LY

Our project page

SRC4VC: Smartphone-Recorded Corpus for Voice Conversion Benchmark

[1]Yuki Saito, [1]Takuto Igarashi, [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.

SRC4VC: New JP Corpus for VC

- 11h of smartphone-recorded speech samples by 100 speakers
- Various styles: read-aloud, expressive, conversational, singing
- Easy way to validate any-to-any VC systems using device-recorded (& degraded) voice by users as the source speech

1. Background

Hello! → VC → Hello!

- **VC**: transforming the voice characteristics of source speech into those of target speech with unchanged phonetic content
- DNN-based VC: training DNNs for VC w/ multi-speaker corpus
- **Degradation-Robust (DR)VC**[1]: performing well even if the input speech is degraded due to recording environment/channel

Goal: promoting DRVC study by the construction of SRC4VC

2. Construction of SRC4VC

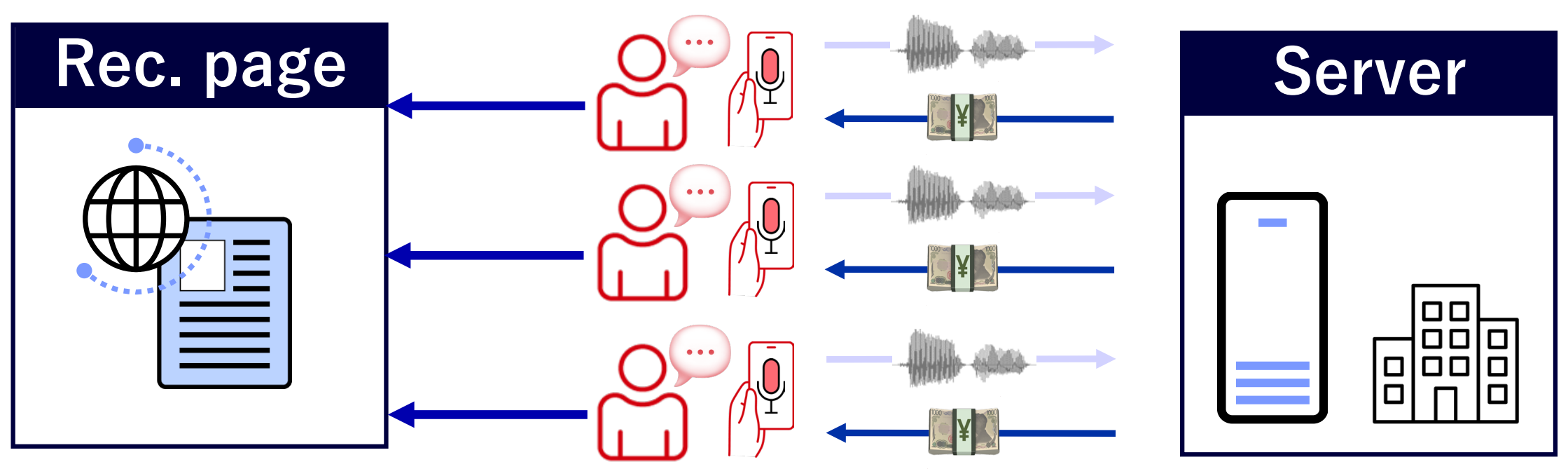
2.1. Core design

SRC4VC aims to advance various VC tasks (e.g., emotional VC[2]) and includes the following four subsets:

- **Read-aloud** 📖: 10 phoneme-balanced sentences from ITA[3]
- **Expressive** 😄😭😡: 5 sentences for each of 6 emotions (Angry, Disgust, Fear, Happy, Sadness, Surprise) from JVNV[4]
- **Conversational** 🗣️: 10 situation-oriented dialogues from STUDIES[5] (teacher-student) & CALLS[6] (operator-customer)
- **Singing** 🎵: 2 Japanese copyright-free songs ("katatsumuri" = child-song & "Shining star" = J-POP)

2.2. Voice recording by crowdworkers

1. Preparing a webpage containing the recording instruction, start/stop button, and text w/ pronunciation
2. Recruiting speakers through crowdsourcing (Lancers)
3. Asking the recruited speakers to record their voice samples using smartphones in a quiet room as possible



2.3. Annotation by crowdworkers

- **Speaker-wise recording quality**: recruiting 400 annotators who rated recording-quality of randomly presented 25 read-aloud samples (i.e., performing recording-quality MOS test)

Q How good is the recording quality of the presented voice?

