# NPU-NTU System for Voice Privacy 2024 Challenge


- **Jixun Yao**[1], Nikita Kuzmin[2], Qing Wang[1], Pengcheng Guo[1], Ziqian Ning[1], Dake Guo[1] , Kong Aik Lee[3], Eng-Siong Chng[2], Lei Xie[1]

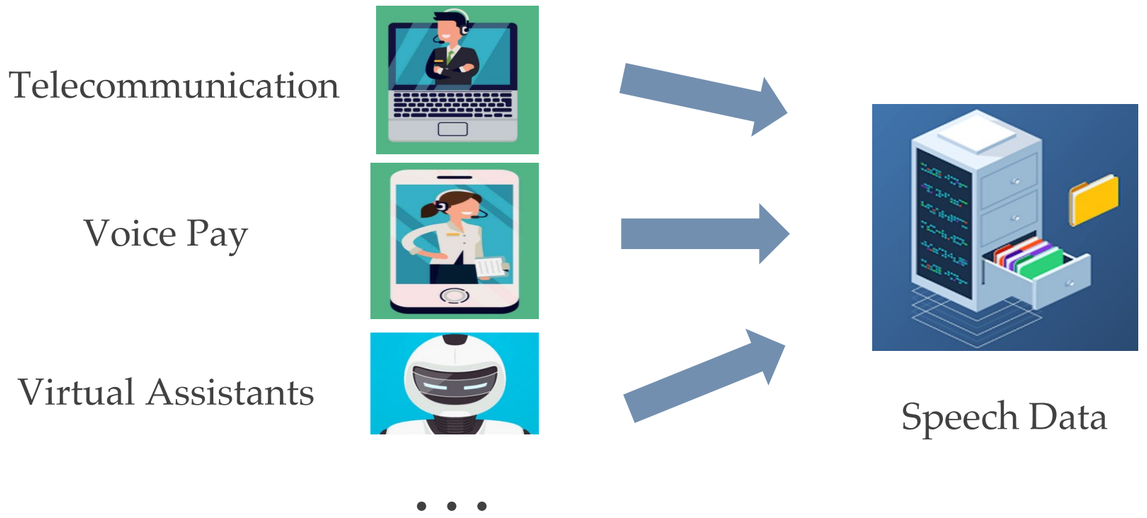
1. Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China
http://www.npu-aslp.org
2. Nanyang Technological University
3. The Hong Kong Polytechnic University



# Background

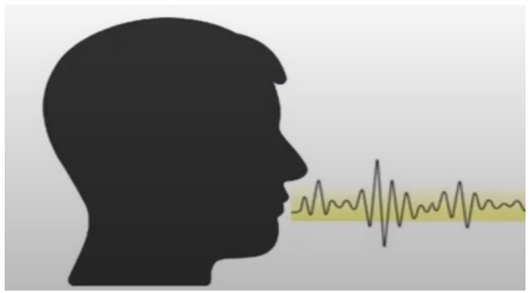
## Speech Data

- Speech data are proliferating exponentially
- Applications record personal speech data which have risk to be stolen by attacker




## Sensitive Information

- Speech data contain rich personal sensitive information



- Age
- Gender
- Health State
- Religious Beliefs
- Other sensitive attributes

# Privacy Protection

- **Speaker Anonymization**
  - Implemented before users share their speech data
  - Effectively remove a speaker's identity while preserving the linguistic information and para-linguistic information
- **VoicePrivacy Challenge**
  - Provide baseline systems, evaluation metrics and pipeline
  - preserving the emotional state, a key paralinguistic attribute


