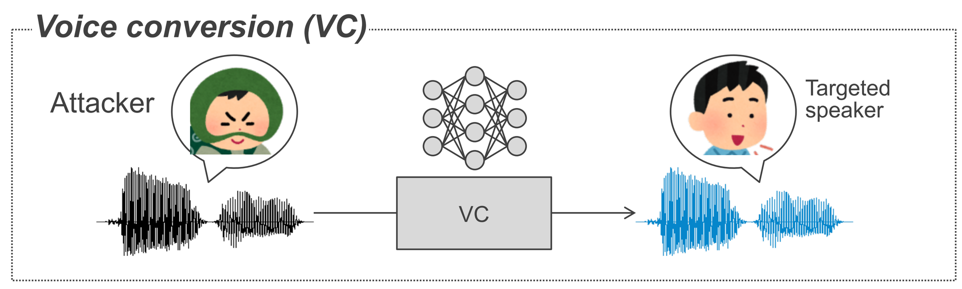
# V2S Attack: Building DNN-Based Voice Conversion from Automatic Speaker Verification

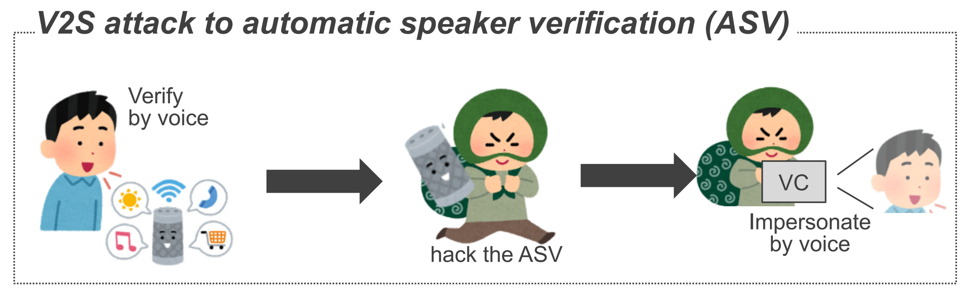
Taiki Nakamura[†], Yuki Saito[†], Shinnosuke Takamichi[†], Yusuke Ijima[‡], and Hiroshi Saruwatari[†]
[†]The University of Tokyo, Japan [‡]NTT Corporation, Japan

## 1. Introduction: Verification-to-Synthesis (V2S) Attack

Automatic speaker verification (ASV)[1]
✔ identifies the speaker of the input voice.
→ If an attacker hacks the ASV, voices of enrolled speakers risk being reproduced.

Voice conversion (VC)
✔ predicts targeted speaker's voice.
→ VC is a possible technique used in impersonation attack.




Deceiving the ASV has some possibility of reproducing the targeted speaker's individuality by VC. We name this attack "verification-to-synthesis (V2S) attack".

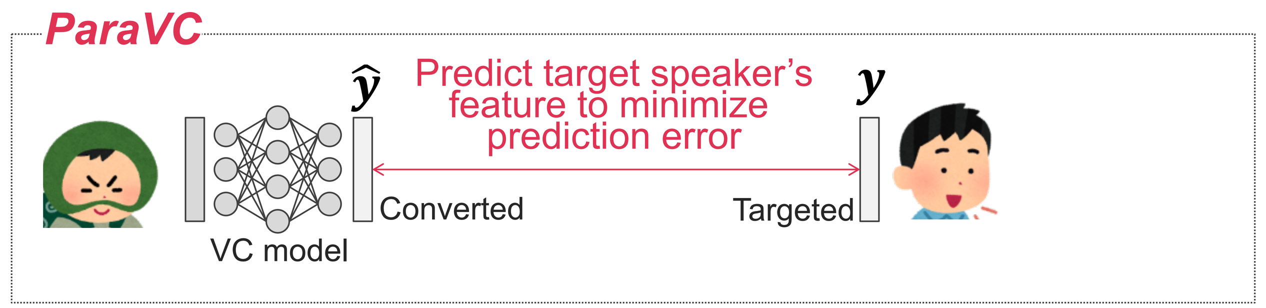
■ Our approach
✔ proposes VC training with the "white-boxed" ASV and pre-trained automatic speech recognition (ASR) models without the targeted speaker's voice data.

**Proposed VC performs comparably to the standard VC methods using a tiny amount of parallel voice data.**

## 2. Standard VC Using Targeted Speaker's Voices

### (1) One-to-one parallel VC (ParaVC)[2]

✔ is trained to minimize mean squared error (MSE) betw. $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ using a targeted speaker's voice.




