Making Partially Fake Audio

CNSL 이요원

1 Introduction

As machine learning continues to advance, deepfake in audio are now becoming a social concern. Therefore, it is necessary to have a technology that can distinguish whether it is a deepfake or a real voice. Since partially fake audio can change the meaning entirely by changing or inserting specific words in a sentence, partially fake audio should also be able to be detected in deepfake detection. In this document, we will discuss how to construct the partially fake audio dataset.

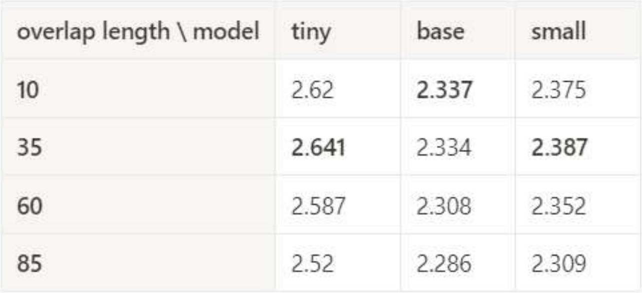
2 How to make partially fake audio

2.1 Tools

To construct partially fake audio, both real and fake datasets are necessary. In this study, we utilized 500 samples from the Korean Single Speaker(KSS) datsaset as the real dataset. Additionally, for the fake dataset, we generated and employed 60 sentences using a pre-trained Text-to-Speech(TTS) model.

Next, 'whisper-timestamp' was used to determine the timestamp intervals between words, enabling us to segment the sentence samples into individual words. Utilizing these timestamps, we employed the 'pydub' library to cut out and concatenate segments with overlaps. To assess the quality of the generated partially fake audio, we utilized the 'Mean Opinion Score(MOS)' as a metric.

2.2 Compare whisper model



overlap length \ model	tiny	base	small
10	2.62	**2.337**	2.375
35	**2.641**	2.334	**2.387**
60	2.587	2.308	2.352
85	2.52	2.286	2.309

Figure 1. The table above shows the MOS score according to the overlap length of each whisper model with 3000 samples.

We conducted tests to determine the average MOS score after generating 3000 partially fake samples using various overlap lengths with three whisper models: tiny, base and small. The summarized test results are presented in *Figure 1*. Notably, the tiny model exhibited the highest performance, with optimal results achieved when the overlap length was set to 35ms. Consequently, we concluded to construct the partially fake audio dataset employing the tiny model with a 35ms overlap length.

2.3 Maintain sentence structure

In an effort to enhance audio quality, we focused on preserving sentence structure. We applied a simpler method. In Korean, sentences typically commence with a subject or adverb and conclude with a concluding ending. Therefore, maintaining this structural convention is anticipated to yield a more natural dataset.

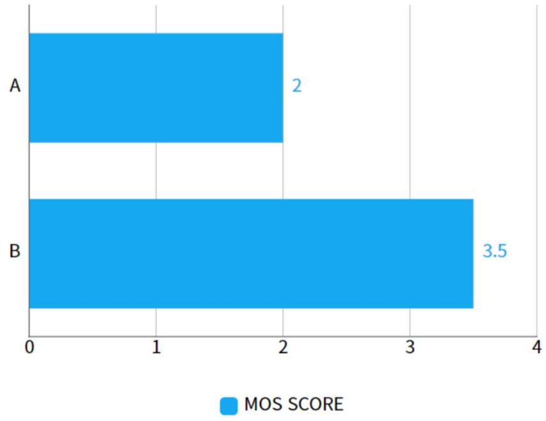


Figure 2. A is an audio sample composed of randomness, B is composed while maintaining structure.

We requested native Korean speakers to assess the Mean Opinion Score(MOS) to compare two sets of samples: one maintaining the sentence structure and the other not. Indeed, it is evident from *Figure2* that the sample preserving the sentence structure scored higher.