- **Utterance-wise perceived emotion**: recruiting 500 annotators who labeled emotion for each of "Expressive" & "Conversational" samples (5 annotations per sample)

Q What do you think the emotion of the presented voice is?
Hap
Sad
Neu
Fea
…

References

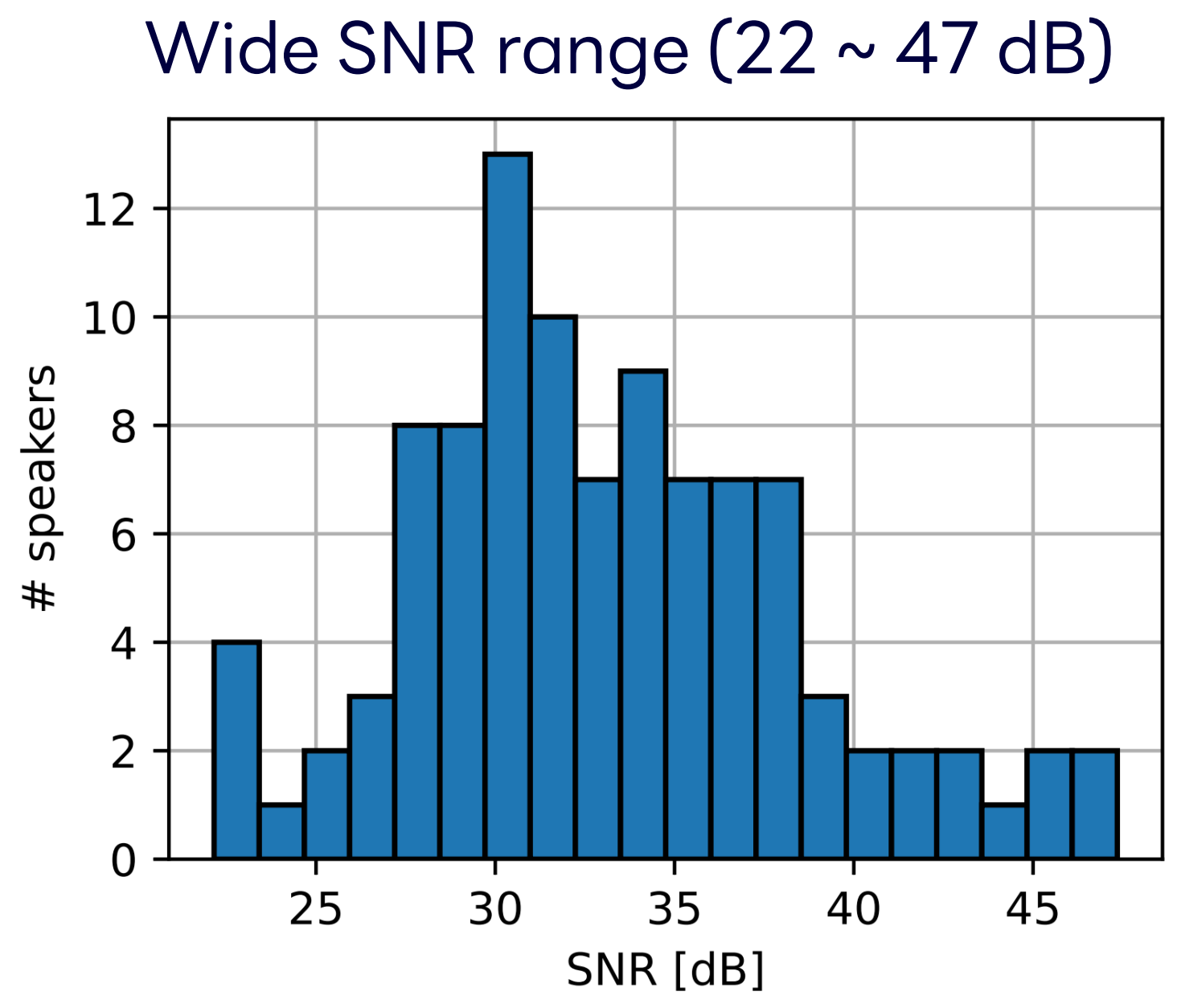
[1] C.-Y. Huang et al., 2022. [2] K. Zhou et al., 2022. [3] J. Koguchi et al., 2021. [4] D. Xin et al., 2024.
[5] Y. Saito et al., 2022. [6] Y. Saito et al., 2023. [7] J. Yamagishi et al., 2019. [8] H. Li et al., 2022.
[9] S. Takamichi et al., 2020. [10] S. Takamichi et al., 2018. [11] G. Mittag et al., 2021. [12] J. Lin et al., 2021.
[13] J. Kong et al., 2020.

