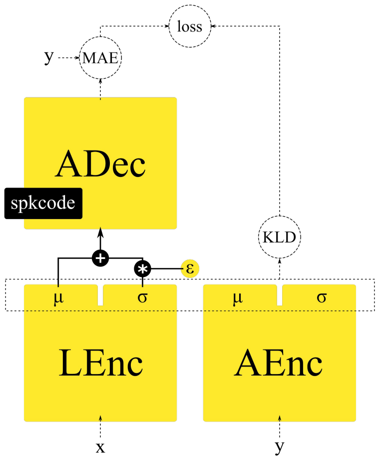
Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech

Hieu-Thi Luong, Junichi Yamagishi (NII, Japan)

NII Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics



Multimodal architecture for speaker-adaptive text-to-speech

Jointly training an auxiliary acoustic encoder with a typical TTS system so it could be used to perform unsupervised speaker adaptation later.

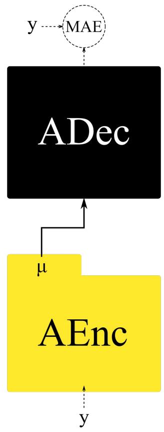
Text-to-speech stack:

$$z^L \sim LEnc(x; \phi^L) = p(z|x)$$
$$\tilde{y}^L = Dec(z^L; \theta^{\text{core}}, \theta^{\text{spk},(k)})$$

Speech-to-speech stack:

$$z^A \sim AEnc(y; \phi^A) = q(z|y)$$
$$\tilde{y}^A = Dec(z^A; \theta^{\text{core}}, \theta^{\text{spk},(k)})$$

"A Unified Speaker Adaptation Method for Speech Synthesis using Transcribed and Untranscribed Speech with Backpropagation" Hieu-Thi Luong and Junichi Yamagishi arXiv preprint arXiv:1906.07414



(M)any-to-one voice conversion

Development procedure:

Step 1. Train the initial TTS model:

$$\text{loss}_{train} = \text{loss}_{tts} + \beta\ \text{loss}_{tie}$$
$$\text{loss}_{tts} = L_{MAE}(\tilde{y}^L, y)$$
$$\text{loss}_{tie} = L_{KLD}(LEnc(x), AEnc(y))$$

Step 2. Adapt to target speaker:

$$\text{loss}_{adapt} = \text{loss}_{sts}$$
$$\text{loss}_{sts} = L_{MAE}(\tilde{y}^A, y)$$

Step 3. Convert speech utterances of arbitrary speakers to the target voice.

Abstract

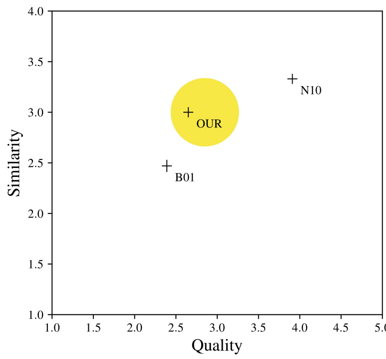
Voice conversion (VC) and text-to-speech (TTS) are two tasks that share a same objective of generating speech with a target voice. However, they are usually developed under vastly different frameworks.

We propose a method to bootstrap a VC system from a pretrained speaker-adaptive TTS model by fine-tuning to untranscribed speech data of the target speaker.

The methodology can also be used to build a VC system for unseen (and without transcript) languages.

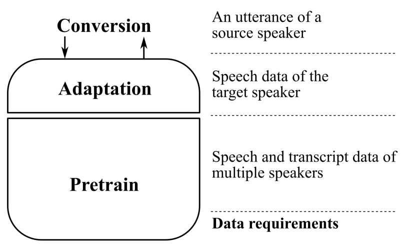
x=linguistic
y=melspectrogram
zdim=64
β=0.25
sr=22050 Hz
corpus=VCTK
nspeaker=72
vocoder=WaveNet
quantization=10bit
nsubject=28

Speech samples

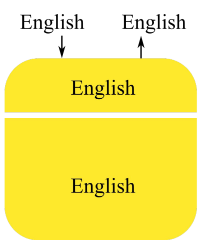


Subjective results for reenactment of VCC2018 non-parallel task

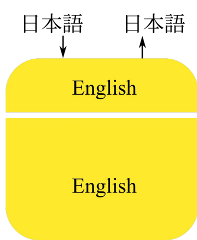
Cross-lingual voice conversion?



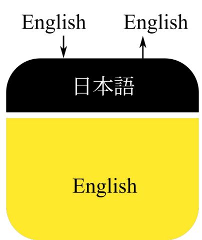
PA-C



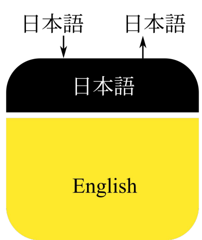
EE-E
Voice conversion



EE-J
Cross-language voice conversion

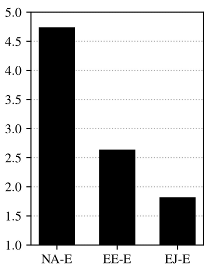


EJ-E
Cross-language speaker adaptation

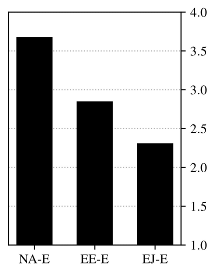


EJ-J
Text-to-speech without text

Two bilingual (Japanese-English) speakers are used to test the performance of the unseen language scenarios.

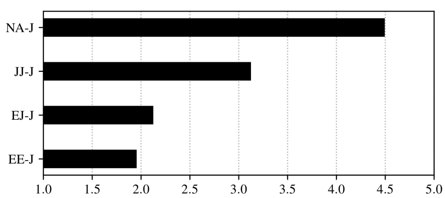


(a) Quality



(a) Similarity

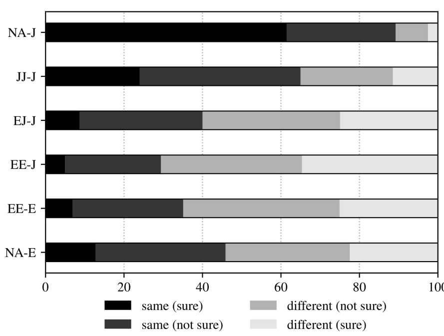
Cross-language speaker adaptation EJ-E is worse than the intra-language scenario EE-E as expected but it is enough to confirm the ability to perform cross-language adaptation and established a solid baseline.



(a) Quality

Quality and similarity of the unseen language scenarios is worse than the intra-language one with the EJ-J is slightly better than EE-J.

When presented with natural Japanese and English speech, Japanese listeners gave low score for similarity. More sophisticated test is needed to evaluate multi-lingual scenarios.



(a) Similarity

The reference is the natural Japanese utterance of the target bilingual speakers