Speech inpainting: Context-based speech synthesis guided by video

Juan F. Montesinos[1], **Daniel Michelsanti**[2,3], Gloria Haro[1], Zheng-Hua Tan[2], Jesper Jensen[2,3]

[1]Universitat Pompeu Fabra, Department of Information and Communication Technologies, Spain
[2]Aalborg University, Department of Electronic Systems, Denmark
[3]Oticon A/S, Denmark

INTERSPEECH 2023 - 24 August 2023

Agenda

- Motivation
- Audio-Visual Speech Inpainting
- Approach
- Experiments
- Results
- Conclusions

Motivation

- Speech is one of the most common multimodal events in our daily life.
- Nowadays, we are constantly exposed to a lot of speech signals from digital content.
- Sometimes, the audio stream is corrupted due to, e.g., muted microphones, external noises or transmission losses.
- If the corrupted audio segment is short (<200 ms), **audio inpainting** approaches can be used to restore the missing information.
- What if the corrupted segments are longer?

Audio-Visual Speech Inpainting

Can we restore the lost speech information from
the **audio context** and the **uncorrupted visual information**?

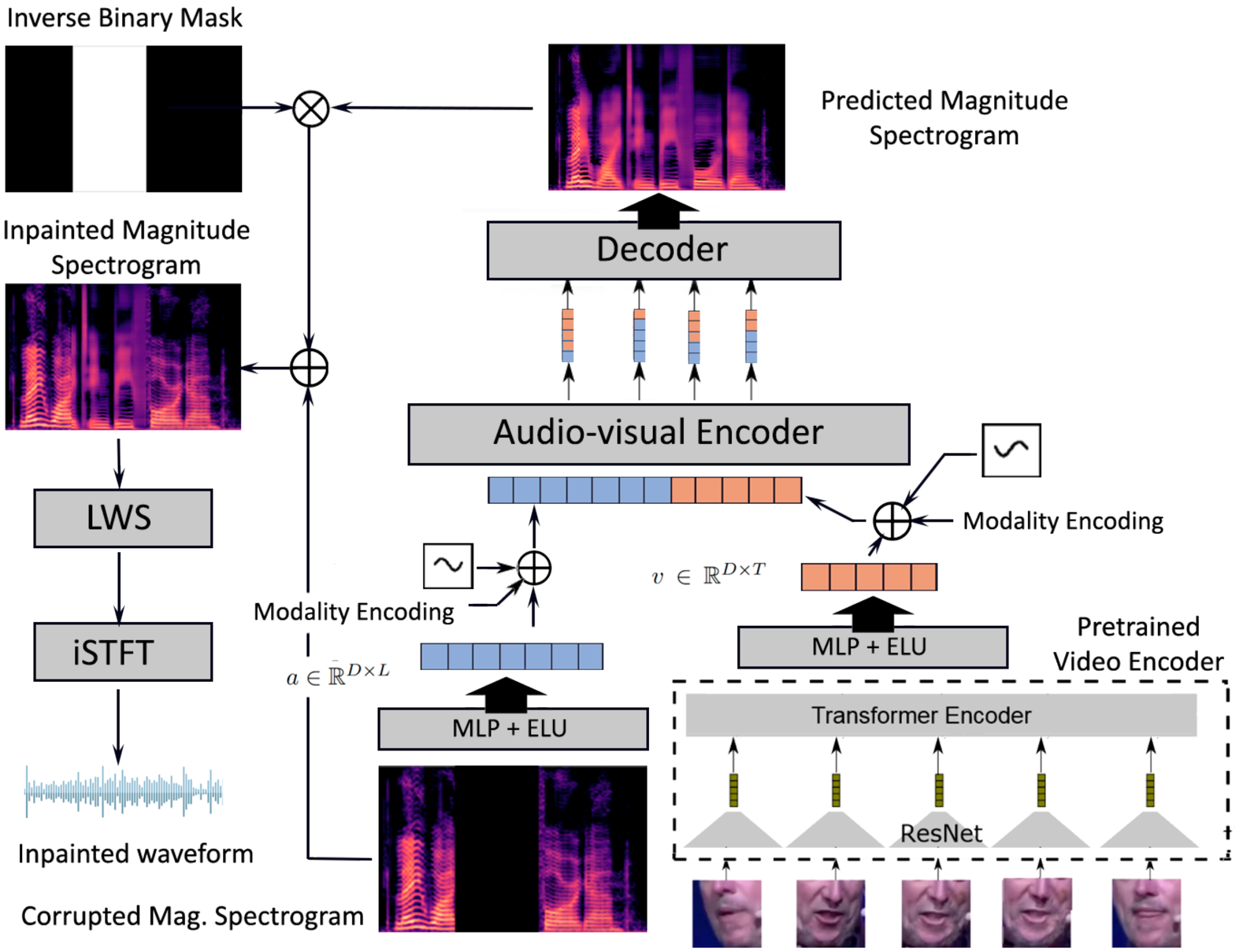
Contributions

1. We propose a **transformer architecture** for audio-visual speech inpainting.

2. We show that speech inpainting can benefit from using **high-level visual features** extracted with the Audio-Visual HuBERT Network (AV-HuBERT) [10], whose effectiveness for related tasks has previously been reported.

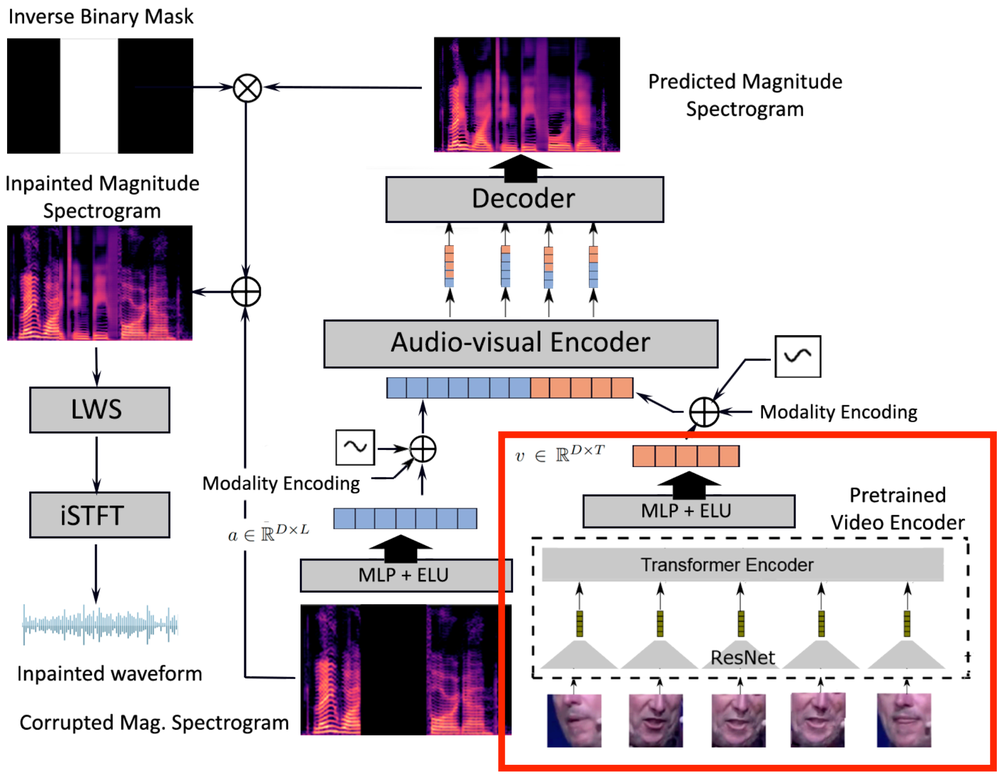
[10] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, “Learning audio-visual speech representation by masked multimodal cluster prediction,” ICLR, 2022.

Approach


Inverse Binary Mask
Predicted Magnitude Spectrogram
Inpainted Magnitude Spectrogram
Decoder
Audio-visual Encoder
LWS
Modality Encoding
iSTFT
Modality Encoding
$v \in \mathbb{R}^{D \times T}$
$a \in \mathbb{R}^{D \times L}$
MLP + ELU
Pretrained Video Encoder
MLP + ELU
Transformer Encoder
ResNet
Inpainted waveform
Corrupted Mag. Spectrogram

Feature Extraction (Video)