To generate partially fake audio while preserving a minimal sentence structure, the first and last words are stored separately when segmenting the real and fake dataset. For the partially fake audio, the first and last words are positioned by randomly selecting between real and fake with a 50% probability, while real and fake words are interspersed alternately within the middle section.

3 Conclusion and Future Work

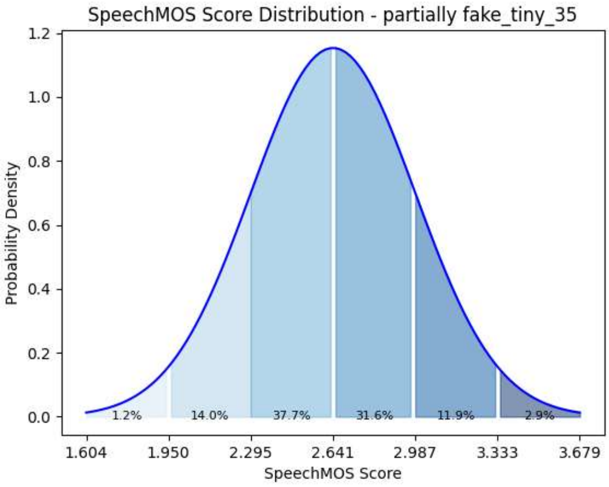


Figure 3. MOS score distribution of 3000 partially fake samples

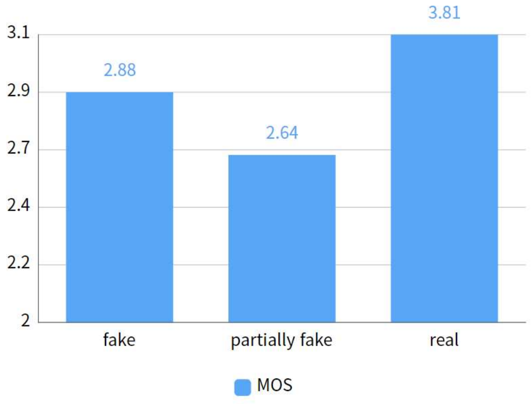


Figure 4. Average MOS of fake, partially fake and real dataset

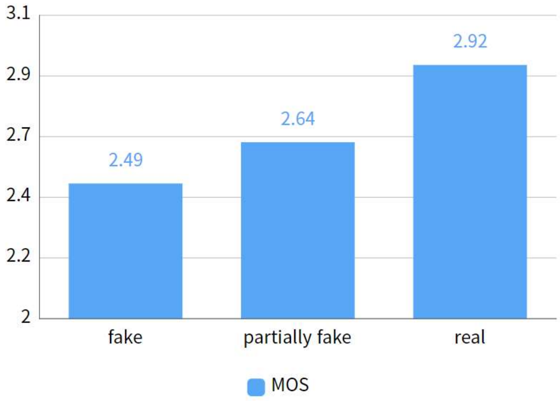


Figure 5. Average MOS for 3000 samples of fake, partially fake, real dataset connected with overlap

We generated 3000 partially fake samples without considering the meaning of sentences. The varying lengths of the words composing these sentences will aid in training a deepfake detection model capable of identifying partially fake of different lengths. The MOS scores distribution of 3000 samples of partially fake audio, generated as described above, is depicted in *Figure 3*. Initially, it was anticipated that the MOS score would fall approximately midway between those of the real dataset and the fake dataset. However, the actual results, as illustrated in *Figure 4*, did not align with this expectation. The anticipated issue lies in the connection problem of word segments. Although efforts were made to enhance naturalness through overlap, limitations persist compared to actual speech. As evidence, constructing a sentence using only the fake and real samples in the same manner as the partially fake audio in *Figure 5* reveals that the MOS score of the partially fake audio falls within the expected range between the two extremes. Currently, there is a necessity to address the issues concerning partially fake audio to create a more natural-sounding output and advance future work aimed at improving the MOS score.