

Appendix

A. Detection of VoiceNoNG-Edited Speech

As demonstrated in the previous subjective evaluation, people find it challenging to distinguish between speech edited by the proposed VoiceNoNG and real speech. To prevent malicious uses, a reliable Deepfake detector is essential. This section explores a practical new threat within the partial spoof scenario. Given the potential for neural codecs to become a new audio format standard, the assumption that all codec-generated speech is fake may soon be unrealistic. Therefore, in the following experiments, we classified codec-resynthesized speech as a new category. The original binary classifier was extended to a ternary classifier with the labels: **real**, **resynthesized**, and **edited**. Additionally, for the audio condition, besides the original VoiceNoNG setting where non-edited segments come from the original audio, we consider a more challenging setting where non-edited segments are also resynthesized from the codec. We refer to this condition as VoiceNoNG (resyn).

A.1. Deepfake speech detector model setup

We built the w2v2-detector using a model architecture similar to the one that won first place in the partially fake audio track at ADD2022 [1], which also demonstrated strong performance in our previous work. Specifically, the w2v2-detector employs a pretrained wav2vec2-base-960h model [2] for feature extraction. We use a weighted sum of the 13 hidden states, including the output from the CNN encoder, as our input features. These features are then projected from 768 dimensions to 128 dimensions using a linear layer. The model then splits into two paths: the **frame-level** branch and the **utterance-level** branch. In the frame-level branch, the projected features pass through a linear layer to produce a 3-dimensional output representing the prediction for each frame. In the utterance-level branch, the projected features undergo attentive statistics pooling [3] to compress the features into a single representation of the entire utterance, which is then passed through a linear layer to generate the utterance-level prediction.

A.2. Training and test sets for Deepfake detector

We used LibriLight medium [4] as the source audio files and generated the edited speech using our proposed VoiceNoNG. The target transcripts were generated by prompting an LLM (i.e., zephyr-7B-beta [5]). However, the edited transcripts may not always adhere to the desired format. To address this, we used word-level Levenshtein distance [6] to identify substitutions, and then randomly selected words for replacement as our method of edit manipulation.

The dataset was first categorized by speaker, prioritizing those with more audio samples for the training set, and then sequentially allocating speakers to the validation and test sets. This approach ensured that no speaker appeared in multiple sets, with the test set containing a significant number of unseen speakers, enabling us to effectively assess the detector’s generalizability at the speaker level. The dataset was divided into training, validation, and test sets, comprising 106,186, 34,744, and 33,974 audio files, respectively. Additionally, the quantities of real, resynthesized, and edited speech in the original VoiceNoNG and VoiceNoNG (resyn) settings were balanced within each set.

Table 1: *Detection and localization results of speech edited by VoiceNoNG.*

# of training sample	frame F1 (%) / utterance accuracy (%)	
	VoiceNoNG	VoiceNoNG (resyn)
106,186	95.84 / 98.08	82.63 / 96.99
80,000	93.59 / 97.17	83.14 / 97.14
40,000	94.44 / 97.17	81.28 / 96.66
20,000	91.25 / 94.85	80.89 / 92.83
5,000	89.39 / 91.32	76.45 / 88.01
1,000	65.97 / 79.07	56.86 / 64.61
500	58.83 / 72.79	47.48 / 66.19

A.3. Experimental results

Table 1 presents the performance of the w2v2-detector in detecting and localizing speech edits made by the proposed VoiceNoNG model. The table reveals that the utterance-level accuracy is nearly 100%, even under the VoiceNoNG (resyn) condition. However, as anticipated, the frame-level F1 score for the VoiceNoNG (resyn) condition is lower than that for VoiceNoNG. This indicates that it is more challenging for the detector to differentiate between **edited** and **non-edited resynthesized** segments at the frame level.

To examine how the amount of training data affects the detector’s performance, we included results for different training set sizes in the table. The table indicates that approximately 40,000 samples are sufficient to train a reliable DeepFake detector. Additionally, thanks to the advantages provided by the self-supervised front-end (i.e., wav2vec2), only 5,000 examples are needed to achieve an utterance-level accuracy of around 90%. Figure 1 presents an example of frame-level detection results for the VoiceNoNG (resyn) condition. Additional examples under different acoustic conditions can be found in Figure 2 to 4. Through this experiment, we discovered that while it is challenging for people to differentiate between speech edited by the proposed VoiceNoNG and real speech, the trained w2v2-detector is still capable of detecting subtle artifacts that distinguish real audio from fake.

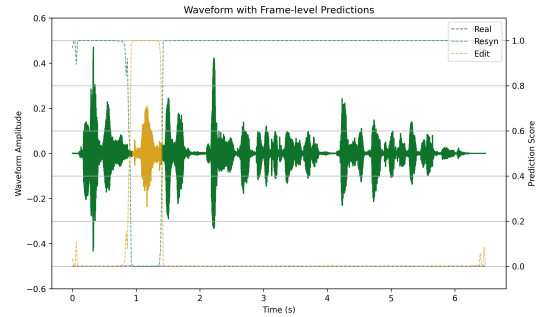


Figure 1: *An example of frame-level detection for VoiceNoNG (resyn) condition. In the waveform, the green sections represent resynthesized speech, while the orange sections indicate edited speech. The dashed lines show the predicted scores of our detector for the three classes.*

B. Subjective Evaluation Details

Instructions

Some of the speeches you will listen to may have been **partially edited**. Your task is to assess the naturalness of the speech **focusing** solely on the **speaker and background audio coherence, prosody, emotion, and speech rate**. Some of the audio may come from internet videos and have background noise. Please **ignore** the noise, grammar, semantics, or other linguistic factors in your evaluation.

Please rate each audio's naturalness (i.e., human-sounding) independently from 1-5. 1 is **least** natural, and 5 is **most** natural.

