

SpecWav-Attack: Leveraging Spectrogram Resizing and Wav2Vec 2.0 for Attacking Anonymized Speech

1st Yuqi Li

Dept. of Large Language Model
Qifu Technology
Shanghai, China
liyqi1-jk@360shuke.com

2nd Yuanzhong Zheng

Dept. of Large Language Model
Qifu Technology
Shanghai, China
zhengyuanzhong-jk@360shuke.com

3rd Zhongtian Guo

Dept. of Computer Science
Fudan University
Shanghai, China
guozt24@m.fudan.edu.cn

4th Yaoxuan Wang

Dept. of Large Language Model
Qifu Technology
Shanghai, China
wangyaoxuan-jk@360shuke.com

5th Jianjun Yin*

School Of Information Science And Technology
Fudan University
Shanghai, China
yinjianjun@fudan.edu.cn

6th Haojun Fei*

Dept. of Large Language Model
Qifu Technology
Shanghai, China
feihaojun-jk@360shuke.com

Abstract—This paper presents SpecWav-Attack, an adversarial model for detecting speakers in anonymized speech. It leverages Wav2Vec2 for feature extraction [1] and incorporates spectrogram resizing and incremental training for improved performance. Evaluated on librispeech-dev and librispeech-test, SpecWav-Attack outperforms conventional attacks, revealing vulnerabilities in anonymized speech systems and emphasizing the need for stronger defenses, benchmarked against the ICASSP 2025 Attacker Challenge [2].

Index Terms—SpecWav-Attack, Wav2Vec2, Spectrogram Resizing, Incremental training

I. INTRODUCTION

This paper introduces SpecWav-Attack, a tailored adversarial model for attacking anonymized speech with a focus on Effective Equal Error Rate (EER). Using the ECAPA-TDNN architecture [3], we integrate the Wav2Vec2 self-supervised model [1] to enrich speech representations, enhancing sensitivity to variations in anonymized data.

We apply a spectrogram resizing technique to the train-clean-360 dataset [4], which boosts robustness and generalization. Evaluations on the librispeech-dev [4] and librispeech-test [4] datasets show that SpecWav-Attack outperforms traditional methods in both attack success rate and robustness.

Our results reveal vulnerabilities in speech anonymization and emphasize the need for more effective defenses in voice privacy, suggesting that advancements in robust countermeasures are crucial to mitigate adversarial risks.

II. METHODOLOGY

A. Data Augmentation

We implement the SR-based data augmentation method proposed in Freevc [5], which adjusts the Mel spectrogram vertically during preprocessing. As outlined in Fig. 3, the augmentation process begins with the extraction of the Mel spectrogram x_{mel} from the source waveform y . This is followed by vertical spectral resampling (SR) to produce a modified

Mel spectrogram x'_{mel} , which distorts speaker-specific features while retaining content-related information, thus enhancing model robustness. The process is completed by reconstructing the waveform y' from x'_{mel} , thereby enhancing the model's adaptability to diverse speech inputs.

B. Incremental Training

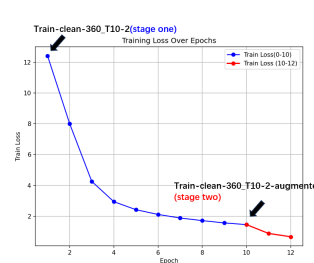


Fig. 1: Training loss changes with epoch

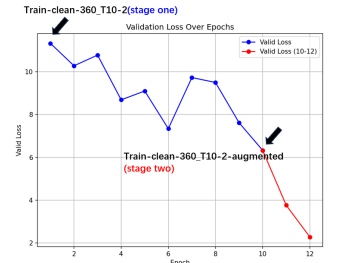


Fig. 2: Validating loss changes with epoch

Initially, a preliminary training phase is conducted for a certain number of epochs (10 epochs on the train-clean-360 dataset). After this, the model is further trained for 2 to 4 epochs on the train-clean-360-augmented dataset, building upon the previously trained model. This approach allows the model to gradually adapt to the data and the task, helping to avoid overfitting at the early stages of training.

C. Feature Extraction

In this work, we replace traditional fbank-based feature extraction with the self-supervised Wav2Vec 2.0 model, which generates 1024-dimensional embeddings from large unlabeled datasets to capture richer and more complex speech patterns. This model encodes both phonetic and speaker-independent features, enhancing performance in scenarios with limited labeled data. The improved feature extraction enhances the

