Imperial College London

ICASSP 2023
4 - 10 JUNE,
RHODES ISLAND, GREECE

Meta

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

Rodrigo Mira [1] Buye Xu [2] Jacob Donley [2] Anurag Kumar [2] Stavros Petridis [1,3]
Vamsi Krishna Ithapu [1] Maja Pantic [1,3]

[1] Imperial College London
[2] Meta Reality Labs Research
[3] Meta

**Check out our project page here (with demos!)**

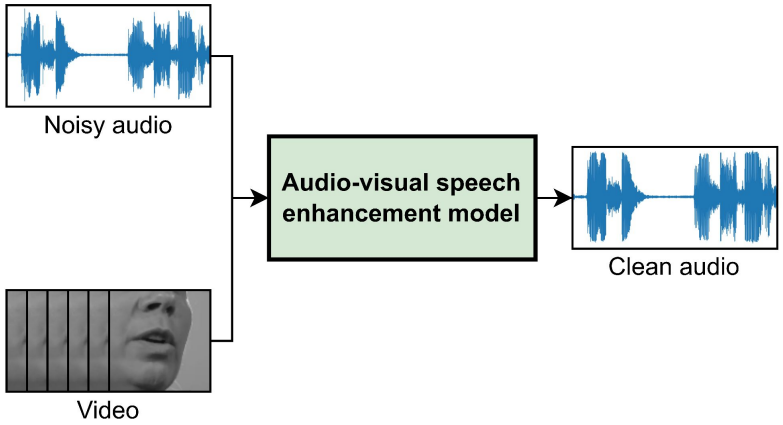
□□□

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

Imperial College London

Meta

# What is Audio-Visual Speech Enhancement?

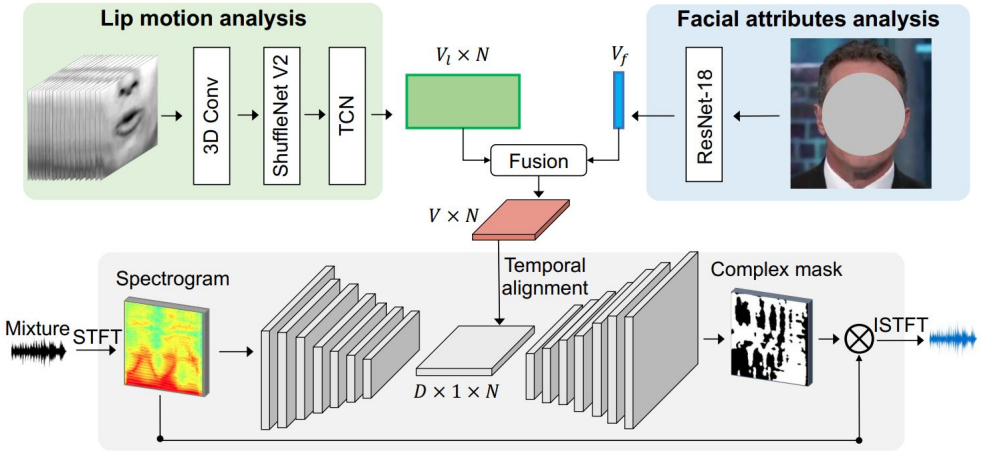



- Audio-visual speech enhancement (AVSE) aims to enhance audio by leveraging the speaker's lip movements.
- Can be trained on raw audio-visual data by combining clean speech with noise on the fly.
- Outperforms audio-only speech enhancement in 2 main scenarios:
  - Low signal-to-noise ratio (SNR)
  - Interfering speech
- Many applications (e.g. videoconferencing)

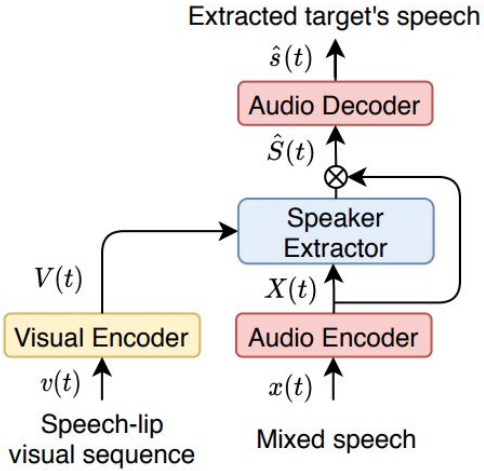
Imperial College London

Meta

# Previous AVSE Approaches




*Gao et al.“VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency” CVPR 2021*

