Speech samples

Noise-Robust Voice Conversion by Conditional Denoising Training Using Latent Variables of Recording Quality and Environment

[1]Takuto Igarashi, [1]Yuki Saito, [1]Kentaro Seki, [1,2]Shinnosuke Takamichi, [3]Ryuichi Yamamoto, [3]Kentaro Tachibana, [1]Hiroshi Saruwatari

[1]The University of Tokyo, Japan, [2]Keio University, Japan, [3]LY Corp., Japan.

CDT: Conditional Denoising Training

- Noise-robust VC conditioning the VC model on speech degradation information: recording quality and environment
- Two conditioning strategies: utterance-wise and frame-wise
- Conditioning frame-wise features is essential for noise-robust VC

1. Background

- VC: converting a source speaker's timbre to that of a target speaker while preserving the linguistic content
- DNN-based VC: training DNNs for VC w/ multi-speaker corpus
- Noise-Robust VC[1]: performing well even if the input speech is degraded due to recording environment/channel
- Denoising Training[1] (DT): an end-to-end learning method for noisy-to-clean VC with noisy data augmentations

2. Conventional DT

2.1. DT algorithm

1. Generate pseudo-noisy speech by adding various noises to clean speech with random SNRs and feed it into the VC model
2. Compute norm between the clean and converted speech

Huang et al.[1] showed that DT improved VC's noise-robustness when autoencoder-based VC models were used like S2VC[2]

2.2. Limitations

- The naturalness of the converted speech is still limited when the noise of the input speech is unseen during the training.

