# Characterizing the Adversarial Vulnerability of Speech Self-Supervised Learning


Haibin Wu[1,2*], Bo Zheng[2,3*], Xu Li[3], Xixin Wu[2,3], Hung-yi Lee[1], Helen Meng[2,3]

[1]Graduate Institute of Communication Engineering, National Taiwan University

[2]Centre for Perceptual and Interactive Intelligence, The Chinese University of Hong Kong

[3]Human-Computer Communications Laboratory, The Chinese University of Hong Kong



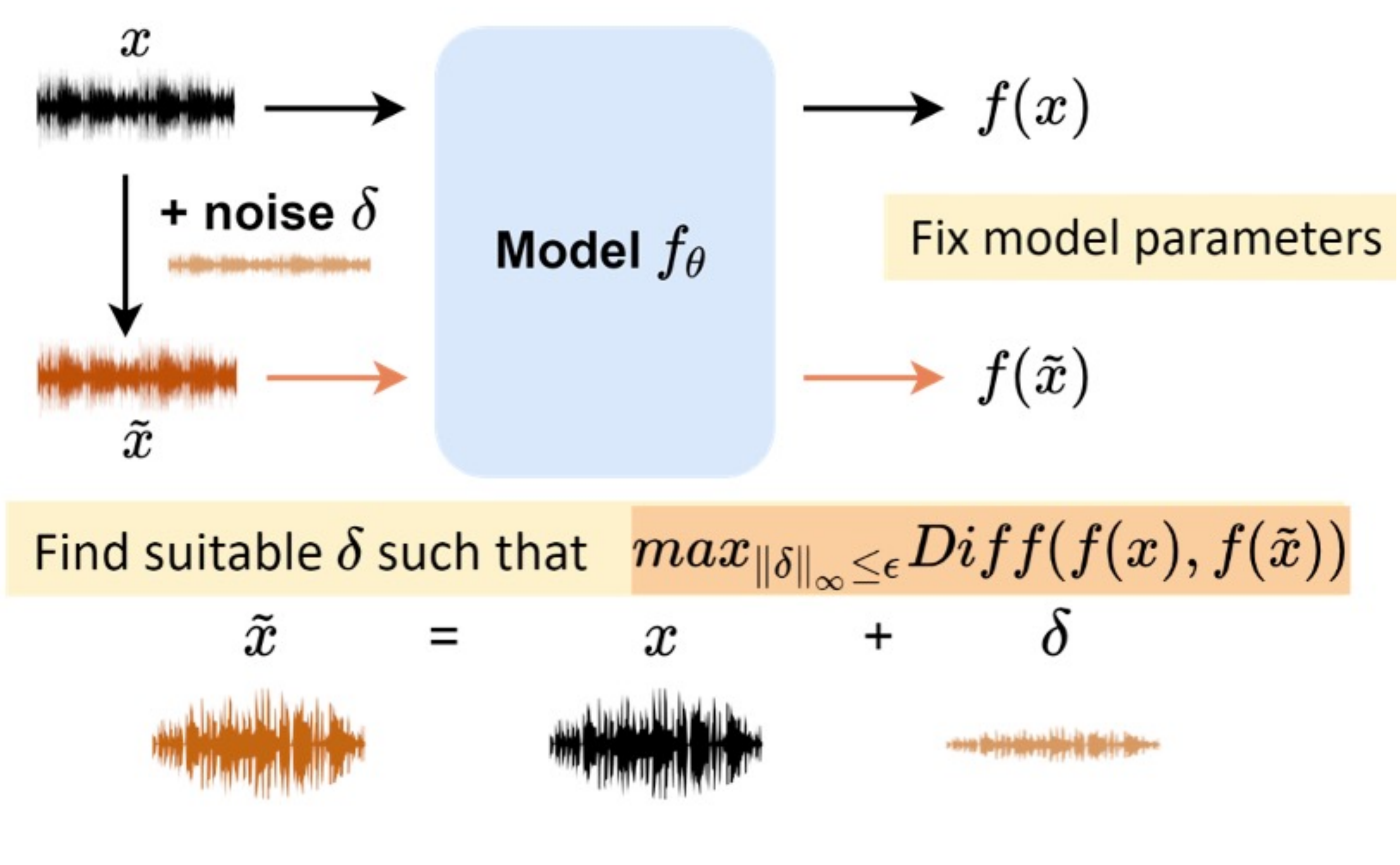
Paper #: 3287

icassp 2022 Singapore


## Introduction

- Motivation
  - The SUPERB shows speech SSL models improve the performance of various downstream tasks
  - The paradigm of the self-supervised learning upstream model followed by downstream tasks arouses more attention in the speech community
  - Characterizing the adversarial robustness of such paradigm is of high priority