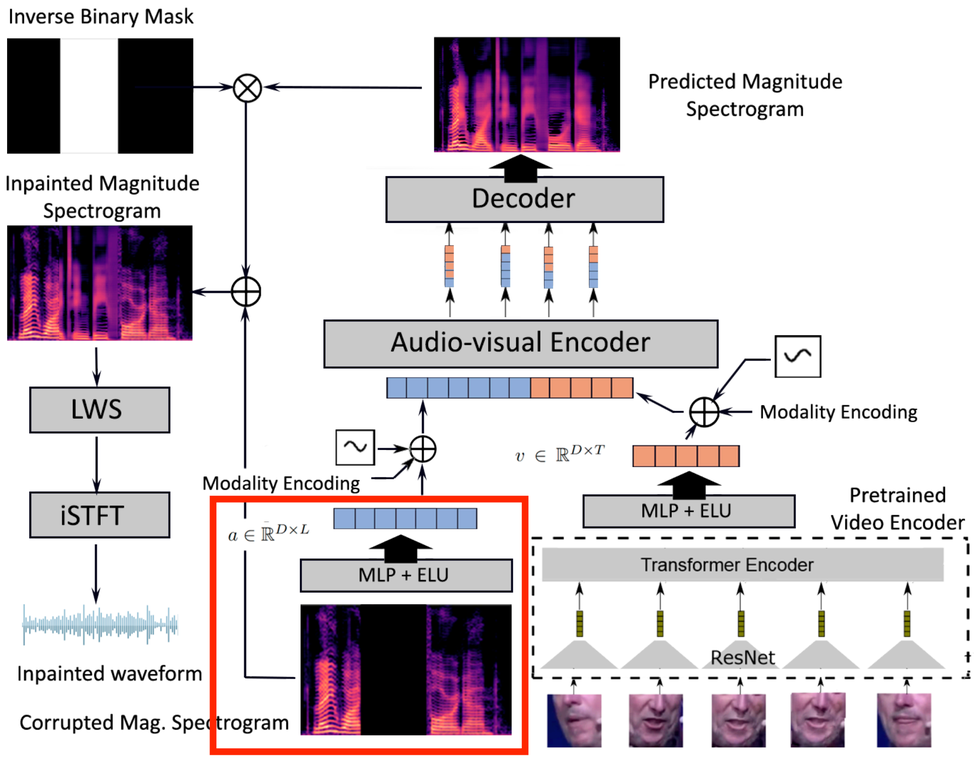
- We extract high-level visual features using the **AV-HuBERT's** [10] **video encoder**, which processes the sequence of video frames using a ResNet followed by a transformer encoder to model the temporal dependencies.
- In addition, we get visual embeddings using a **multi-layer perceptron** (MLP) with **exponential linear unit** (ELU) activation on top.



[10] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," ICLR, 2022.

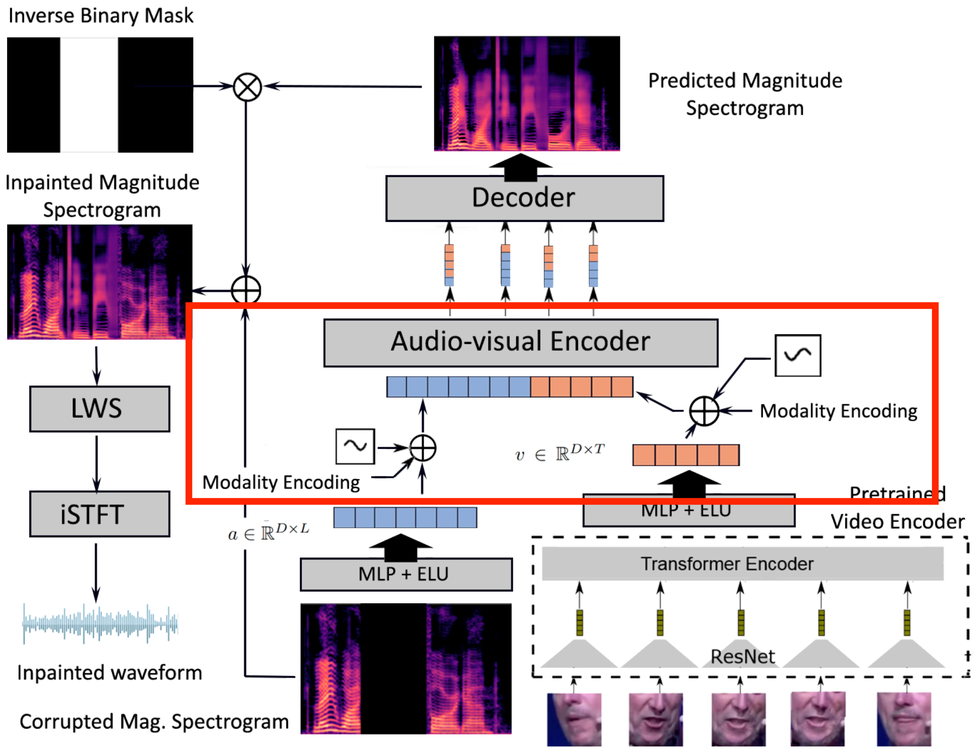
Feature Extraction (Audio)

- In order to extract learned acoustic features, we use a similar **MLP** that takes as input the masked spectrogram.