**Original Speaker**: *Today is a nice day* (Happy)

**Pseudo Speaker** : *Today is a nice day* (Happy)

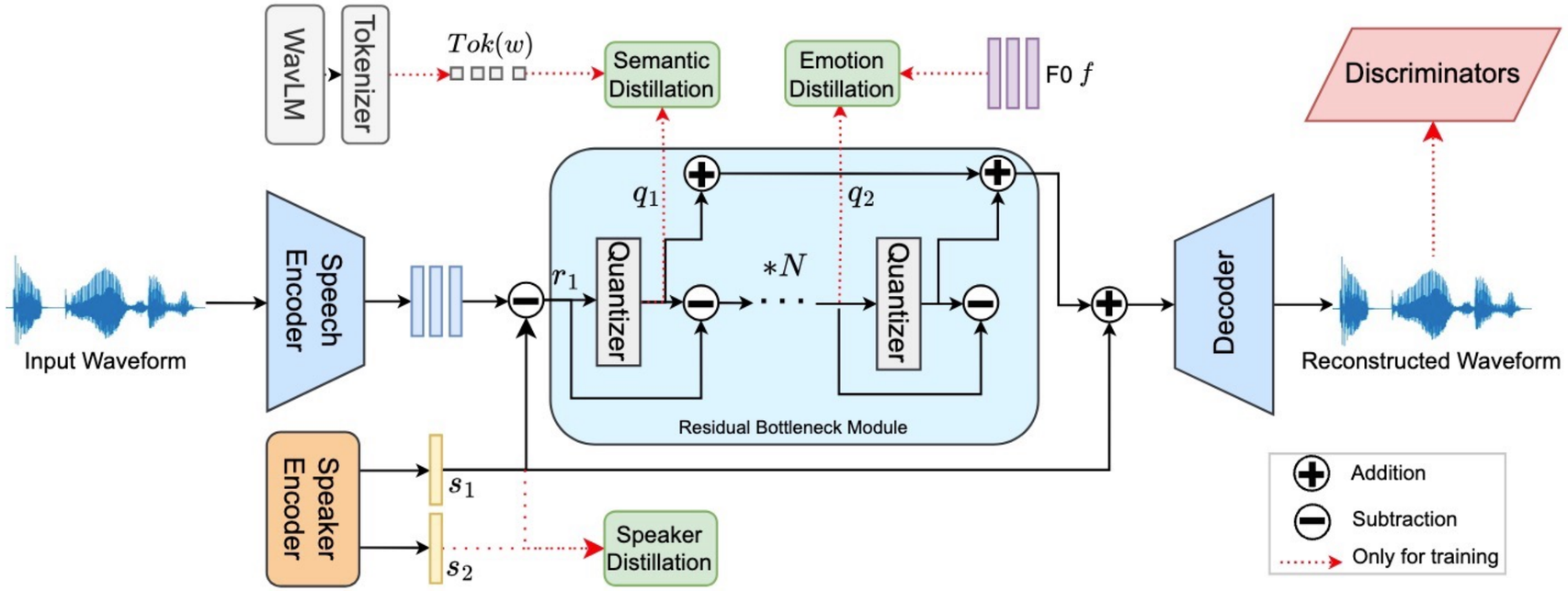
https://www.voiceprivacychallenge.org/

# Our Proposed

- **A speaker anonymization system based on a disentangled neural codec**
  - We propose a serial disentanglement strategy to perform step-by-step disentanglement
    - From a global time-invariant representation (speaker identity)
    - To a temporal time-variant representation (linguistic content and fundamental frequency)
  - We introduce three distillation method to disentangle each speech attribute:
    - Linguistic content, speaker identity and emotion state