### (2) One-to-many non-parallel VC (NonparaVC)[3]

✔ can convert an attacker voice to any arbitrary speaker's voice.
✔ is often trained using multi-speaker corpora in advance.

Two standard VC models are trained using a targeted speaker's voice. These are used as references to evaluate the performances of the proposed speaker V2S attack.

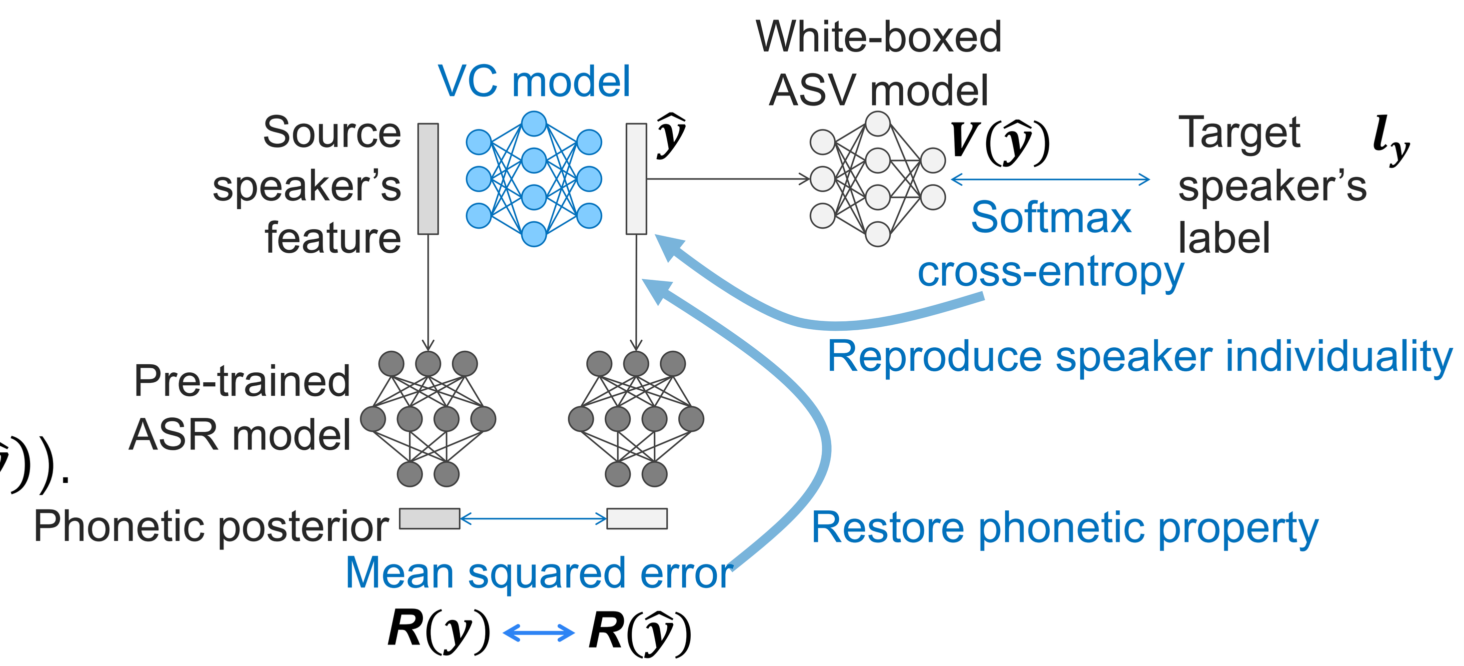
## 3. V2S Attack: VC without Using Targeted Speaker's Voices

### V2S attack model

✔ uses two DNN models for training the VC
→ Loss can be backpropagated to VC model.

■ White-boxed ASV model $\boldsymbol{V}(\cdot)$
→ Targeted speaker's label ($\boldsymbol{l}_{\boldsymbol{y}}$) is given.
→ $\boldsymbol{V}(\cdot)$ estimates speaker similarity between input voices ($\hat{\boldsymbol{y}}$) & targeted speaker's voices as softmax cross-entropy, $L_{SCE}(\boldsymbol{l}_{\boldsymbol{y}}, \boldsymbol{V}(\hat{\boldsymbol{y}}))$.
→ It helps to reproduce the targeted speaker's individuality, but does not keep the phonetic property of the input voice.