Multimodal Fusion

- We sum a **positional encoding** that reflects the temporal sorting of the elements in the sequence [16] and a **modality encoding** that transmits whether each element is an acoustic or a visual feature [15].
- We construct the AV embedding by concatenating both modalities temporally.
- **Temporal concatenation** is robust to out-of-sync signals.
- Then, a six-block **transformer encoder** ingests the AV embedding.

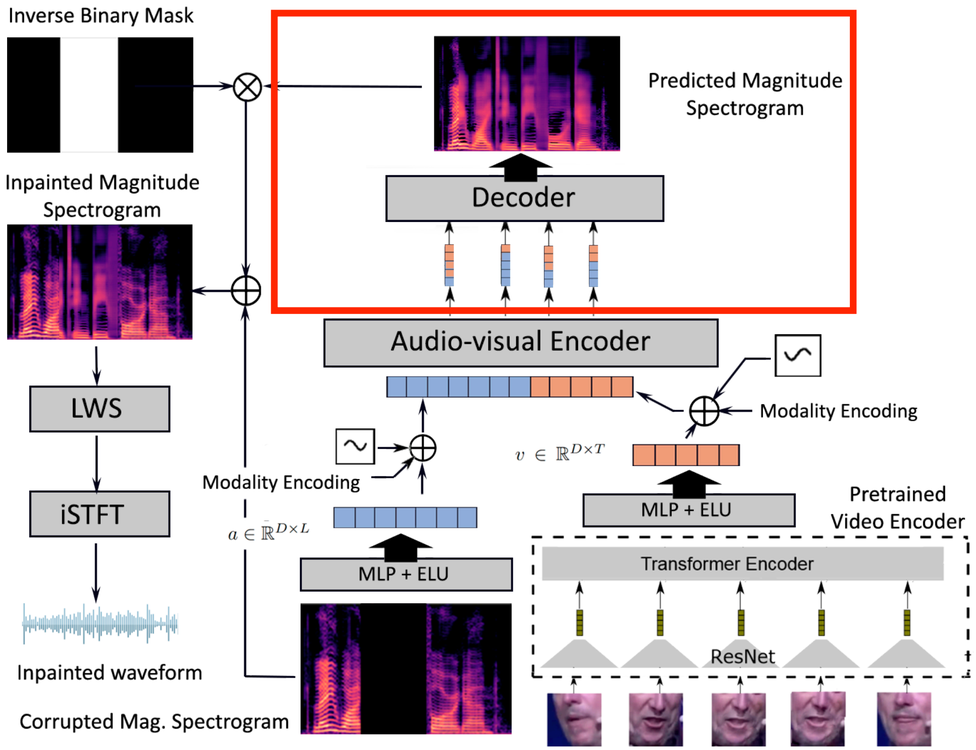


[15] H.Chen,W.Xie,T.Afouras,A.Nagrani,A.Vedaldi,andA.Zisserman, "Audio-visual synchronisation in the wild," 32nd British Machine Vision Conference (BMVC), 2021.

[16] J. F. Montesinos, V. S. Kadandale, and G. Haro, "Vovit: Low latency graph-based audio-visual voice separation transformer," in European Conference on Computer Vision (ECCV), 2022.

