22 Aug. 2023

E2E-S2S-VC:
End-to-end sequence-to-sequence voice conversion

Takuma Okamoto[1], Tomoki Toda[2,1], and Hisashi Kawai[1]

[1]National Institute of Information and Communications Technology, Japan, [2]Nagoya University, Japan

Demo samples and source code

Demo samples: Hi-Fi-CAPTAIN corpus for Japanese used in experiments

Source code based on ESPnet2-TTS
- Recipe for CMU-ARCTIC corpus
- Recipe for Hi-Fi-CAPTAIN corpus used in experiments

https://ast-astrec.nict.go.jp/demo_samples/e2e-s2s-vc/

Hi-Fi-CAPTAIN: Released!

High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT

1 female and 1 male (English): 14K utts (parallel: 13K)
1 female and 1 male (Japanese): 19K utts (parallel: 18.5K)
ESPnet2-TTS recipe for JETS-based E2E-TTS

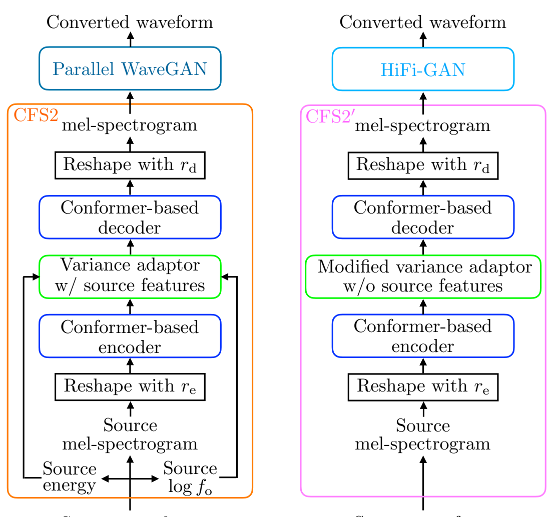
https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/

1. Introduction

- **Voice conversion (VC) methods**
  - Framewise VC based on frame-by frame conversion
    - Parallel data not required
    - Difficult to convert duration and prosody between source and target speakers
    - End-to-end models have been investigated (e.g. NVC-Net)
  - Sequence-to-sequence (S2S) VC
    - Parallel data required
    - Can convert duration and prosody by S2S manner
- **Baseline: non-autoregressive (AR) S2S-VC: CFS2+PWG**
  - Features
    - Conformer-Fastspeech 2 (CFS2)-based non-AR conversion model with Parallel WaveGAN (PWG) neural vocoder
    - Faster and more stable by non-AR structure compared with conventional Transformer-based AR models
  - Four problems
    1. Three models (teacher Transformer, CFS2, PWG) are separately trained -> they cannot be jointly optimized
    2. Unstable alignment due to teacher AR-Transformer
    3. HiFi-GAN neural vocoder outperforms PWG
    4. Energy and fundamental frequency of source speaker required

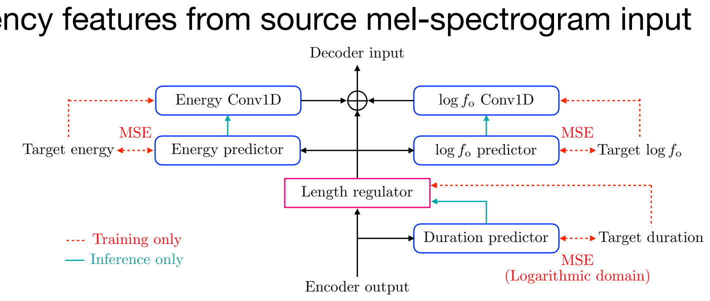
2. Extended model

- **CFS2'+HiFi-GAN**
  - CFS2': CFS2 with modified variance adapter without source energy and fundamental frequency features