Please use a headset to listen and adjust the volume level to your comfort. Each audio should only be replayed at most **twice**.



Figure 5: Instruction of the subjective evaluation.

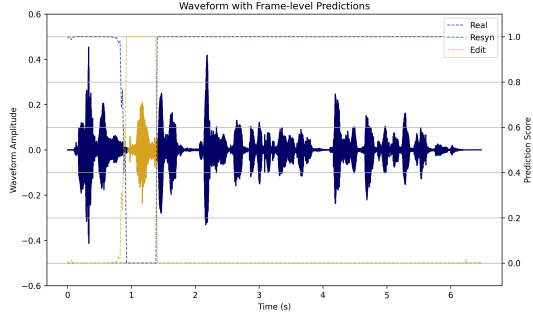


Figure 2: An example of frame-level detection for **VoiceNoNG** condition. In the waveform, the blue sections represent real speech, while the orange sections indicate edited speech.

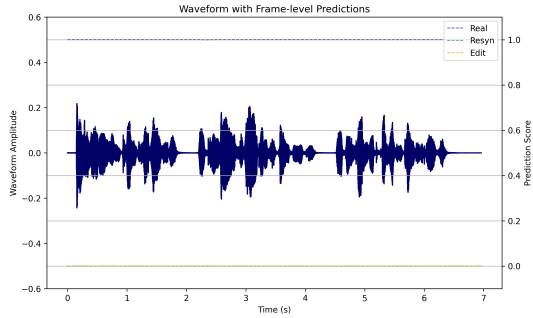


Figure 3: An example of frame-level detection for **real** condition.

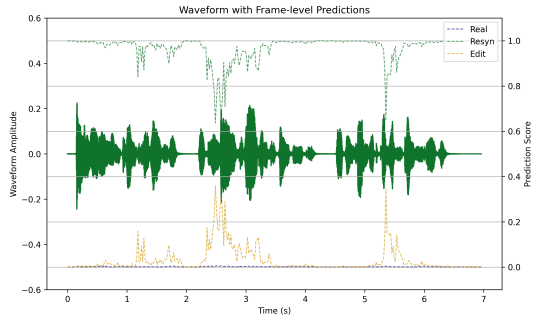


Figure 4: An example of frame-level detection for **resynthesized** condition.

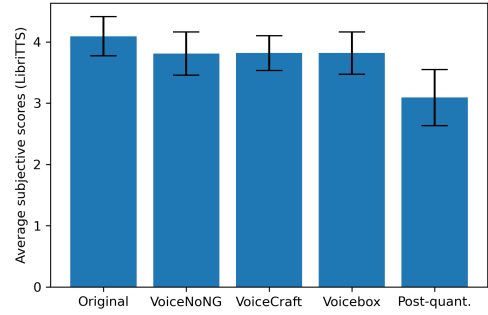


Figure 6: Subjective scores of different speech editing methods in the LibriTTS subset.

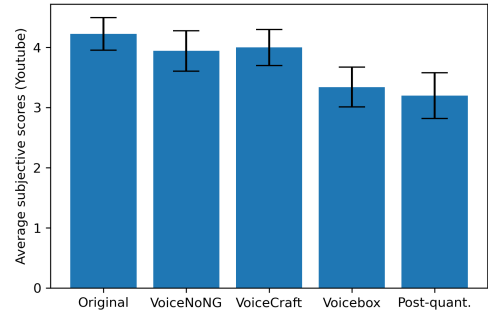


Figure 7: Subjective scores of different speech editing methods in the Youtube subset.

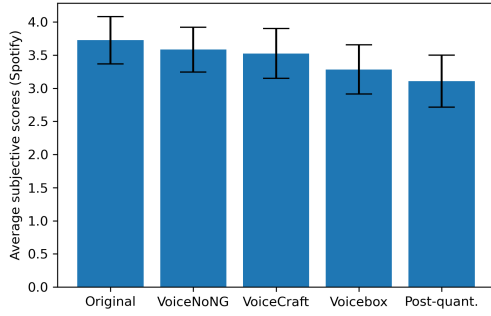


Figure 8: Subjective scores of different speech editing methods in the Spotify subset.

C. Hallucination-like Problem Example

Table 2: An example of VoiceCraft’s attention errors (hallucinations) in the RealEdit dataset. Words highlighted in red indicate missing or deleted words, while those in blue denote extra or inserted words. Parentheses signify simple substitution errors. Bolded texts indicate hallucinations (missing or repeated words).

Ground truth:
”yet anytime you and i question the schemes of the dogooders or dare to dig into any of their motives were denounced as being against their humanitarian goals they say we are always against things we are never for anything”
VoiceCraft:
”yet anytime you and i question the schemes of the (-dogooders +dog) or dare to dig into any of their motives -were +we are denounced as gooders we are denounced as being against their humanitarian goals they say we are always against things we are never for anything”
Proposed VoiceNoNG:
”yet anytime you and i question the schemes of the (-dogooders +dog eaters) or dare to dig into any of their (-motives were +we are) denounced as being against their humanitarian goals they say we are always against things (-we are +were) never for anything +and”

D. References

- [1] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, “Add 2022: the first audio deep synthesis detection challenge,” in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.
- [2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *CoRR*, vol. abs/2006.11477, 2020. [Online]. Available: <https://arxiv.org/abs/2006.11477>
- [3] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” in *Interspeech*, 2018, pp. 2252–2256.
- [4] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, “Libri-light: A benchmark for asr with limited or no supervision,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.
- [5] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib *et al.*, “Zephyr: Direct distillation of LM alignment,” The H4 Team at Hugging Face, Tech. Rep., October 2023.
- [6] V. I. Levenshtein *et al.*, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.