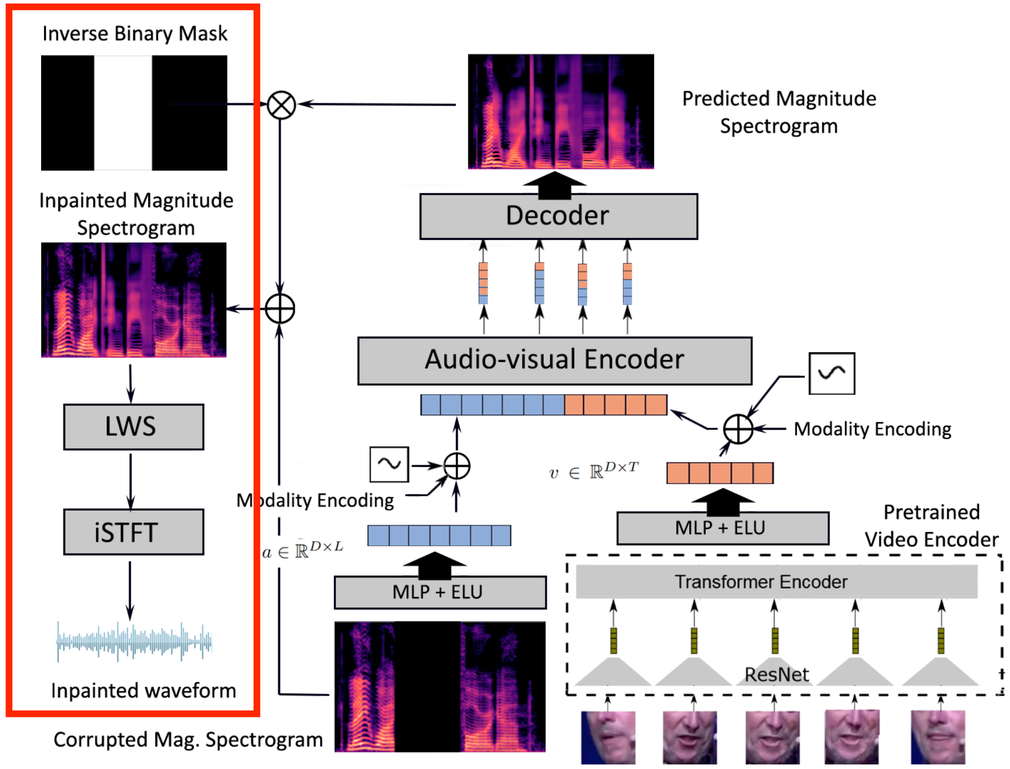
Inpainting

- We use a seven-block **transformer** that processes the high-level features generated by the encoder to provide an estimate of the underlying uncorrupted speech magnitude spectrogram.
- The transformer's role is two-fold:
  1. It acts as an **auto-encoder**, i.e. it reconstructs the uncorrupted segment of the audio.
  2. It **inpaints** the corrupted segment.



Waveform Reconstruction

- We estimate the phase of the underlying uncorrupted speech spectrogram using **Local Weighted Sums** (LWS) [17] and then compute the inverse STFT to recover the waveform.
- Other approaches (e.g. GAN-based) can be used to reconstruct the waveform.



[17] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency," in Proc. DAFx, vol. 10, 2010, pp. 397–403.

Training Parameters

- **Training loss:** $L(A, \hat{A}) = \alpha\, MAE(\hat{A}^c, A^c) + \beta\, MAE(\hat{A}^u, A^u)]$.
  We set $\alpha = 10$ and $\beta = 1$, so the network is forced to focus on the inpainting task, as it is much harder than the auto-encoding task.
- **Batch size:** 10.
- **Learning rate:** 10^{-4}.

Experiments

Datasets

- **GRID** [8] - Controlled environment and small vocabulary.
  Speaker-independent setting:
  - *Training set*: 25 speakers, 1000 utterances per speaker (3 s each).
  - *Validation set*: 4 speakers, 1000 utterances per speaker (3 s each).
  - *Test set*: 4 speakers, 1000 utterances per speaker (3 s each).
- **Voxceleb2** [9] - In-the-wild recordings and unconstrained vocabulary.
  Speaker-independent setting using 2 s excerpts.

- We corrupt the speech data with fullband temporal gaps of a duration between 160 and 1600 ms.

[8] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421–2424, 2006.

[9] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in Interspeech, 2018.

Baselines

- **Audio-Only Baseline** - Model trained without using visual information as input.
- **Audio-visual Baseline** [7] - Framework consisting of a stack of three Bi-LSTM layers.
  - *Acoustic features*: normalised log magnitude spectrograms.
  - *Visual features*: landmark-based motion vectors.
  - *Fusion*: Concatenation of features.
  - *Loss*: Mean squared error between the predicted log magnitude spectrogram and the ground-truth one in the corrupted segment.

[7] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen, “Audio-visual speech inpainting with deep learning,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6653–6657.

