Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device

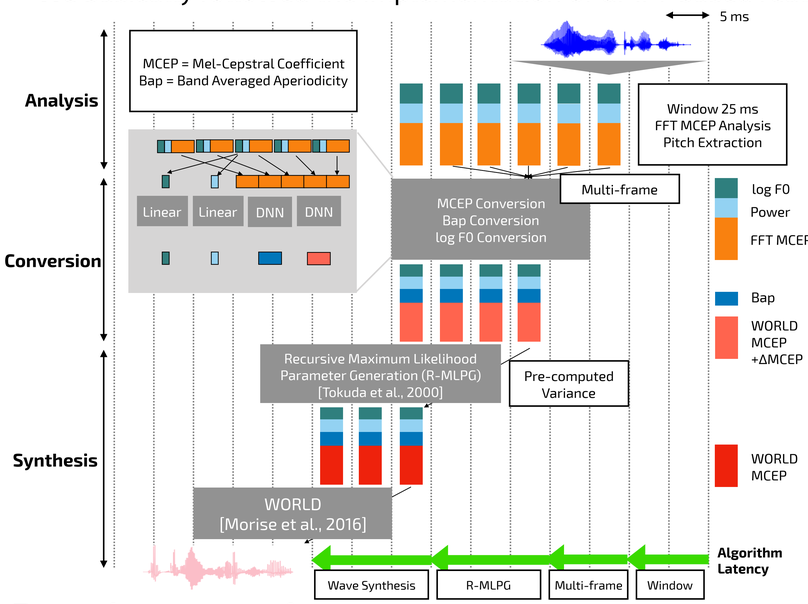
Riku Arakawa, Shinnosuke Takamichi and Hiroshi Saruwatari
The University of Tokyo, e-mail: arakawa-riku428@g.ecc.u-tokyo.ac.jp, shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

Synopsis

(1) Present an implementation for DNN-based real-time voice conversion (VC).
(2) Propose audio data augmentation for improving quality.
(3) Introduce a mask-shaped device for noise robustness.

1. Implementation of DNN-based real-time voice conversion

→ We basically followed the implementation of GMM-based real-time VC [Toda et al., 2012].



Parameters

- frame length: **25 ms**
- frame shift: **5 ms**
- multi-frame: **C = 2**
- R-MLPG buffer frames: **M = 3**
- WORLD buffer frames: **W=3**

Latency

- analysis step: **10 ms**
- conversion step: **10 ms**
- synthesis step: **30 ms**

The total latency is 50 ms, not interfering speakers.

Experiment settings

dataset	Voice Actress Corpus [y_benjo et al, 2017] 100 utterances, ~12 min, two male speakers
# of train and test	90 / 10 utterances
sampling frequency	16 kHz
speech feature	F0 / Power/ MCEP (up to 40 dim) / BAP
DNN	Three-layer Perceptron, Input-Output Residual [Saito et al., 2017]
CPU	Intel (R) Core (TM) i7-3770K CPU @ 3.50 GHz

Computation cost for the proposed VC

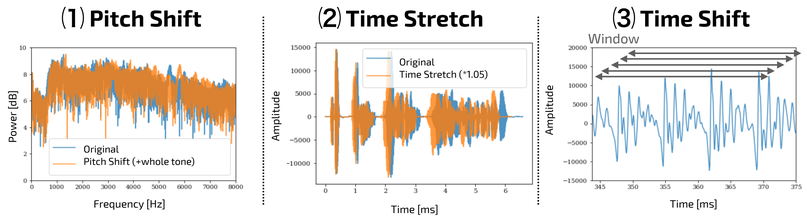
RTF (Real-Time Factor): a value of the computation time divided by the length of the input waveform

	Analysis	Conversion	Synthesis
Processing Time [ms]	0.85 ± 0.092	1.12 ± 0.51	0.57 ± 0.10
RTF	0.17	0.22	0.11

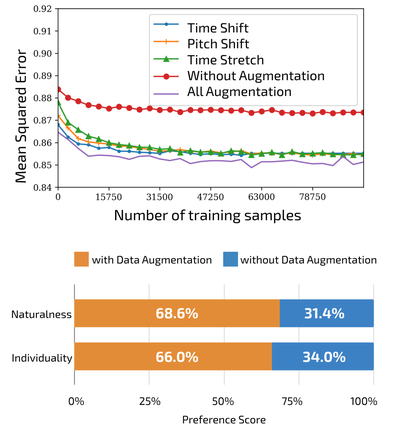
→The total RTF was smaller than 1.0, so the proposed VC system operates in real time.

2. Audio data augmentation for a quality improvement

→ To artificially augment training data by signal processing



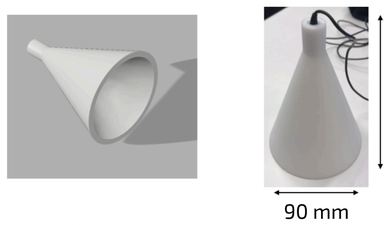
Pitch Shift : four patterns of {semitone, whole tone} {up, down}
Time Stretch: two patterns of multiplying by {0.95, 1.05}
Time Shift: four patterns of offset {-2.5, -1.25, 0, 1.25, 2.5} ms



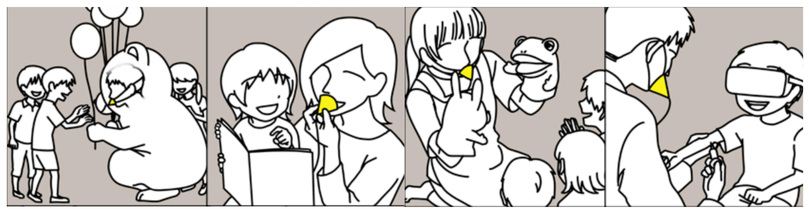
→Data augmentation improves the quality both quantitatively and qualitatively.

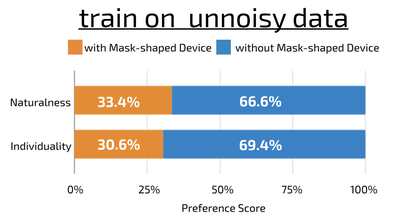
3. Mask-shaped device for noise robustness

→ To make the system robust for noisy environment



Physically block noise & confine the original speech in real-time.





35 listeners participated in evaluation on 10 utterances recorded at a noisy place.
(preference AB / XAB)

→ The physical mask is effective for naturalness of speech under noisy environment.

We'll present a further application of this mask in ACM UIST 2019 posters.
preprint: https://rikky0611.github.io/resource/transvoice_uist2019poster_paper.pdf