3. Corpus Analysis

3.1. Corpus specification

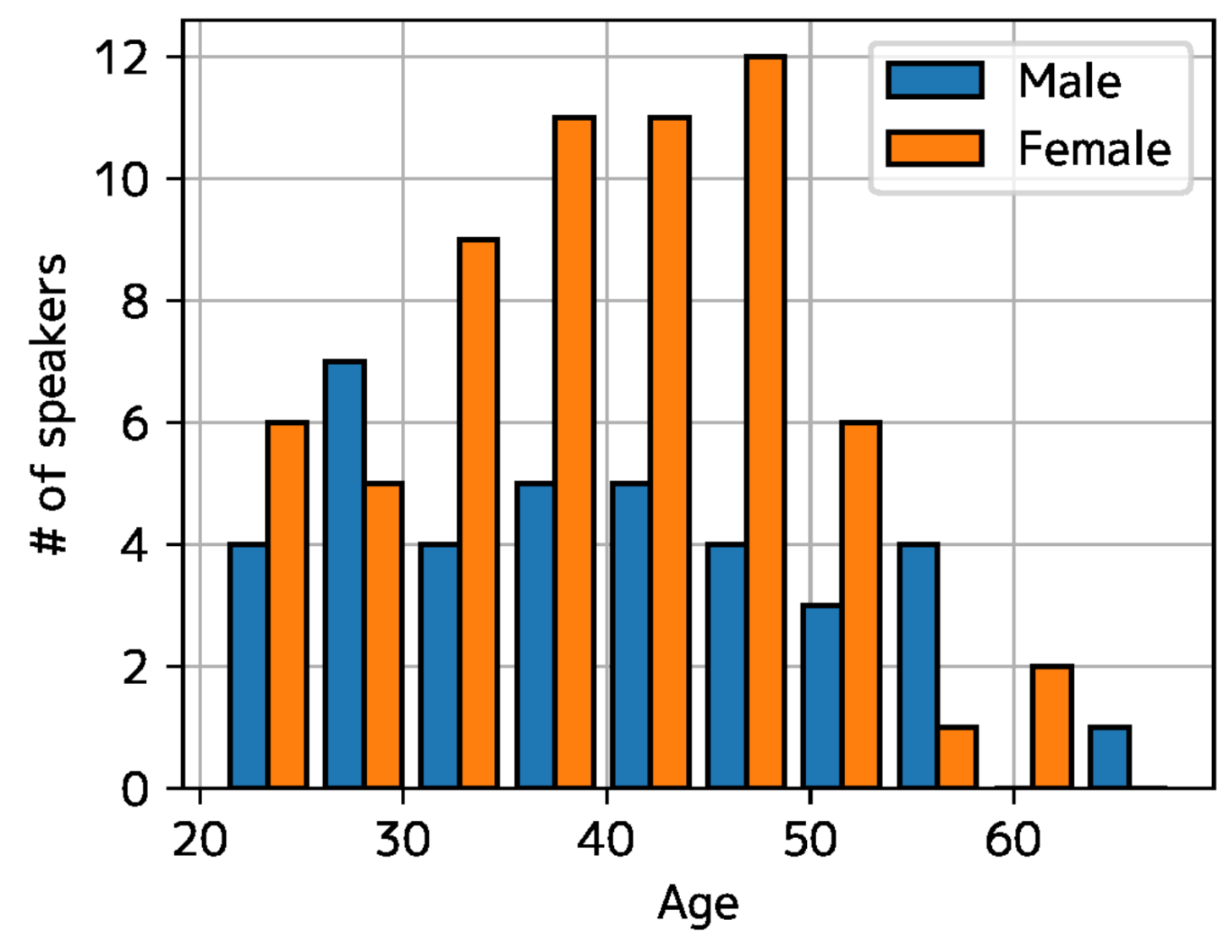
Subset	# samples	Hours
Read-aloud	1,000	1.46
Expressive	3,000	7.16
Conversational	1,000	1.66
Singing	200	0.87
Total	**5,200**	**11.14**

Wide SNR range (22 ~ 47 dB)