3. Proposed CDT

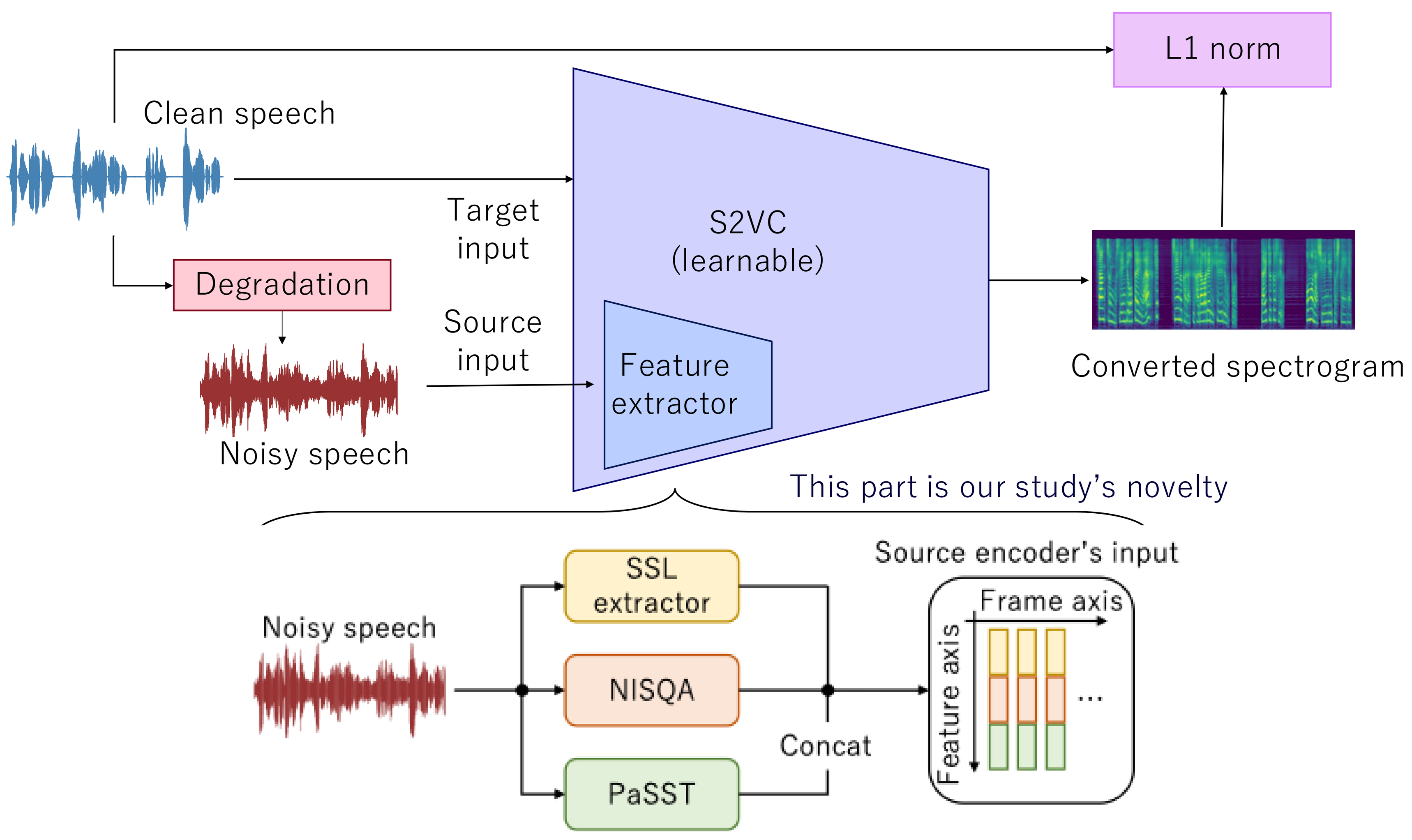
3.1. Motivation

- Make sure that the VC model explicitly learns information about speech degradation such as noise characteristics and levels

3.2 CDT algorithm

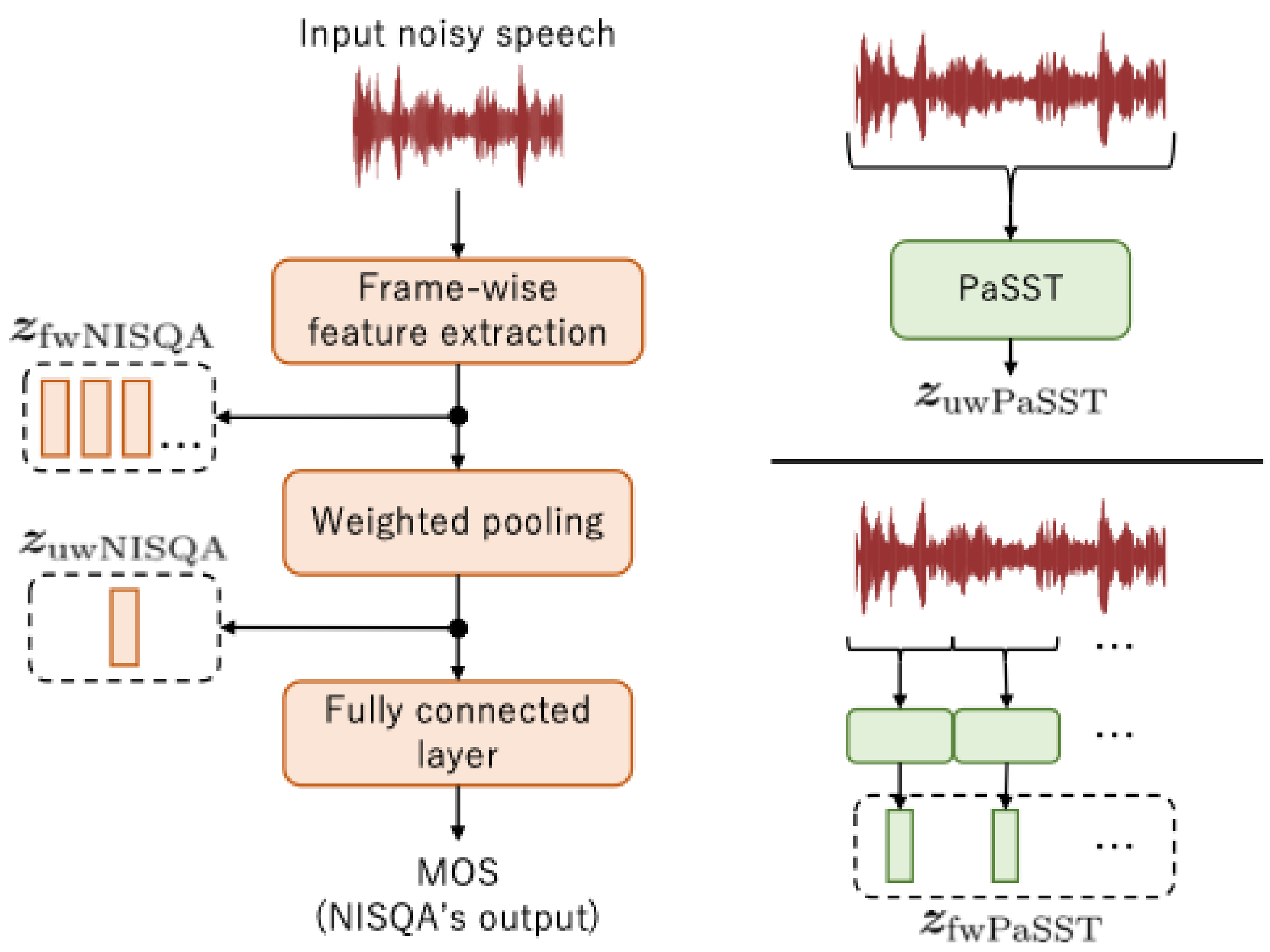
Condition the VC model on speech degradation information: recording quality, environment extracted by NISQA[3], PaSST[4]

