Emotional Speech Anonymization: Preserving Emotion Characteristics in Pseudo-speaker Speech Generation

Team Anemone @ VPC 2024

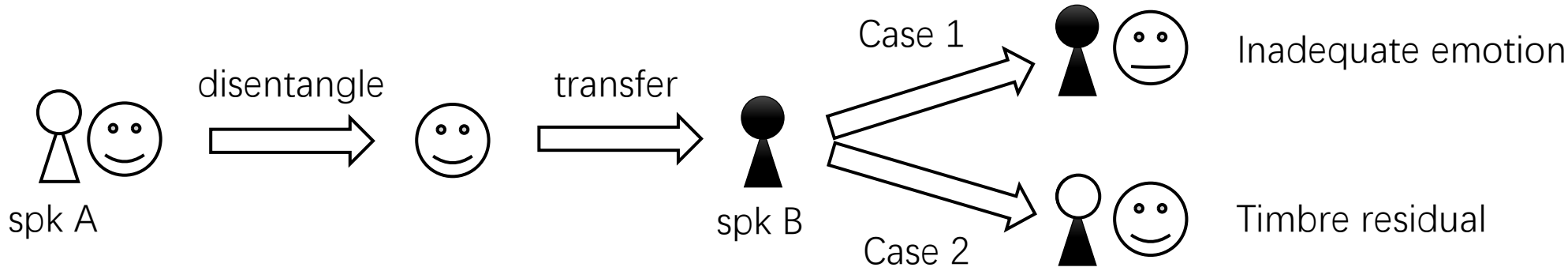
Background and task

- Perform anonymization while retaining the original speaker's style and emotional characteristics

Equal error rate (EER)	Unweighted average recall (UAR)	Word error rate (WER)
Privacy	Emotion	Content

Why is there a trade-off?

