WAVEFORM RE-SYNTHESIS FROM AV-HUBERT REPRESENTATIONS

Ju-Chieh Chou, Karen Livesu

Toyota Technological Institute at Chicago

1. PROPOSED APPROACH

We proposed a re-synthesis framework for audio-visual speech enhancement. As shown in Fig. 1, we use a pre-trained AV-HuBERT [1] as the backend model for speech enhancement. A WaveGrad-like [2] model is then trained to generate waveform conditioning on the final layer of AV-HuBERT representations. To make sure the generated output is clean speech, we filter the dataset (LRS3+ VoxCeleb2 [3, 4]) using a neural quality estimator (NQE) [5]. We filter the utterances using SI-SDR [6] predicted by the NQE, which results in about 305 hours of audio-visual speech.

With this vocoding task alone, the model can perform speech enhancement already. To further improve the result, we fine-tune the model on AVSE dataset. The condition vectors are generated from noisy utterances, and the WaveGrad model is trained to generate clean waveform.

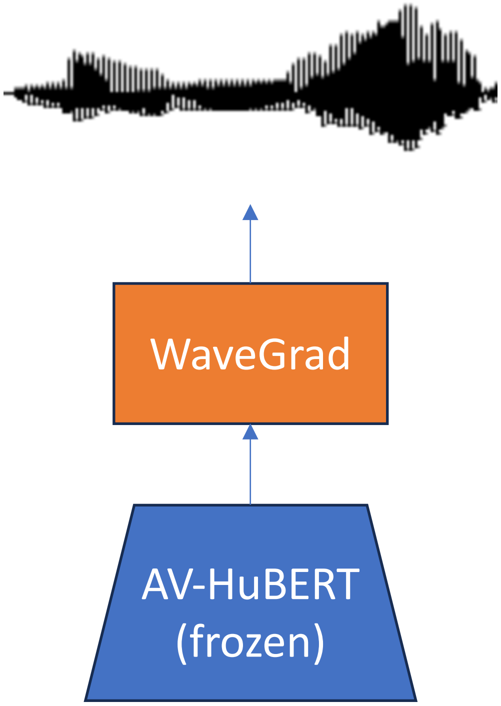

Fig. 1. Proposed framework.

Model	WER	MOS
Pre-trained	21.39	4.13
Pre-trained + FT on AVSE	20.26	4.16

Table 1. Results on AVSE dev set.

2. RESULTS

As our approach is based on generative models, objective evaluation on signal level cannot reasonably reflect the quality. We thus use Whisper [7] small on the AVSE dev set to select our model. We also use MOS score predicted from the neural quality estimator [5] as a reference. The results can be found in Table 1.

3. REFERENCES

[1] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[2] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[4] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[5] Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?,"

in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.