■ Pre-trained automatic speech recognition (ASR) model $\boldsymbol{R}(\cdot)$
→ $\boldsymbol{R}(\cdot)$ estimates the discrepancy as MSE between $\boldsymbol{R}(\boldsymbol{x})$ and $\boldsymbol{R}(\hat{\boldsymbol{y}})$.
→ It helps to restore the phonetic property of the input voice.




### Loss function

$$L\left(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{l}_{\boldsymbol{y}}\right) = L_{SCE}\left(\boldsymbol{l}_{\boldsymbol{y}}, \boldsymbol{V}(\hat{\boldsymbol{y}})\right) + \underset{\text{hyperparameter}}{\underset{\uparrow}{\omega}} L_{MSE}(\boldsymbol{R}(\boldsymbol{x}), \boldsymbol{R}(\hat{\boldsymbol{y}}))$$

## 4. Experimental Evaluation

### Experimental conditions

| | |
|---|---|
| Compared model | (a) ParaVC: trained by {5, 10, 30} utterances<br>(b) NonparaVC: trained by 260 pre-stored speakers<br>(c) V2S: trained by 200 utterances of attacker |
| The number of enrolled speakers | 260 Japanese speakers (130 males and 130 females) |
| Speech params. (including Δ) | 39-dim. mel-cepstral coefficients, Log F0, 10-dim. bap |
| DNN architectures | Feed-Forward (see our paper) |
| Attacker and Targeted speakers | one attacker (one male) & four targeted speakers (two males and two females) |
| Evaluation data | 25 parallel voices |

### Subjective evaluation

Naturalness (preference AB tests)

**male-to-male**

| A | Scores | *p*-value | B |
|---|---|---|---|
| ParaVC ( 5 utts) | 0.388 vs. **0.612** | $1.221 \times 10^{-10}$ | V2S |
| ParaVC (10 utts) | 0.475 vs. 0.525 | 0.158 | V2S |
| ParaVC (30 utts) | 0.458 vs. **0.542** | 0.016 | V2S |
| NonparaVC | **0.598** vs. 0.402 | $2.694 \times 10^{-8}$ | V2S |

**male-to-female**

| A | Scores | *p*-value | B |
|---|---|---|---|
| ParaVC ( 5 utts) | 0.490 vs. 0.510 | 0.572 | V2S |
| ParaVC (10 utts) | **0.593** vs. 0.407 | $1.365 \times 10^{-7}$ | V2S |
| ParaVC (30 utts) | **0.610** vs. 0.390 | $3.174 \times 10^{-10}$ | V2S |
| NonparaVC | **0.538** vs. 0.462 | 0.034 | V2S |

V2S attack ≧ ParaVC (5 utts)

Speaker individuality (preference XAB tests)

**male-to-male**

| A | Scores | *p*-value | B |
|---|---|---|---|
| ParaVC ( 5 utts) | 0.530 vs. 0.470 | 0.090 | V2S |
| ParaVC (10 utts) | **0.615** vs. 0.385 | $< 10^{-10}$ | V2S |
| ParaVC (30 utts) | **0.675** vs. 0.325 | $< 10^{-10}$ | V2S |
| NonparaVC | **0.660** vs. 0.340 | $< 10^{-10}$ | V2S |

**male-to-female**

| A | Scores | *p*-value | B |
|---|---|---|---|
| ParaVC ( 5 utts) | **0.585** vs. 0.415 | $1.324 \times 10^{-6}$ | V2S |
| ParaVC (10 utts) | **0.713** vs. 0.287 | $< 10^{-10}$ | V2S |
| ParaVC (30 utts) | **0.705** vs. 0.295 | $< 10^{-10}$ | V2S |
| NonparaVC | **0.588** vs. 0.412 | $< 10^{-10}$ | V2S |

V2S attack = ParaVC (5 utts)

## 5. Conclusion

V2S attack: voice impersonation attack using VC
✔ uses ASV, and ASR model for VC training.
✔ is trained without the targeted speaker's voices.

Experimental result
→ V2S attack can synthesize voices that has naturalness and speaker individuality comparable to a standard parallel VC with a tiny amount of data.

We are planning to
✔ improve the performances of the V2S attack.
✔ investigate ways of preventing the V2S attack.

[References]
[1] Dehak, N. et al., IEEE Transactions on ASLP, 2011 [2] Toda, T. et al., IEEE Transactions on ASLP, 2007 [3] Saito, Y. et al., Proc. ICASSP, 2018