- NN methods -- the creation of pseudo-speakers brings sparsity



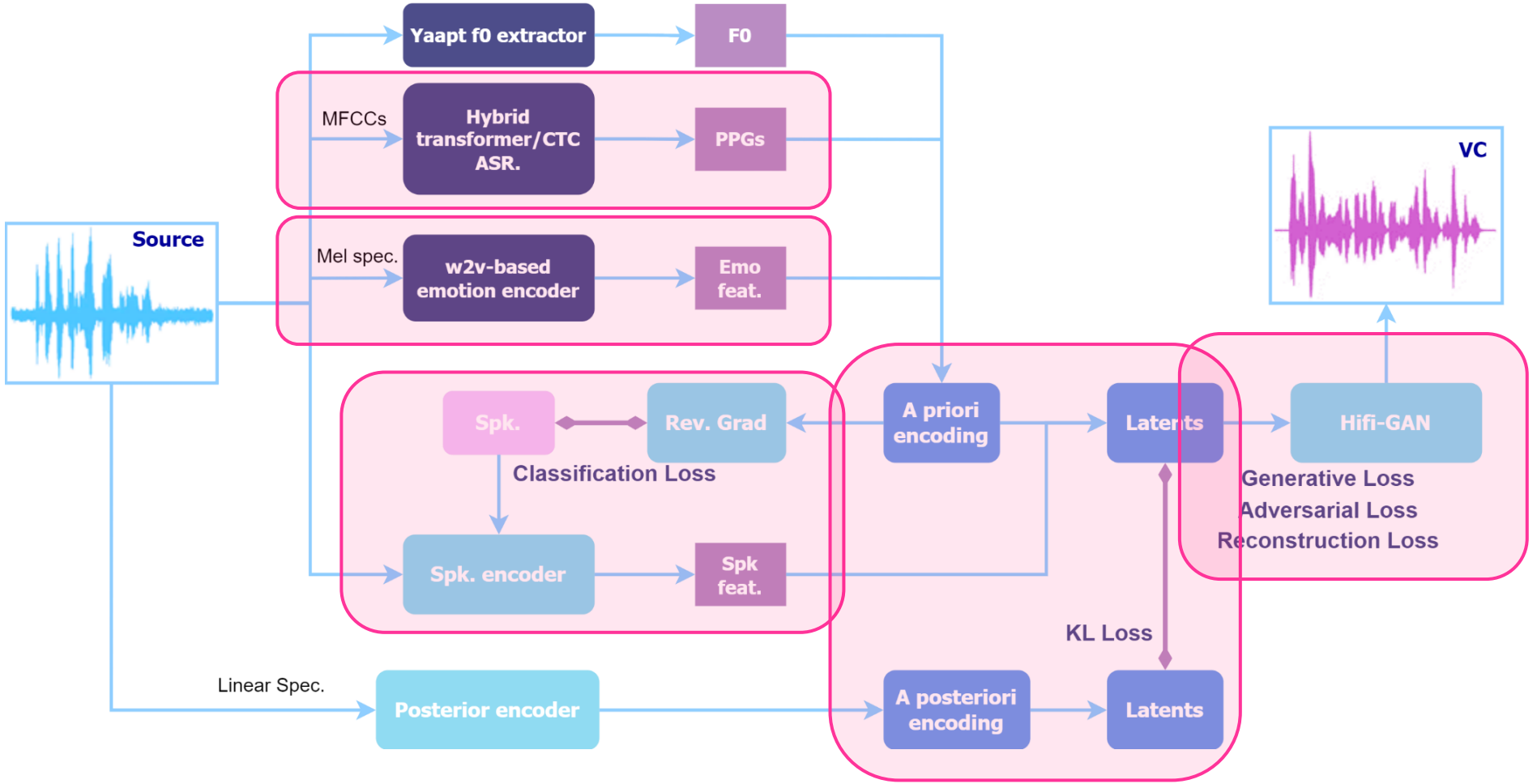
- DSP methods -- non-vocoder

1. Low speech quality
2. High chance of being detected as fake voices

⟹ Not practical

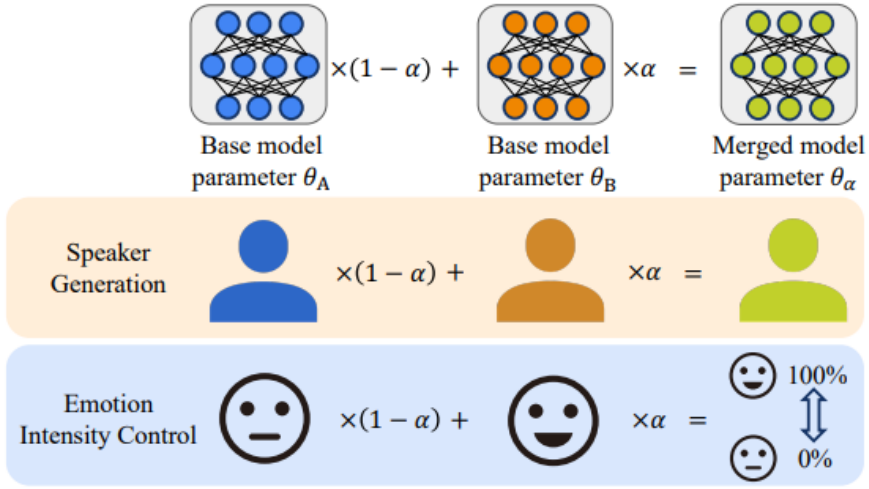
Our methodology

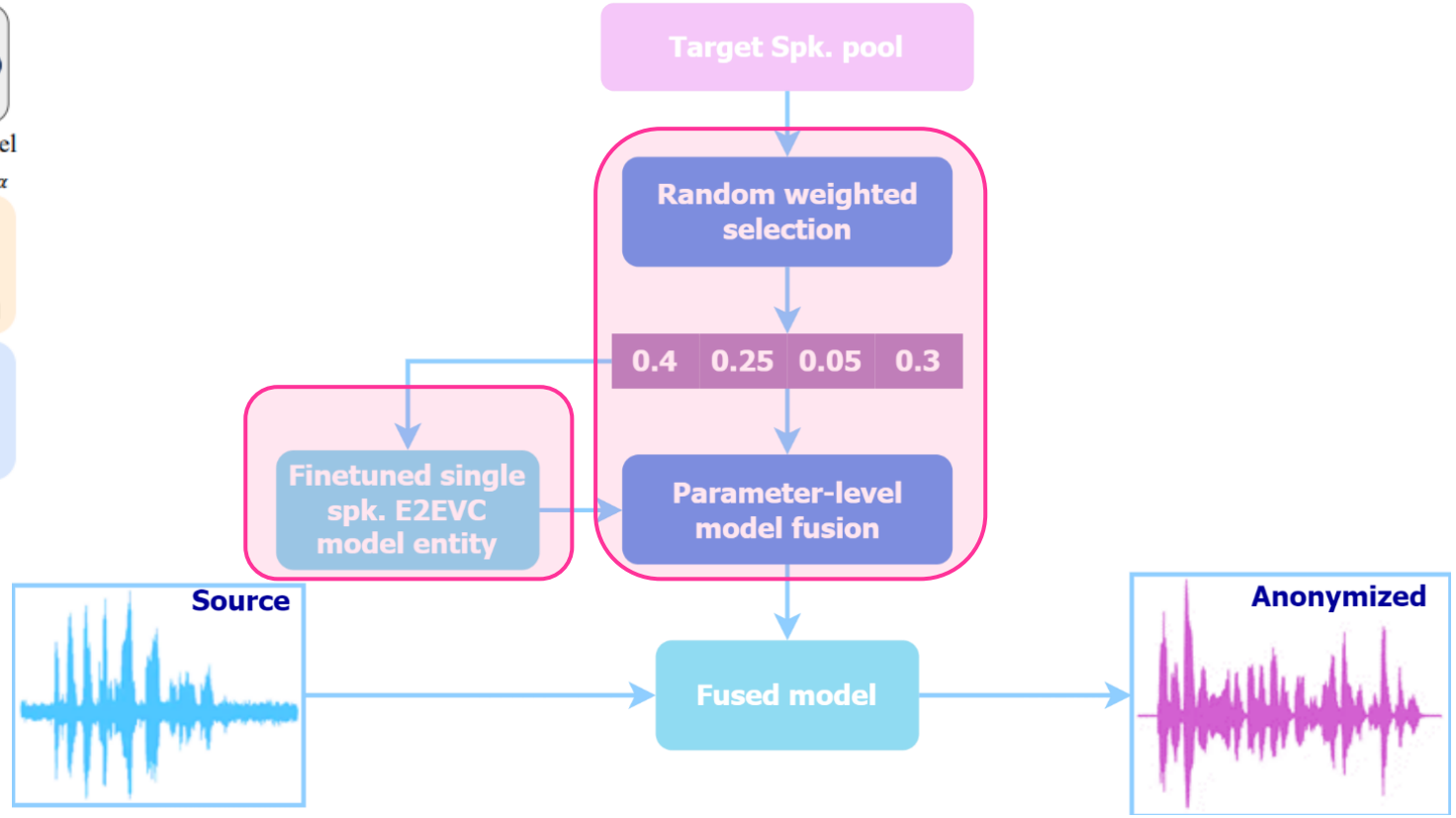
- Backbone: A fully end-to-end VC-model



Our methodology

- Model fusion





Results

- Emotion (UAR%)

System	ave.↑
proposed	**49.96**
Origin	70.07
Given B1	42.74
Given B2	**54.55**
Given B3	37.83
Given B4	42.38
Given B5	38.13
Given B6	36.26

Results

- Content (WER%)

System	ave.↓
proposed	**4.72**
Origin	1.82
Given B1	**2.99**
Given B2	10.20
Given B3	4.32
Given B4	6.02
Given B5	4.55
Given B6	9.39

Results

- Privacy (EER%)

System	ave.↑
proposed	**31.55**
Origin	5.16
Given B1	7.64
Given B2	6.00
Given B3	26.28
Given B4	31.48
Given B5	**34.36**
Given B6	22.10

Limitations

- Model fusion technique lacks interpretability
- Emotion modeling remains inadequate

Thank you for your attention!

About VPC 2024

- Fantastic program!

- Our team focuses on personalized TTS, and mainly studies areas such as emotion synthesis and large speech models. Therefore, we are new to this competition and do not have a thorough understanding of the tasks and goals of anonymization. Emotion generation itself is a difficult topic because it involves the study of internal mechanisms related to speech disentanglement, and there has actually been a lack of very effective means. After the entire competition, We think that the topic of emotion + anonymization is indeed novel and interesting. It does cover a lot of content and there are many things for research and exploration.

- Perhaps you have considered performing both ‘speech deepfake detection’ and ‘speaker privacy protection’ tasks simultaneously? We noticed that some of the baseline synthesized sounds in B1~B6 may sound fake. Even if the performance indicators look pretty good, they have little practical value. Perhaps we can consider adding some detection baselines, and the generated speech is required to have a high chance of being recognized as ‘TRUE voice’. On that basis, we further explore the topic of anonymization.