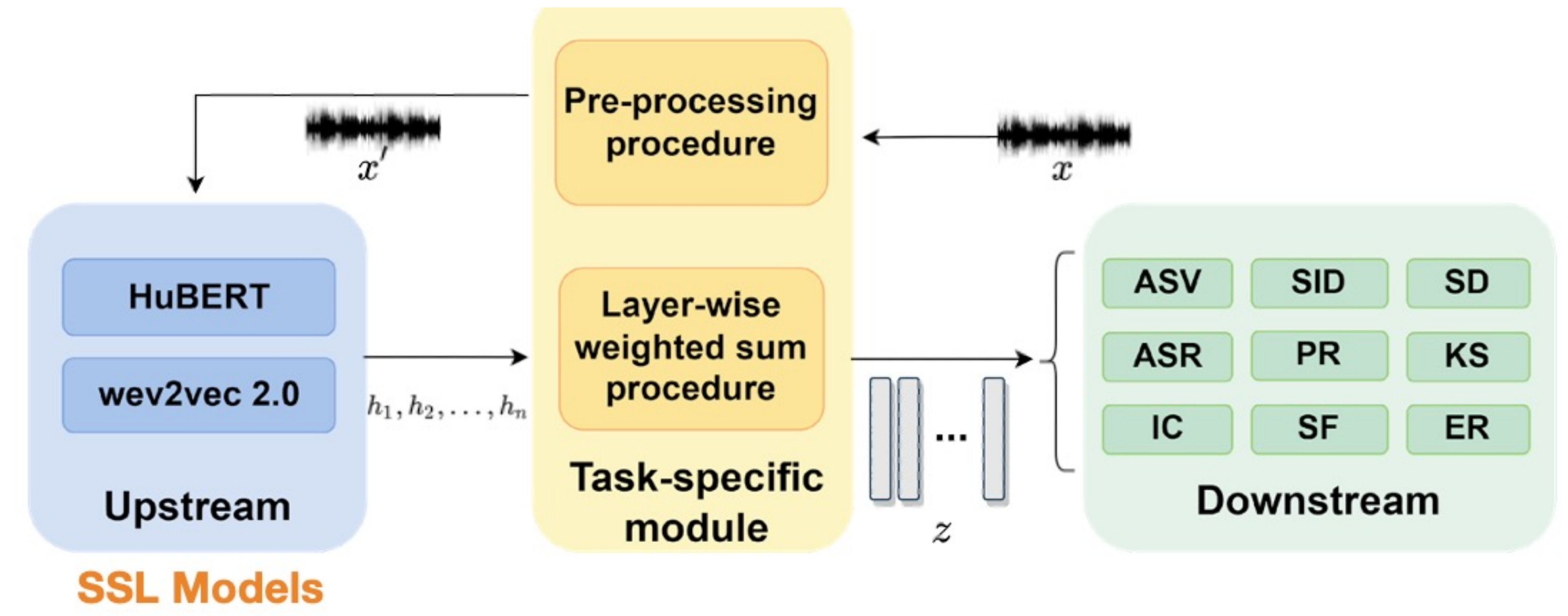
## Adversarial Attack




## Self-supervised learning models

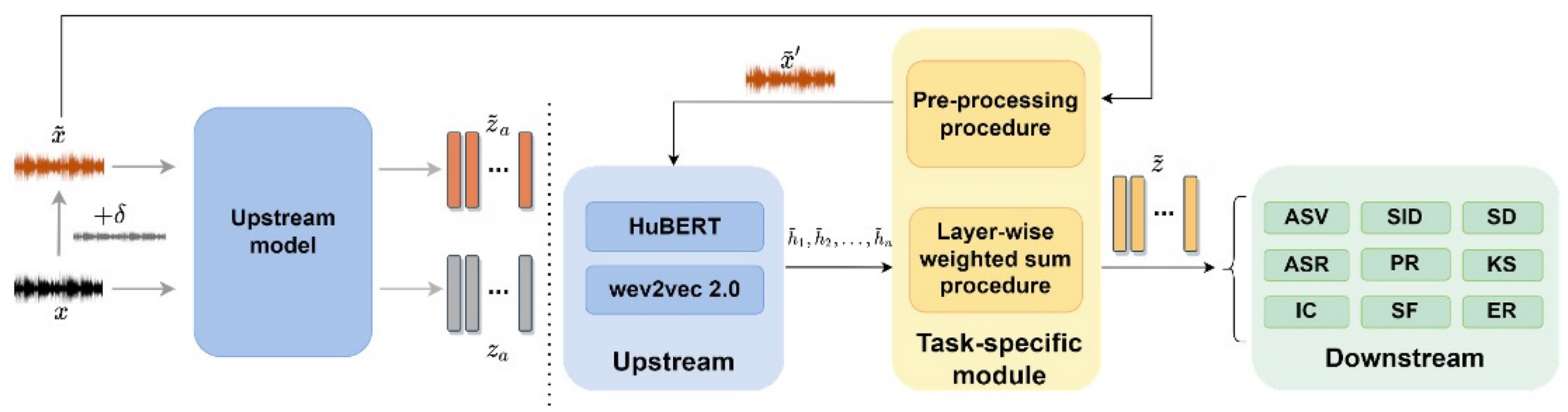
- HuBERT adopts BERT-style token classification for pre-training
- wav2vec 2.0 learns general-purpose knowledge by contrastive loss
- Both HuBERT and wav2vec 2.0 get the excellent performance in all the downstream tasks in the settings of SUPERB