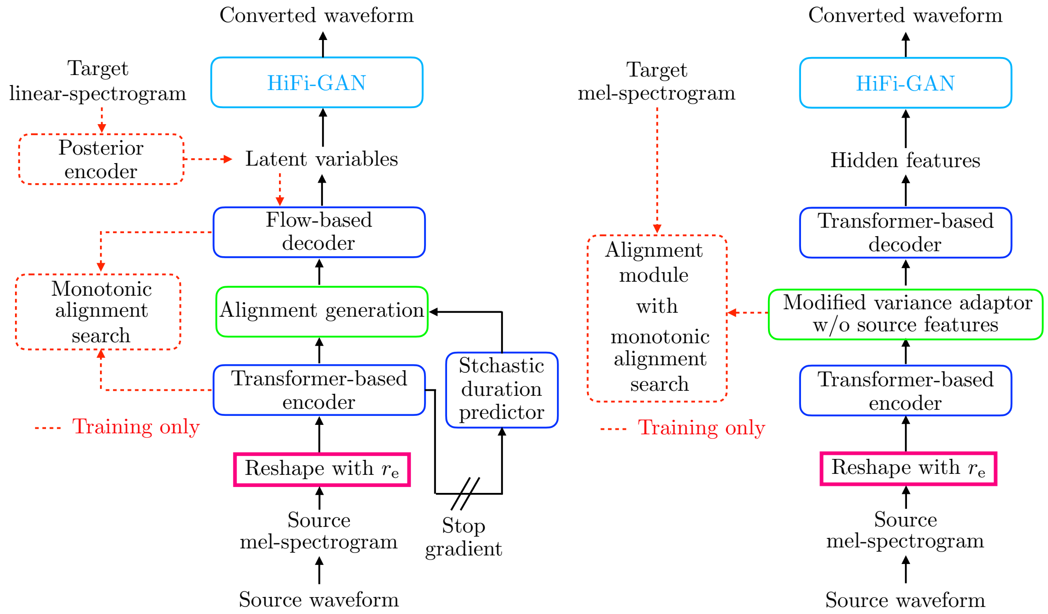
(a) CFS2+PWG (Baseline)　(b) CFS2'+HiFi-GAN (Proposed)

  - Modified valiance adapter predicts target energy and fundamental frequency features from source mel-spectrogram input



3. Proposed methods

- **End-to-end text-to-speech (E2E-TTS) models: VITS and JETS**
  - VITS: VAE + flow + HiFi-GAN + monotonic alignment search (MAS)
  - JETS: Fastspeech 2 + HiFi-GAN + MAS
- **Proposed E2E-S2S-VC: VITS-VC and JETS-VC**
  - Introducing VITS and JETS for E2E-TTS models into S2S-VC
    - Source mel-spectrogram input can be directly converted to target speech waveform with a single neural network
    - Using a reduction factor only for encoder to successfully train MAS for VC
    - Can solve all the four problems in baseline model



(c) VITS-VC (Proposed)　(d) JETS-VC (Proposed)

4. Experiments

- **Experimental conditions**
  - Dataset: Parallel 1,000 utterances for Japanese in Hi-Fi-CAPTAIN
    - Training: 950 utts, Validation: 25 utts, Evaluation: 25 utts
  - Sampling frequency: 24 kHz
  - Objective evaluation criteria: MCD, logf_oRMSE, CER and RTF
  - Subjective evaluation criteria (*N*=20): MOS, speaker similarity
- **Results of experiments**

	Male ⟶ Female			Female ⟶ Male			
Method	MCD [dB]	log f_o RMSE	CER [%]	MCD [dB]	log f_o RMSE	CER [%]	RTF
Original	N/A	N/A	1.0	N/A	N/A	1.2	
(Baseline) CFS2+PWG	5.83 ± 0.52	0.25 ± 0.07	3.4	4.74 ± 0.26	0.20 ± 0.04	4.4	3.44
CFS2'+PWG	5.50 ± 0.45	0.24 ± 0.08	3.0	4.76 ± 0.23	0.18 ± 0.06	6.8	3.41
CFS2'+HiFi-GAN (ft)	5.31 ± 0.58	**0.22 ± 0.07**	4.4	**4.49 ± 0.31**	0.19 ± 0.08	5.8	**0.72**
CFS2'+HiFi-GAN (jt)	5.95 ± 0.60	0.25 ± 0.06	12.7	4.80 ± 0.32	0.22 ± 0.08	12.5	**0.72**
VITS-VC ($r_e = 2$)	5.31 ± 0.43	0.23 ± 0.08	5.2	4.50 ± 0.30	**0.18 ± 0.05**	3.2	0.77
VITS-VC ($r_e = 3$)	5.36 ± 0.43	**0.22 ± 0.07**	5.4	4.58 ± 0.28	0.19 ± 0.06	5.8	0.76
JETS-VC ($r_e = 2$)	**5.28 ± 0.42**	0.23 ± 0.07	**2.2**	4.78 ± 0.36	0.21 ± 0.09	**2.2**	0.79
JETS-VC ($r_e = 3$)	5.38 ± 0.41	0.25 ± 0.09	2.8	4.59 ± 0.25	0.21 ± 0.09	3.0	0.78