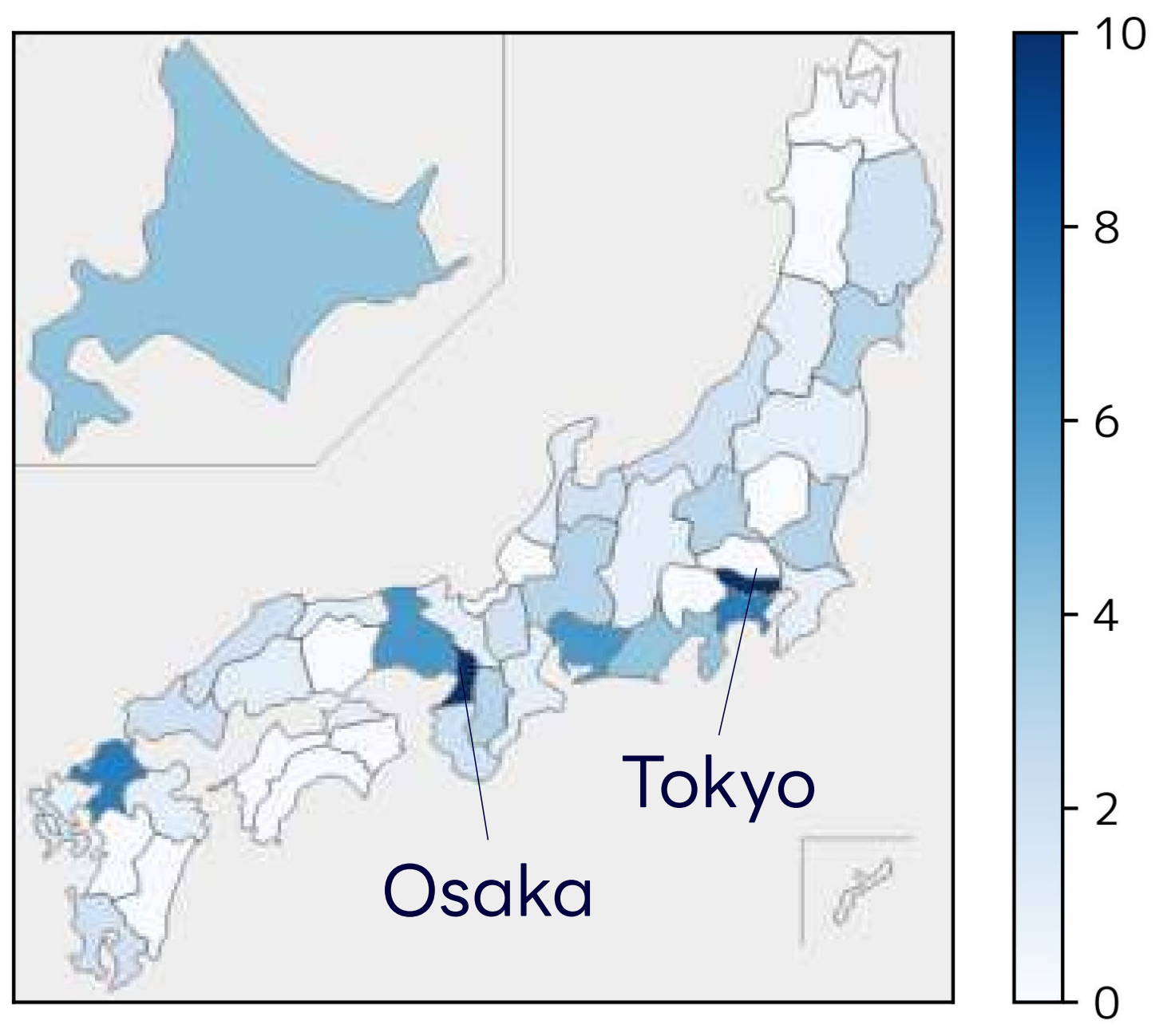


3.2. Speaker distribution

37 males & 63 females
of various ages (21 ~ 68)



From 34/47 JP prefectures



3.3. Comparison w/ existing corpora

Corpus	# styles	Lang.	Dur. [h]	# spkrs.	Recording
VCTK[7]	1	EN	44	109	Studio
DDS[8]	1	EN	2,000	48	Device
JVS[9]	3	JP	30	100	Studio
CPJD[10]	1	JP	7	22	Device
SRC4VC	4	JP	11	100	Smartphone

3.4. Annotation results

Spearman's **R**ank **C**orrelation **C**oefficient (**SRCC**) between human-annotated recording-quality MOS & each NISQA score[11]

Noisiness	Coloration	Discontinuity	Loudness	Naturalness
0.15	**0.67**	**0.62**	0.36	0.54

Due to frequency response & non-linear distortion

% of agreed emotional samples (see below for the definition)

Subset	Ang	Dis	Fea	Hap	Sad	Sur	Neu
😄😭😡	14.6	17.8	14.4	16.7	15.7	17.3	0.35
🗣️	0.45	0.29	0.08	0.59	0.56	0.28	1.08

→ (Ang, Ang, Ang, Ang, Hap) → **Agreed: "Ang"**

4. Any-to-Any VC Experiment

4.1. Setup (see our paper for the details)

- Baseline VC model: S2VC[12] + HiFi-GAN vocoder[13] (following the same setup as existing DRVC study[1])
- Data: JVS for training, SRC4VC for evaluation

4.2. Naturalness/similarity MOS tests

- # listeners: 200 for each (30 samples/listener)

	Method	Nat.	Sim.
	Baseline (B)	2.54	2.17
Training the Baseline w/ **D**ata **A**ugmentation	B+DA (noise)	2.59	2.18
	B+DA (reverb)	2.66	2.22
	B+DA (band)	2.62	2.17
Cascading the Baseline w/ **S**peech **E**nhancement	B+SE (Demucs[14])	2.53	2.17
	B+SE (Miipher[15])	2.74	2.21

SE reasonably works as preproceesing for DRVC!