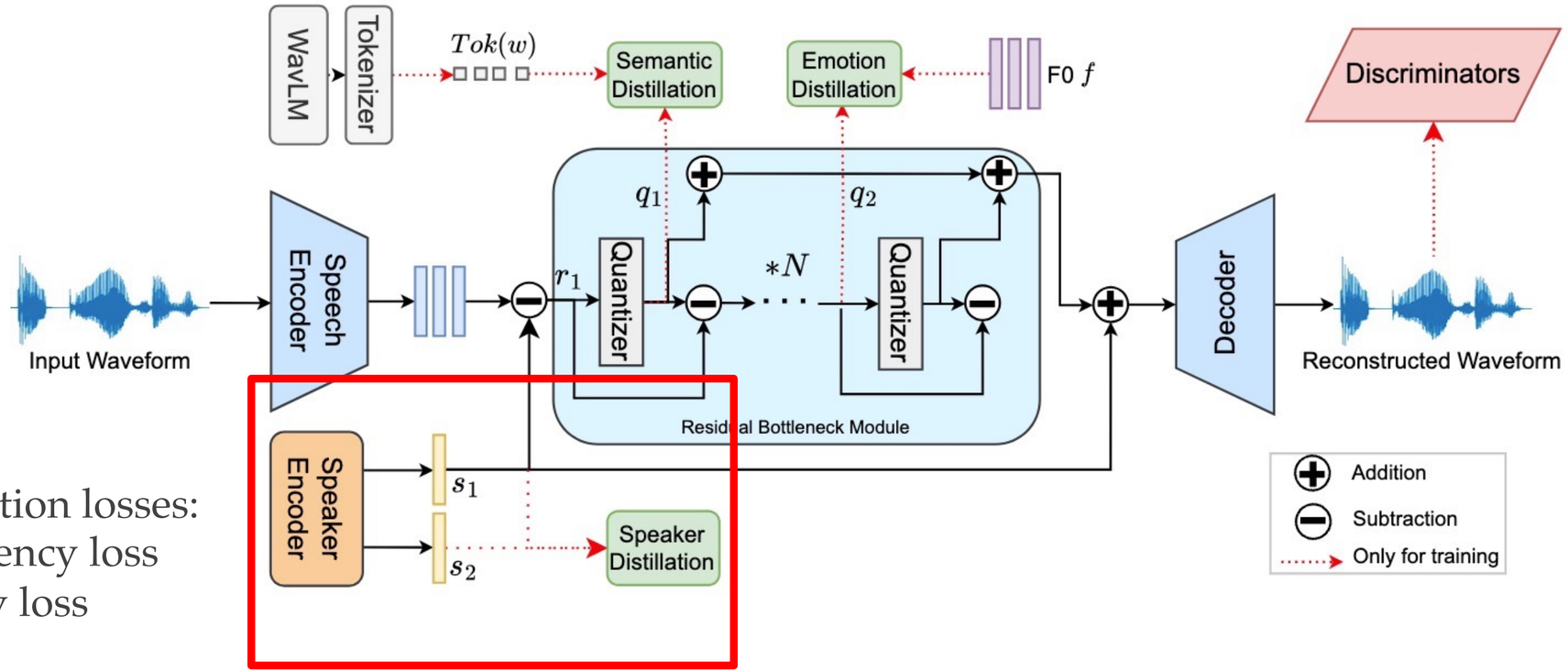
# System Overview

- Auto-encoder architecture consist of:
  - Speech encoder: compress the speech samples into frame-level representations
  - Speaker encoder: extract global speaker representation
  - Residual bottleneck module: disentangle frame-level representation
  - Decoder: reconstruct the input speech waveform



# Factor Distillation

## Speaker distillation




Two speaker distillation losses:

- speaker consistency loss
- speaker identity loss

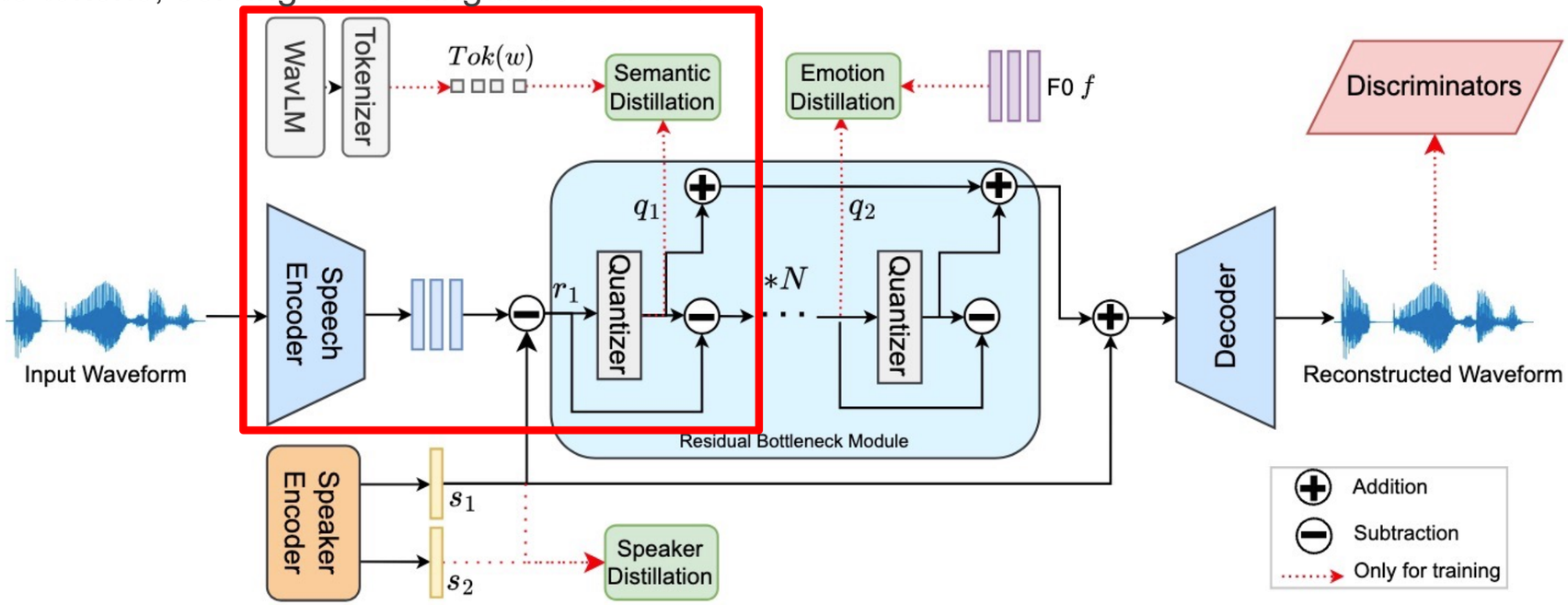
$$\mathcal{L}_{\text{spk}} = \mathbb{E}[-log(C(I \mid s_1))] + \mathbb{E}[-log(C(I \mid s_2))] - \cos(s_1, s_2),$$

# Factor Distillation

## Linguistic distillation

We extract 6th layer's output from pre-trained WavLM model and employ a K-means cluster transfer the representation into discrete tokens, serving as the linguistic teacher

