AttS2S-VC: Sequence-to-Sequence Voice Conversion with Attention and Context Preservation Mechanisms

←Speech samples

Kou TANAKA, Hirokazu KAMEOKA, Takuhiro KANEKO, Nobukatsu HOJO (NTT corp., Japan)

Innovative R&D by NTT

⌘ Overview

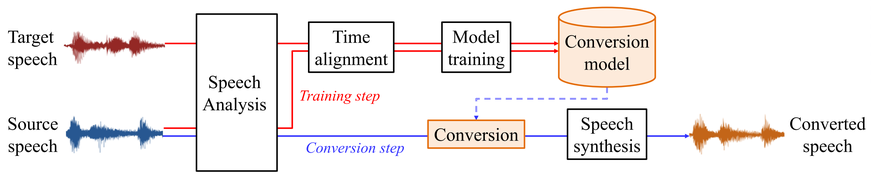
✓ Voice conversion (VC)
- Convert the para/non- linguistic information such as speaker identity, emotion, and pronunciation/accent
- Preserve the linguistic information

✓ Our contribution
- Stabilize and accelerate the training procedure by guided attention and proposed context preservation losses
- Convert not only spectral envelopes but also F0 contours and durations of speech to be converted
- Require no context information such as phoneme labels
- Require no time-aligned parallel data in advance

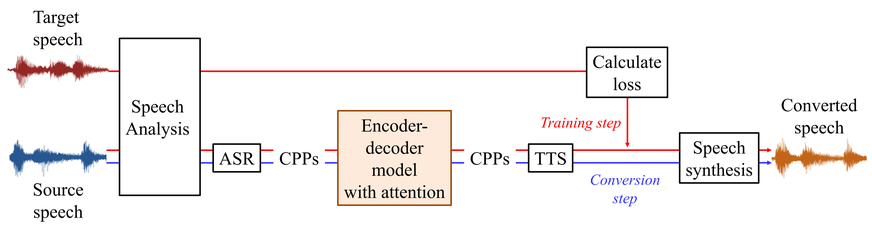
	Frame/sequence based	Sequence-to-sequence based	
	Conventional VC		Proposed VC
Acoustic feature	Converted	Converted	Converted
Duration	--	Converted	Converted
Transcript	--	Required	--

⌘ Conventional VC

✓ Frame/sequence based VC [Stylianou+, 1998] [Toda+, 2007]
- Convert spectral envelope and aperiodicity
- Use the duration of source speech as that of target speech due to no time-warping function in the conversion step
- Linear conversion of F0



✓ Sequence-to-sequence (Seq2Seq) VC [Miyoshi+, 2017] [Liu+, 2018]
- Convert all of acoustic features and durations information
- Use the same flow in both of the training and conversion step
- Automatic speech recognition (ASR) and text-to-speech (TTS)
  - Use context posterior probabilities (CPPs) corresponding to heuristically defined context information
  - Require a large number of transcripts in the training

