Weekly Work Report

May 20 - May 31

Arth J. Shah

arth.j@parther.samsung.in
arth123shah@gmail.com

May 31, 2024

1 Research problems

I explored and examined the theoretical as well as existing potentials available in Voice Liveness Detection, and Signfake tasks.

2 Research Approach

Now-a-days, use of pre-trained models formed from huge data, and with billions of parameters, have been boosted abundantly. Their potential for one or another task keep on growing due to its variety of factors and aspects to learn properties of an audio file. More deeply created and more fine-turned models can perform better on variety of tasks, due to it's capability to capture more detailed information about particular audio sample.

3 Progress in this week

Have gone through few previously accepted papers in ICASSP, and INTERSPEECH, on topic of deepfake, and VLD.

Learned basics of a few pre-trained models, such as

WHISPER [1].

WAV2VEC 2.0 [2]

XLSR (advanced w2v2) [3]

HUBERT [4]

RAWNET (partially) [5]

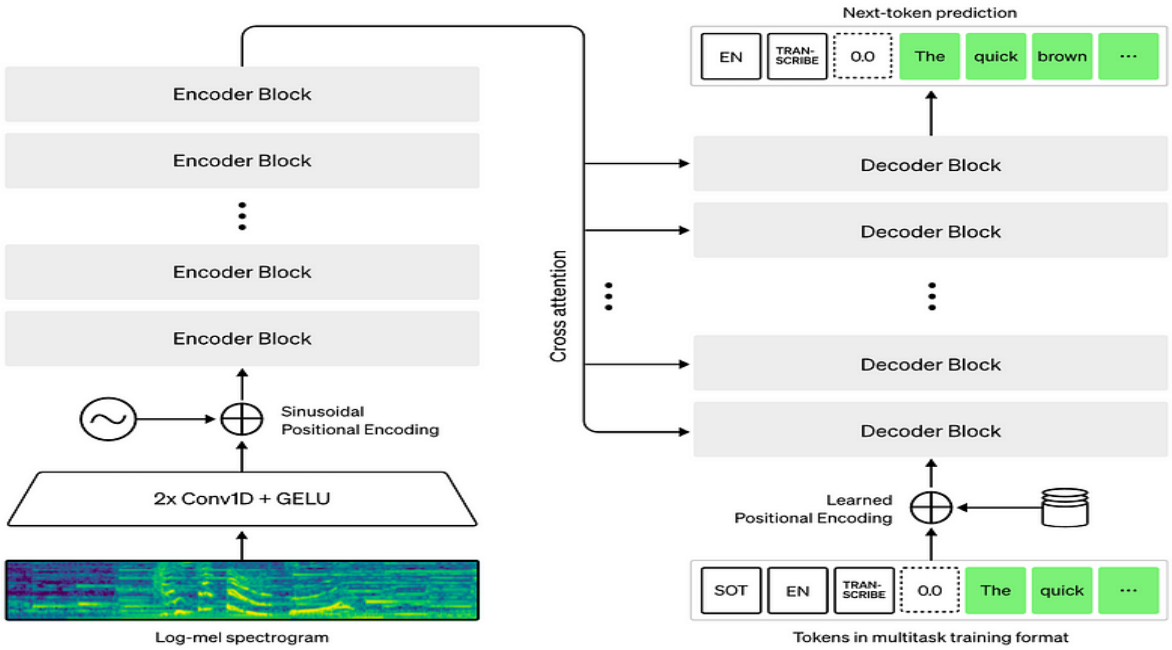

Figure 1: WHISPER feature extraction process.

Fig. 1 displays the feature extraction process through WHISPER modelon basis of encoder and decoder blocks.

4 Future Plan

Objective: To get on one of two topic.
Deadline: End of the fortnight.

2024.06.01 — 2024.06.05 : find an relevant approach.

2024.06.05 — 2024.06.10 : Learn selected approach deeply.

2024.06.10 — 2024.06.15 : Code and try implementing the approach.

2024.06.15 : Representation and discussion of the obtained results.

References

[1] H. Ma, Z. Peng, M. Shao, J. Li, and J. Liu, "Extending whisper with prompt tuning to target-speaker asr," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 516–12 520.

[2] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhang, "A study on fine-tuning wav2vec2. 0 model for the task of mispronunciation detection and diagnosis." in *Interspeech*, 2021, pp. 4448–4452.

[4] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?" in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6533–6537.

[5] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.