$$\mathcal{L}_{\text{lin}} = \mathbb{E}[-log(\text{Tok}(w) \mid q_1))],$$



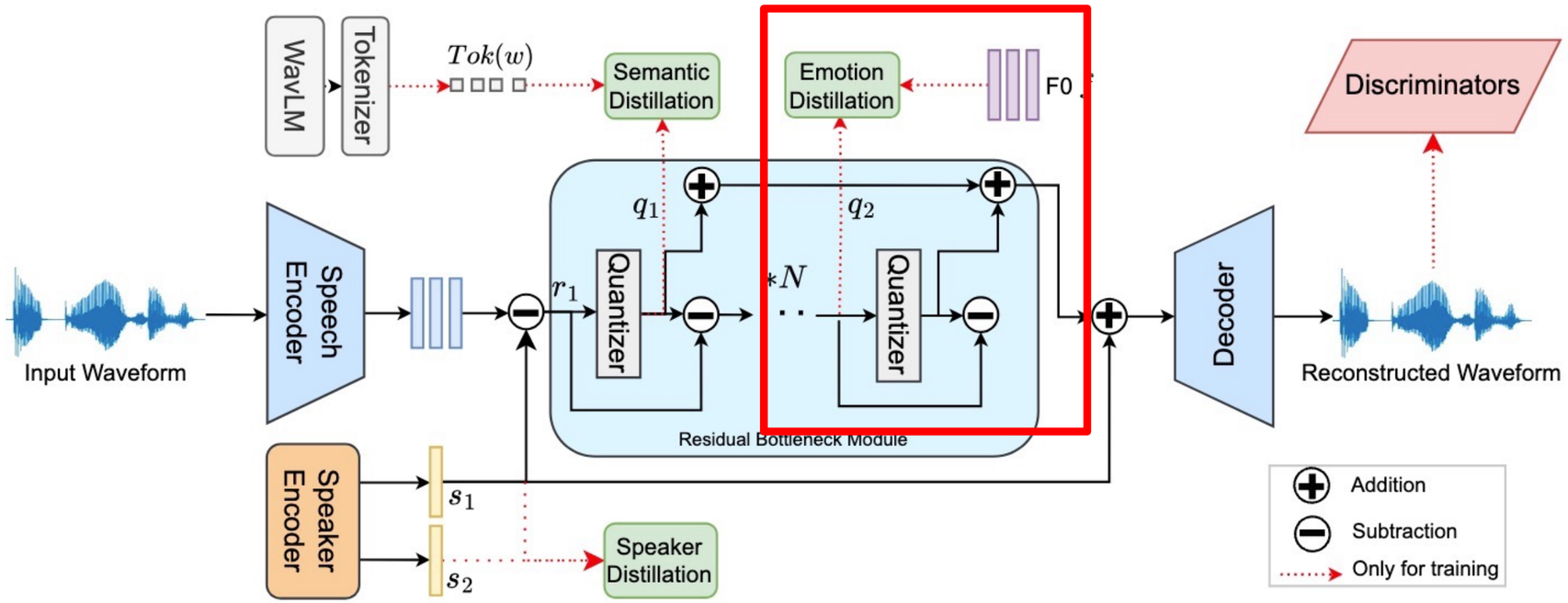

Residual bottleneck module employed for hierarchical disentanglement

# Factor Distillation

## Emotion distillation

We further constrain the residual quantizer with the fundamental frequency (F0)

$$\mathcal{L}_{\text{emo}} = cos(f, \text{Proj}(q_2)),$$



# Training and Inference

- **Training objective**
  - Reconstruction loss: frequency domain, both L1 and L2
  - Adversarial loss: same configuration with HiFi-GAN
  - Distillation loss: speaker, linguistic, emotion
- **Inference**
  - Replace the original speaker identity extracted from the original speech with anonymized identity
  - Anonymized identity: weighted sum the averaged speaker identity and a randomly generated speaker identity

# Experimental Results

- Our proposed system achieves 37.69% and 36.99% on dev and test datasets in condition 3, while the EER results for condition 4 are 42.45% and 40.46%
- Utility metrics **outperform all baseline systems**
  - WER for linguistic preservation
  - UAR for emotion preservation

| | EER, % (↑) | | | | | | WER, % (↓) | | UAR, % (↑) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LibriSpeech-dev | | | LibriSpeech-test | | | LibriSpeech-dev | LibriSpeech-test | IEMOCAP-DEV | IEMOCAP-TEST |
| | F | M | Avg | F | M | Avg | | | | |
| Orig. | 10.51 | 0.93 | 5.72 | 8.76 | 0.42 | 4.59 | 1.80 | 1.85 | 69.08 | 71.06 |
| B1 | 10.94 | 7.45 | 9.20 | 7.47 | 4.68 | 6.07 | 3.07 | 2.91 | 42.71 | 42.78 |
| B2 | 12.91 | 2.05 | 7.48 | 7.48 | 1.56 | 4.52 | 10.44 | 9.95 | 55.61 | 53.49 |
| B3 | 28.43 | 22.04 | 25.24 | 27.92 | 26.72 | 27.32 | 4.29 | 4.35 | 38.09 | 37.57 |
| B4 | 34.37 | 31.06 | 32.71 | 29.37 | 31.16 | 30.26 | 6.15 | 5.90 | 41.97 | 42.78 |
| B5 | 35.82 | 32.92 | 34.37 | 33.95 | 34.73 | 34.34 | 4.73 | 4.37 | 38.08 | 38.17 |
| B6 | 25.14 | 20.96 | 23.05 | 21.15 | 21.14 | 21.14 | 9.69 | 9.09 | 36.39 | 36.13 |
| C3 | 44.18 | 31.20 | 37.69 | 37.96 | 36.03 | 36.99 | 2.56 | 2.66 | 65.98 | 64.48 |
| C4 | 45.31 | 39.60 | 42.45 | 40.66 | 40.26 | 40.46 | 3.51 | 3.19 | 62.93 | 60.87 |

# Conclusion

- We propose a codec based speaker anonymization system with serial disentanglement strategy
- We introduce three distillation method to disentangle the linguistic content, speaker identity and emotion state
- Experiments on VPC official evaluation pipeline demonstrate our proposed speaker anonymization system outperform all baseline systems


Jixun Yao (姚继珣)
Audio, Speech & Language Processing Group (ASLP@NPU)
www.npu-aslp.org
Email: yaojx@mail.nwpu.edu.cn

# Thank You!