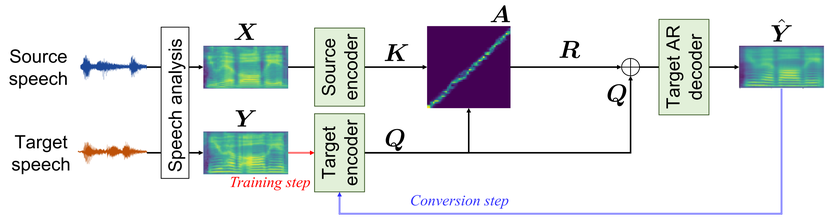


⌘ Proposed AttS2S-VC

✓ Seq2Seq model with attention mechanism
- Seq2Seq loss

$$\mathcal{L}_{\text{Seq2Seq}} = ||\hat{\boldsymbol{Y}} - \boldsymbol{Y}||_1$$

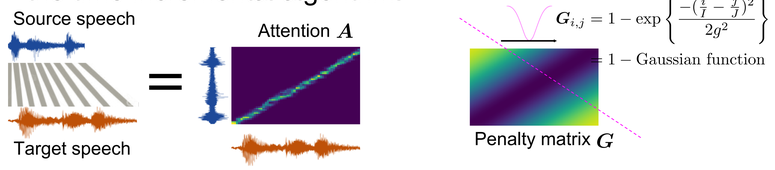
- Source & target speech: $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_I], \boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_J]$
- Source encoder: $\boldsymbol{K} = f_{\text{SrcEnc}}(\boldsymbol{X})$
- Target encoder: $\boldsymbol{Q} = f_{\text{TarEnc}}(\boldsymbol{Y})$
- Attention: $e_{i,j} = f_{\text{FFNN}}(\boldsymbol{k}_i, \boldsymbol{q}_j)$ $\quad a_{i,j} = \dfrac{\exp(e_{i,j})}{\sum_i \exp(e_{i,j})}$
  - Consider the long-range temporal dependencies between source and target sequences: $\boldsymbol{R} = \boldsymbol{K}\boldsymbol{A}$
- Target AR decoder: $\hat{\boldsymbol{Y}} = f_{\text{TarDecAR}}(\text{concat}(\boldsymbol{R}, \boldsymbol{Q}))$



✓ Stabilizing and accelerating the training procedure
- Guided attention loss [Tachibana+, 2017]

$$\mathcal{L}_{\text{ga}} = ||\boldsymbol{G} \odot \boldsymbol{A}||_1$$

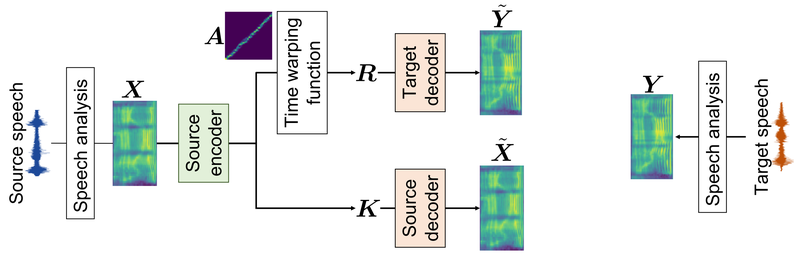
  - Accelerate the training of an attention module
  - Attention should be “nearly diagonal”. Most of the speech signal processing applications follow the time incremental algorithm.



- Proposed context preservation loss

$$\mathcal{L}_{\text{cp}} = ||\tilde{\boldsymbol{X}} - \boldsymbol{X}||_1 + ||\tilde{\boldsymbol{Y}} - \boldsymbol{Y}||_1$$

  - Guided attention loss often makes the training fail.
  - Preserve the linguistic information of source speech
  - Encode the source speech to a shared space of source and target speech



✓ Objective function: $\mathcal{L} = \mathcal{L}_{\text{Seq2Seq}} + \lambda_{\text{ga}}\mathcal{L}_{\text{ga}} + \lambda_{\text{cp}}\mathcal{L}_{\text{cp}}$

⌘ Experimental evaluation

✓ Conditions
- Dataset

Speaker	2 male speakers (rms and bdl) 2 female speakers (clb and slt)
Training	1,000 utterances per a speaker
Evaluation	Remaining 132 utterances

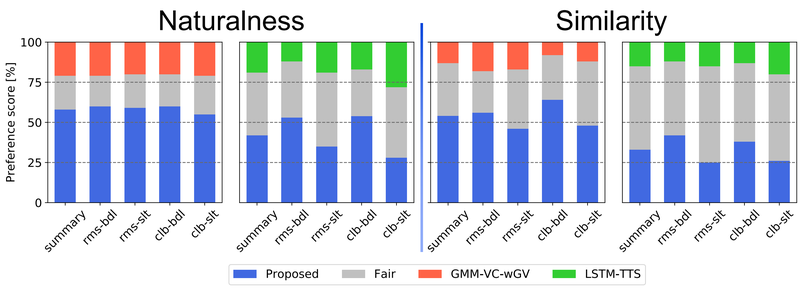
- Conventional systems

Frame/sequence based	Gaussian mixture model based VC considering global variance [Toda+, 2007]
Seq2Seq2 based	3-stacked LSTMs-based TTS assuming that ASR works perfectly [Miyoshi+, 2017]

- Model architecture of AttS2S-VC

Source encoder	CBHG inspired by Tacotron [Wang+, 2017]
Target encoder & AR decoder	3 stacked LSTMs
Attention mechanism	Additive attention [Bahdanau+, 2014]
Source decoder	CBHG
Target decoder	CBHG

✓ Results of preference tests



- Proposed method outperforms GMM based VC.
  - Durations unseen in the target speech make the conversion errors larger.
- Proposed method is comparable to LSTM based TTS.
  - Hand-crafted contextual features as input is insufficient.

※This work was supported by JSPS KAKENHI 17H01763.