	:	Aims to extract voice characteristic $\hat{\mathbf{c}}$
Face Decoder NN	:	Generates face image $\hat{\mathbf{y}}$ from $\mathbf{c}$ or $\hat{\mathbf{c}}$

References

[1] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. 2018 IEEE Workshop on Spoken Language Technology (SLT 2018)*, pp. 266-273, Dec. 2018.

[2] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, " ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder, " *arXiv*:1808.05092 [stat.ML], 2018.

[3] H. Kameoka, K. Tanaka, A. Valero Puche, Y. Ohishi, T. Kaneko, "Crossmodal Voice Conversion," *arXiv*:1904.04540 [cs.SD], 2019.

Contact

Hirokazu Kameoka Email: cs-liaison-ml at hco.ntt.co.jp

Media Recognition Research Group, Media Information Laboratory

Innovative R&D by NTT

Open House 2019