- Thus, we select HuBERT and wav2vec 2.0 to conduct adversarial attacks

## Upstream-downstream paradigm

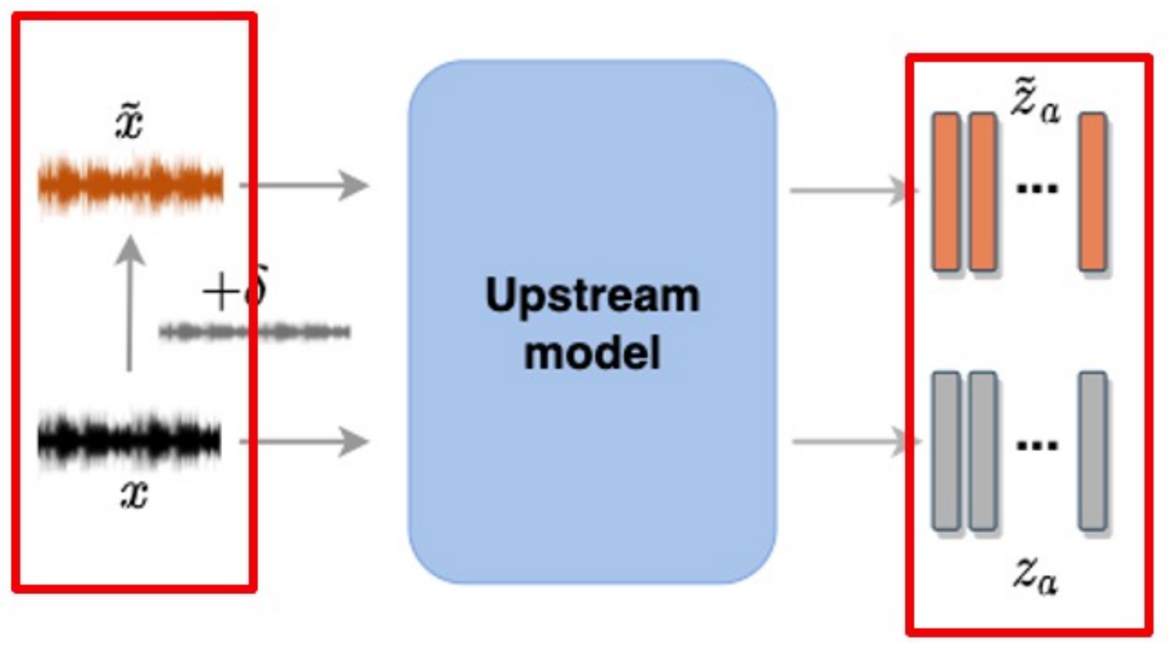



## Attack Framework

| | Zero-knowledge attackers | Limited-knowledge attackers |
|---|---|---|
| Target upstream model | ✗ | ✓ |
| Target downstream model | ✗ | ✗ |
| Layer-wise weighted sum procedure | ✗ | ✗ |
| Preprocessing procedure | ✗ | ✗ |