- NISQA: automatically estimating recording quality scores
- PaSST: obtaining audio tags from input sounds



3.3 Frame-wise conditioning

- NISQA and PaSST are originally designed to output an utterance-wise (uw) prediction: MOS and audio tag
- For conditioning a VC model, frame-wise (fw) features reflecting the noise's non-stationarity are essential; obtained as follows
- 4 proposed CDT methods: xxNISQA-xxPaSST (xx=uw or fw)



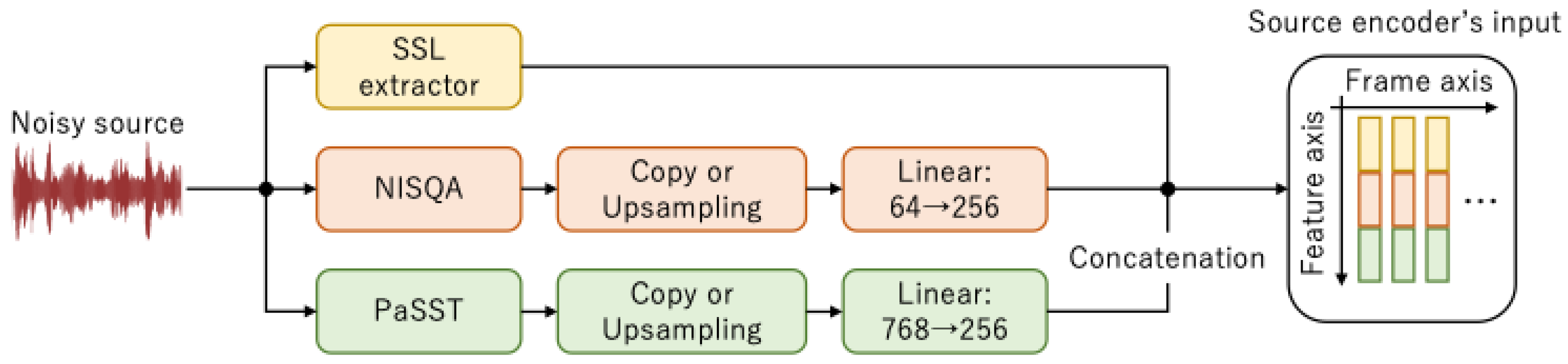
4. Experimental evaluation

4.1. Experimental conditions

Corpus: JVS[5] containing 22 hrs speech for 100 Japanese speakers

Noise: DEMAND[6] for training and WHAM![7] for test w/ [0,20]dB

- Unify the size of latent variables and concat them as follows



4.2. Objective evalution

VC on 250 test pairs of source and target speech samples

- CER: Character Error Rate showing intelligibility
- SECS: Speaker Embedding Cosine Similarity showing speaker similarity between source and target

Method		CER[%]	SECS
NISQA	PaSST		
-	-	26.3	0.935
uw	uw	27.2	**0.938**
uw	fw	24.6	0.935
fw	uw	24.7	0.931
fw	fw	**23.3**	0.934

4.3. Subjective evalution

MOS test: naturalness

Q How good is the perceptual quality of the presented VC sample?

MOS test: simlarity

Q How similar the speakers were who produced the target and VC sample?

AB test: natunalness

Q Which is the natural VC sample, A or B?

Method		Nat.	Sim.
NISQA	PaSST		
-	-	2.75	2.43
uw	uw	2.67	2.39
uw	fw	2.84	**2.50**
fw	uw	2.74	2.43
fw	fw	**2.85**	2.47

A vs B	Nat.
DT vs uw-fw	0.414 vs **0.586**
DT vs fw-fw	0.442 vs **0.558**
uw-fw vs fw-fw	0.514 vs 0.486

Conditioning on frame-wise features is effective; they represent the non-stationary characteristics of noise in the noisy source speech

References

[1] C.-Y. Huang et al., 2022. [2] J. Lin et al., 2021. [3] B. Mittag et al., 2021. [4] K. Koutini et al., 2022.
[5] S. Takamichi et al., 2020. [6] J. Thiemann et al., 2013. [7] G. Wichern et al., 2019.