LA-VocE: Low-SNR Audio-visual Speech Enhancement using Neural Vocoders

Rodrigo Mira\textsuperscript{1} Buye Xu\textsuperscript{2} Jacob Donley\textsuperscript{2} Anurag Kumar\textsuperscript{2} Stavros Petridis\textsuperscript{1,3} Vamsi Krishna Ithapu\textsuperscript{2} Maja Pantic\textsuperscript{1,3}

\textsuperscript{1}iBUG, Imperial College London, UK \textsuperscript{2}Meta Reality Labs Research, USA \textsuperscript{3}Meta, UK

### Motivation

- Audio-visual speech enhancement (AVSE) aims to enhance audio by leveraging the speaker’s lip movements.
- Can be trained on raw unlabeled audio-visual data by combining clean speech with noise on the fly.
- Has many applications, including video conferencing and hearing augmentation in noisy/crowded environments.

### Previous Approaches

- Typically combine lipreading backbones with existing audio-only speech enhancement models.
- Often rely on Griffin-Lim or re-use the noisy phase.

### Our Method - LA-VocE

**Training - Stage 1**

- Lipreading backbone
  - Noisy Audio
  - Video
  - AVSE Model

**Training - Stage 2**

- Video Encoder
  - Video
  - Audio Encoder
  - Audio Decoder
  - Griffin-Lim / Noisy Phase / Direct Pred.

**Inference**

- Enhanced mel-spec.
- Discriminators (multi-period + multi-scale)

### Noise and Interference Study

- We train a transformer-based spectrogram enhancer inspired by recent audio-visual speech recognition models.
- Then, we train a neural vocoder (HiFi-GAN) on the same corpus to generate raw audio from spectrograms.
- Finally, we combine both during inference to perform end-to-end enhancement.
- We train our model by combining clean speech from AVSpeech (~4,700 hours, 11+ languages) with noise from the DNS Challenge noise dataset (~70,000 samples, ~150 classes).

### Comparison with Other Works

- Our HiFi-GAN (trained on AVSpeech) outperforms pre-trained models, Griffin-Lim and noisy phase reconstruction.

### Spec. Inversion Comparison

- Our HiFi-GAN (trained on AVSpeech) outperforms pre-trained models, Griffin-Lim and noisy phase reconstruction.