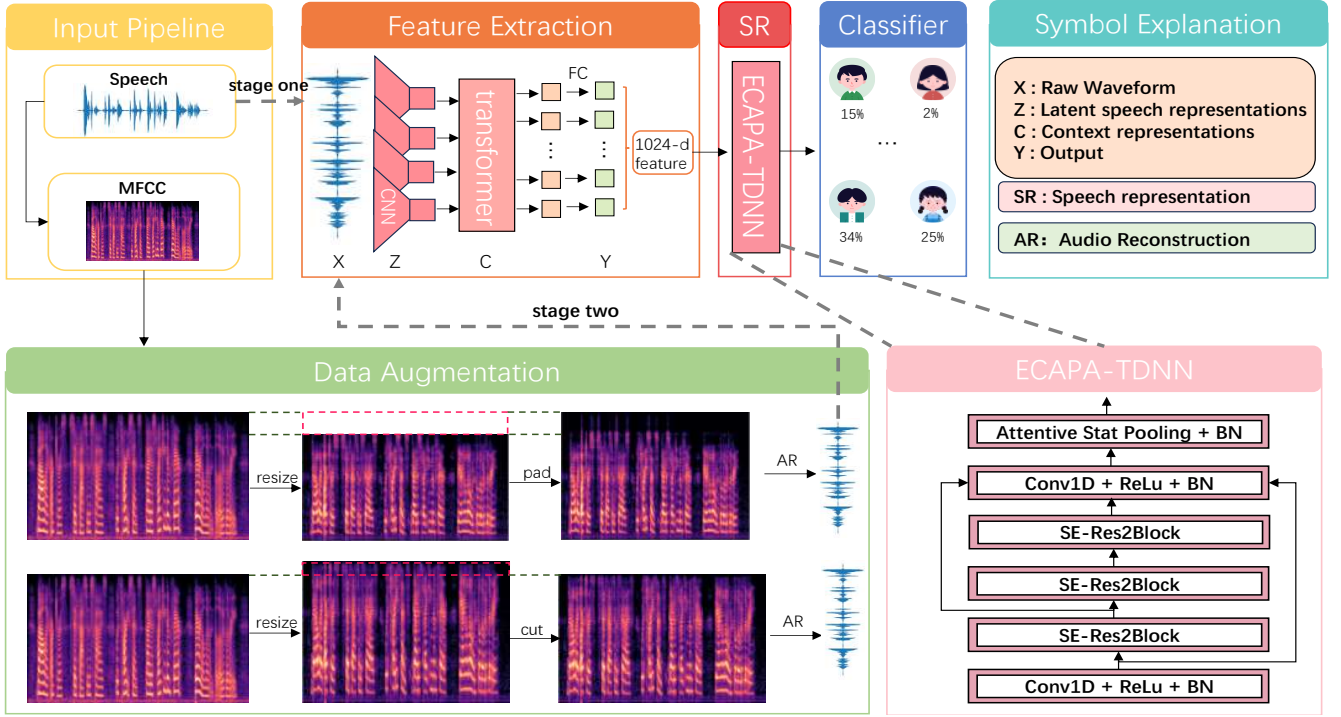


Fig. 3: SpecWav-Attack Architecture: Stage one represents the first 10 epochs of training, and stage two represents 10-12 epochs of training. In this stage, the training set is replaced with the augmented dataset.

Dataset	Gender	EER(%)								
		Orig.	T8-5	T8-5(SW)	T10-2	T10-2(SW)	T12-5	T12-5(SW)	T25-1	T25-1(SW)
LibriSpeech-dev	female	10.51	39.63	28.69	43.63	33.92	43.32	35.09	42.65	35.80
	male	0.93	40.84	32.14	40.04	28.76	44.10	34.80	40.06	37.12
Average dev		5.72	40.24	30.42	41.83	31.34	43.71	34.95	41.36	36.46
LibriSpeech-test	female	8.76	42.50	28.09	41.97	25.91	43.61	34.85	42.34	35.58
	male	0.42	40.05	29.63	29.85	27.17	41.88	34.52	41.92	35.19
Average eval		4.59	41.28	28.86	40.36	26.54	42.75	34.69	41.35	36.39

TABLE I: EER(Equal Error Rate) results for LibriSpeech-dev and LibriSpeech-test datasets,the SW(SpecWav) is the solution we proposed

model's robustness and adaptability, making it ideal for tasks like anonymized speech recognition and other privacy-sensitive applications.

III. RESULTS

Please refer to TABLE I.

IV. CONCLUSION

In summary, the SpecWav model significantly improves the Equal Error Rate (EER) over all baseline methods (T8-5, T10-2, T12-5, T25-1) on both the LibriSpeech-dev and LibriSpeech-test datasets. The improvements are particularly noticeable for T10-2, where a 13.82% reduction in EER is observed. These results demonstrate the effectiveness of SpecWav in enhancing the performance of voice-based tasks by reducing EER across different configurations.

REFERENCES

- [1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.
- [2] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The first VoicePrivacy attacker challenge evaluation plan," *arXiv preprint arXiv:2410.07428*, 2024.
- [3] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.
- [4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.
- [5] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.