## Attack Strategy




**Basic Iterative Method (BIM)**

$$x^{n+1} = clip_{x,\epsilon}(x^n + \delta), \quad for\ n = 0, \ldots, N-1$$

$$\delta = \alpha \times sign(\nabla_{x^n} \|z_a - \tilde{z}_a\|_2)$$

## Experimentals

### Experimental Setup

- Randomly select 100 genuine samples for attack, and repeat the experiments three times.
- Gaussian noise of the same noise-to-signal ratio (NSR) with adversarial perturbations is introduced as baseline for comparison

**Table 1**. Adversarial attack performance on SSL representations for various downstream tasks.

| | | ASR | PR | KS | IC | SF | | SID | ER | SD | | ASV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | WER ↓ | PER ↓ | Acc ↑ | Acc ↑ | F1 ↑ | CER ↓ | Acc ↑ | Acc ↑ | Acc ↑ | DER ↓ | Acc ↑ |
| (a) | w2v2-w2v2 | 19.20[1] (±2.01) | 28.32 (±2.03) | 65.67 (±6.51) | 55.67 (±5.77) | 88.55 (±1.33) | 20.19 (±2.05) | 81.33 (±3.06) | 79.33 (±3.79) | 88.48 (±0.19) | 17.48 (±0.55) | 91.67 (±2.31) |
| (b) | HuBERT-w2v2 | 5.54 (±0.71) | 5.09 (±0.47) | 91.00 (±3.00) | 88.33 (±1.15) | 95.36 (±1.26) | 8.70 (±0.55) | 87.67 (±4.16) | 87.33 (±6.03) | 94.56 (±0.36) | 8.08 (±0.41) | 97.00 (±2.00) |
| (c) | gau-w2v2 | 0.48 (±0.06) | 1.11 (±0.05) | 98.67 (±0.58) | 93.67 (±1.15) | 99.71 (±0.27) | 0.71 (±0.50) | 97.67 (±2.08) | 95.67 (±3.06) | 98.24 (±0.09) | 2.51 (±0.11) | 99 (±0.00) |
| (d) | Clean-w2v2 | 0 | 0 | 100 | 100 | 100 | 0 | 100 | 100 | 98.24 | 2.51 | 100 |
| (e) | HuBERT-HuBERT | 26.76 (±0.82) | 18.67 (±1.54) | 64.33 (±0.58) | 69.67 (±5.03) | 76.91 (±1.67) | 36.54 (±1.83) | 76.33 (±4.93) | 78.33 (±2.08) | 87.78 (±0.83) | 18.39 (±1.65) | 88.33 (±2.08) |
| (f) | w2v2-HuBERT | 1.94 (±0.06) | 2.21 (±0.28) | 96.67 (±1.15) | 98.33 (±1.15) | 99.42 (±0.37) | 1.62 (±0.16) | 93.67 (±1.15) | 91.00 (±2.65) | 95.13 (±0.20) | 7.17 (±0.47) | 96.67 (±1.53) |
| (g) | gau-HuBERT | 0.05 (±0.08) | 0.42 (±0.12) | 99.67 (±0.58) | 99.67 (±0.58) | 99.89 (±0.19) | 0.25 (±0.24) | 98.67 (±2.31) | 99.00 (±0.00) | 98.36 (±0.09) | 2.32 (±0.13) | 99.67 (±0.58) |
| (h) | Clean-HuBERT | 0 | 0 | 100 | 100 | 100 | 0 | 100 | 100 | 98.37 | 2.31 | 100 |

- Simply adding Gaussian noise cannot degrade a well-trained system for the attack purpose
- Zero-knowledge attackers achieve relatively weaker attacks on downstream tasks than limited-knowledge attackers
- Limited-knowledge attackers achieve the most effective attack as shown in (a) and (e)

### XAB test

- Five listeners take part in the XAB listening test.
- The XAB test has a classification accuracy of 58.89%, which shows the adversarial samples are hard to be distinguished from genuine samples.
- Demo: https://bzheng1024.github.io/adv-audio-demo/index.html

## Conclusion

- In this paper, we do some preliminary works to expose the vulnerability of the SUPERB paradigm to adversarial attacks
- For the future work, we will investigate attacks with higher transferability and less imperceptibility
- The long-term goal is to come up with adaptive defense methods that offer protection against dangerous attacks.

## Acknowledgement


- The Centre for Perceptual and Interactive Intelligence, an InnoCentre of The Chinese University of Hong Kong
- This work was done while Haibin Wu was a visiting student at the CUHK
- Google PHD Fellowship Scholarship