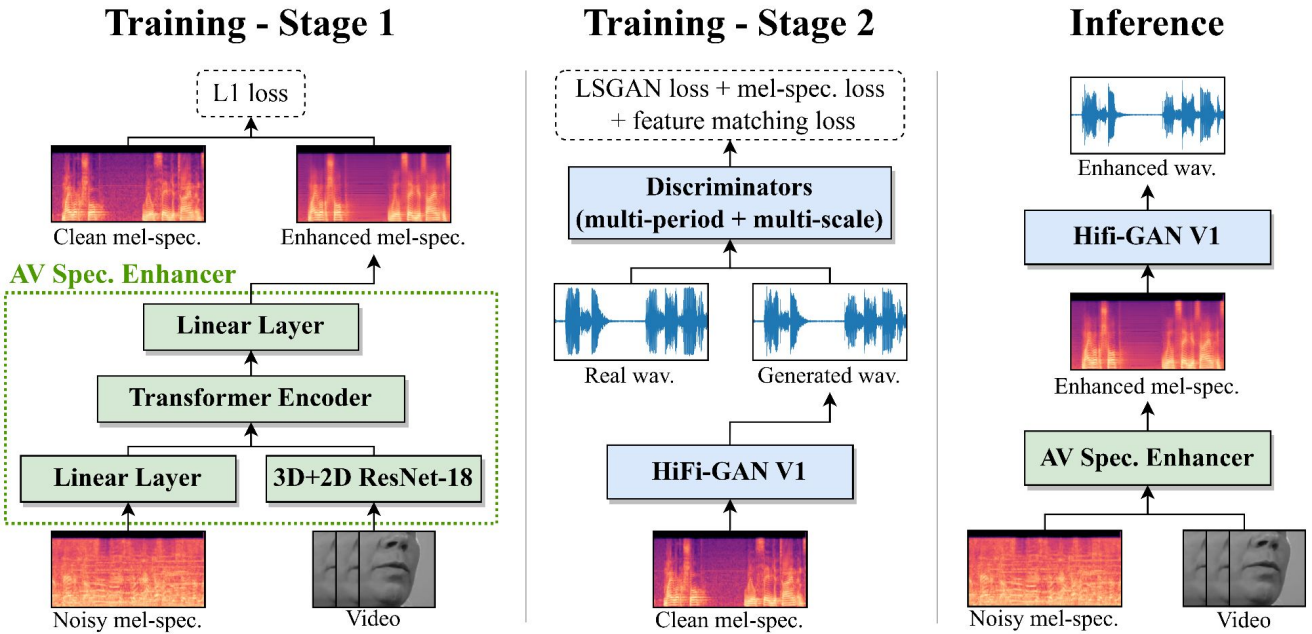



*Pan et al.“Muse: Multi-modal target speaker extraction with visual cues” ICASSP 2021*

Imperial College London

Meta

# Our Method - LA-VocE




Rodrigo Mira et al.

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

Imperial College London

Meta

# Datasets

## AVSpeech

- **Clean audio + video**
- Extracted from Youtube Videos
- **11+** Languages
- Around **4700** hours
- **>150,000** Speakers
- **>2,000,000** clips in total

+

## DNS Challenge Noise

- **Audio only**
- Collected from AudioSet and Freesound
- **~70,000** clips
- **>150** noise classes (e.g., cars, machines , music, etc.)

