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

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

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:
\[ z^L \sim LEnc(x; \phi^L) = p(z|x) \]
\[ \hat{y}^L = \text{Dec}(z^L; \theta_{\text{core}}, \theta_{\text{spk}}(k)) \]

Speech-to-speech stack:
\[ z^A \sim AEnc(y; \phi^A) = q(z|y) \]
\[ \hat{y}^A = \text{Dec}(z^A; \theta_{\text{core}}, \theta_{\text{spk}}(k)) \]


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.

Development procedure:

Step 1. Train the initial TTS model:
\[ \text{loss}_{\text{train}} = \text{loss}_{\text{sta}} + \beta \text{loss}_{\text{tie}} \]
\[ \text{loss}_{\text{sta}} = L_{\text{MAE}}(\hat{y}^L, y) \]
\[ \text{loss}_{\text{tie}} = L_{KLLD}(LEnc(x), AEnc(y)) \]

Step 2. Adapt to target speaker:
\[ \text{loss}_{\text{adapt}} = \text{loss}_{\text{sta}} \]
\[ \text{loss}_{\text{sta}} = L_{\text{MAE}}(\hat{y}^A; y) \]

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

Cross-lingual voice conversion?

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

(b) Similarity

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

website: www.hieuthi.com