Imperial College London

Meta

# Results

| Method | Input | MCD ↓ | PESQ-WB ↑ | ViSQOL ↑ | STOI ↑ | ESTOI ↑ |
|---|---|---|---|---|---|---|
| **Noise condition 1 (1 background noise at 0 dB SNR + 1 interfering speaker at 0 dB SIR)** | | | | | | |
| GCRN [2] | A | 0.410 | 0.044 | 0.093 | -0.052 | -0.038 |
| AV-GCRN [2] | AV | -1.193 | 0.394 | 0.499 | 0.220 | 0.235 |
| AV-Demucs [3] | AV | -5.581 | 0.738 | 0.688 | 0.270 | 0.298 |
| MuSE [8] | AV | -5.528 | 0.787 | 0.679 | 0.276 | 0.299 |
| VisualVoice [9] | AV | -3.781 | 0.606 | 0.645 | 0.249 | 0.270 |
| LA-VocE (audio-only) | A | -3.189 | 0.248 | 0.135 | 0.055 | 0.047 |
| LA-VocE | AV | **-6.653** | **0.931** | **1.100** | **0.294** | **0.333** |
| **Noise condition 2 (3 background noises at -5 dB SNR + 2 interfering speakers at -5 dB SIR)** | | | | | | |
| GCRN [2] | A | -0.416 | -0.010 | 0.163 | -0.015 | -0.015 |
| AV-GCRN [2] | AV | -1.354 | 0.096 | 0.398 | 0.234 | 0.214 |
| AV-Demucs [3] | AV | -5.548 | 0.274 | 0.426 | 0.308 | 0.300 |
| MuSE [8] | AV | -5.314 | 0.297 | 0.409 | 0.308 | 0.289 |
| VisualVoice [9] | AV | -3.388 | 0.164 | 0.367 | 0.253 | 0.237 |
| LA-VocE (audio-only) | A | -2.817 | 0.056 | 0.087 | 0.066 | 0.043 |
| LA-VocE | AV | **-6.863** | **0.511** | **0.700** | **0.379** | **0.397** |
| **Noise condition 3 (5 background noises at -10 dB SNR + 3 interfering speakers at -10 dB SIR)** | | | | | | |
| GCRN [2] | A | -0.414 | -0.015 | 0.210 | -0.020 | -0.005 |
| AV-GCRN [2] | AV | -1.263 | -0.043 | 0.217 | 0.171 | 0.139 |
| AV-Demucs [3] | AV | -4.866 | 0.013 | 0.298 | 0.262 | 0.230 |
| MuSE [8] | AV | -4.185 | 0.011 | 0.242 | 0.231 | 0.182 |
| VisualVoice [9] | AV | -2.518 | -0.045 | 0.248 | 0.181 | 0.160 |
| LA-VocE (audio-only) | A | -1.982 | -0.015 | 0.073 | 0.032 | 0.008 |
| LA-VocE | AV | **-6.170** | **0.159** | **0.447** | **0.371** | **0.358** |

Imperial College London

Meta

# Varying noise/interference

**Table 2**. LA-VocE's performance for different SNR / SIR conditions with 3 background noises and 2 interfering speakers.

| | | PESQ-WB i ↑ | | | | | ESTOI i ↑ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | SNR (dB) | 5 | 0 | -5 | -10 | -15 | 5 | 0 | -5 | -10 | -15 |
| SIR (dB) | 5 | 0.970 | 0.876 | 0.715 | 0.486 | 0.245 | 0.269 | 0.316 | 0.356 | 0.375 | 0.362 |
| | 0 | 0.904 | 0.795 | 0.630 | 0.411 | 0.210 | 0.327 | 0.354 | 0.375 | 0.378 | 0.355 |
| | -5 | 0.789 | 0.679 | 0.511 | 0.319 | 0.136 | 0.386 | 0.394 | 0.397 | 0.383 | 0.349 |
| | -10 | 0.617 | 0.523 | 0.405 | 0.248 | 0.092 | 0.429 | 0.426 | 0.414 | 0.388 | 0.344 |
| | -15 | 0.438 | 0.383 | 0.289 | 0.195 | 0.081 | 0.443 | 0.433 | 0.414 | 0.381 | 0.330 |

**Table 3**. LA-VocE's performance for different numbers of background noises and interfering speakers (-5 dB SNR / SIR).

| | | PESQ-WB i ↑ | | | | | ESTOI i ↑ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | # noises | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| # spk. | 1 | 0.709 | 0.642 | 0.601 | 0.580 | 0.557 | 0.396 | 0.402 | 0.404 | 0.404 | 0.403 |
| | 2 | 0.602 | 0.553 | 0.511 | 0.497 | 0.482 | 0.396 | 0.398 | 0.397 | 0.395 | 0.393 |
| | 3 | 0.539 | 0.490 | 0.462 | 0.455 | 0.431 | 0.390 | 0.390 | 0.388 | 0.387 | 0.384 |

Imperial College London

Meta

# Demo


Rodrigo Mira et al.

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

Imperial College London

Meta

# Conclusion

**Check out our project page here (with demos!)**

- We propose LA-VocE, a new AVSE model that excels in low-SNR scenarios with interfering speech.
- Check out more demos on our project page https://sites.google.com/view/la-voce-avse (or scan the QR code on the left)
- The full paper is available on arXiv https://arxiv.org/abs/2211.10999
- In the future, it would be promising to adapt this framework for real-time synthesis, to enable AV enhancement in live video calls, for example.
- Thank you for watching! 😀