Results

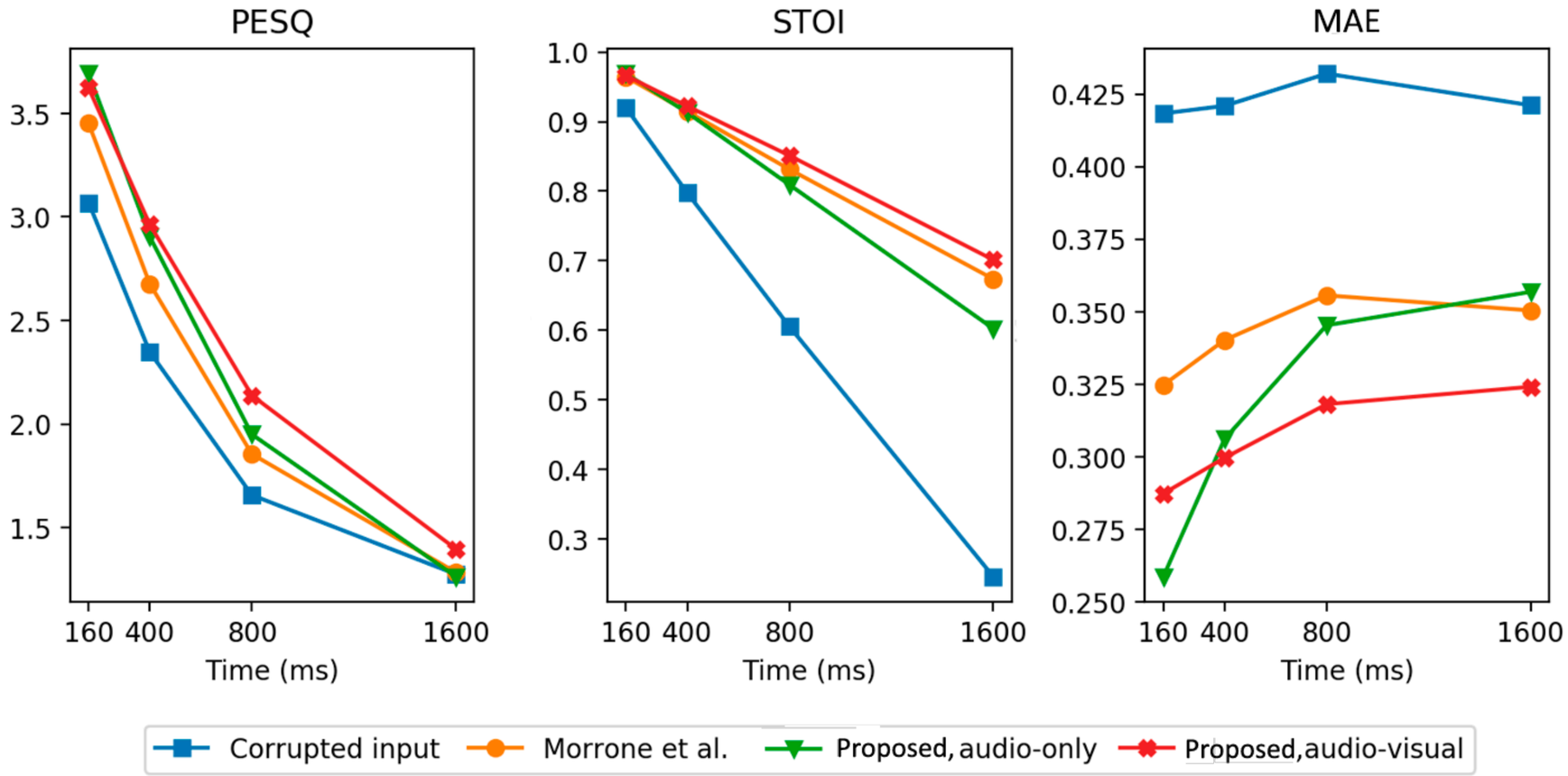
GRID

- Performance are measured in terms of estimated **speech quality** (PESQ), estimated **speech intelligibility** (STOI), and **mean absolute error** (MAE).
- PESQ and STOI are computed using the whole signal, while MAE is computed only within the corrupted segment.
- Our audio-visual approach outperforms both its audio-only counterpart and the previous state-of-the-art audio-visual model.

	PESQ ↑	STOI ↑	MAE ↓
Corrupted input	1.78	0.58	0.43
Morrone et al. [7]	1.98	0.79	0.39
Proposed, audio-only	2.07	0.79	0.34
Proposed, audio-visual	**2.21**	**0.84**	**0.31**

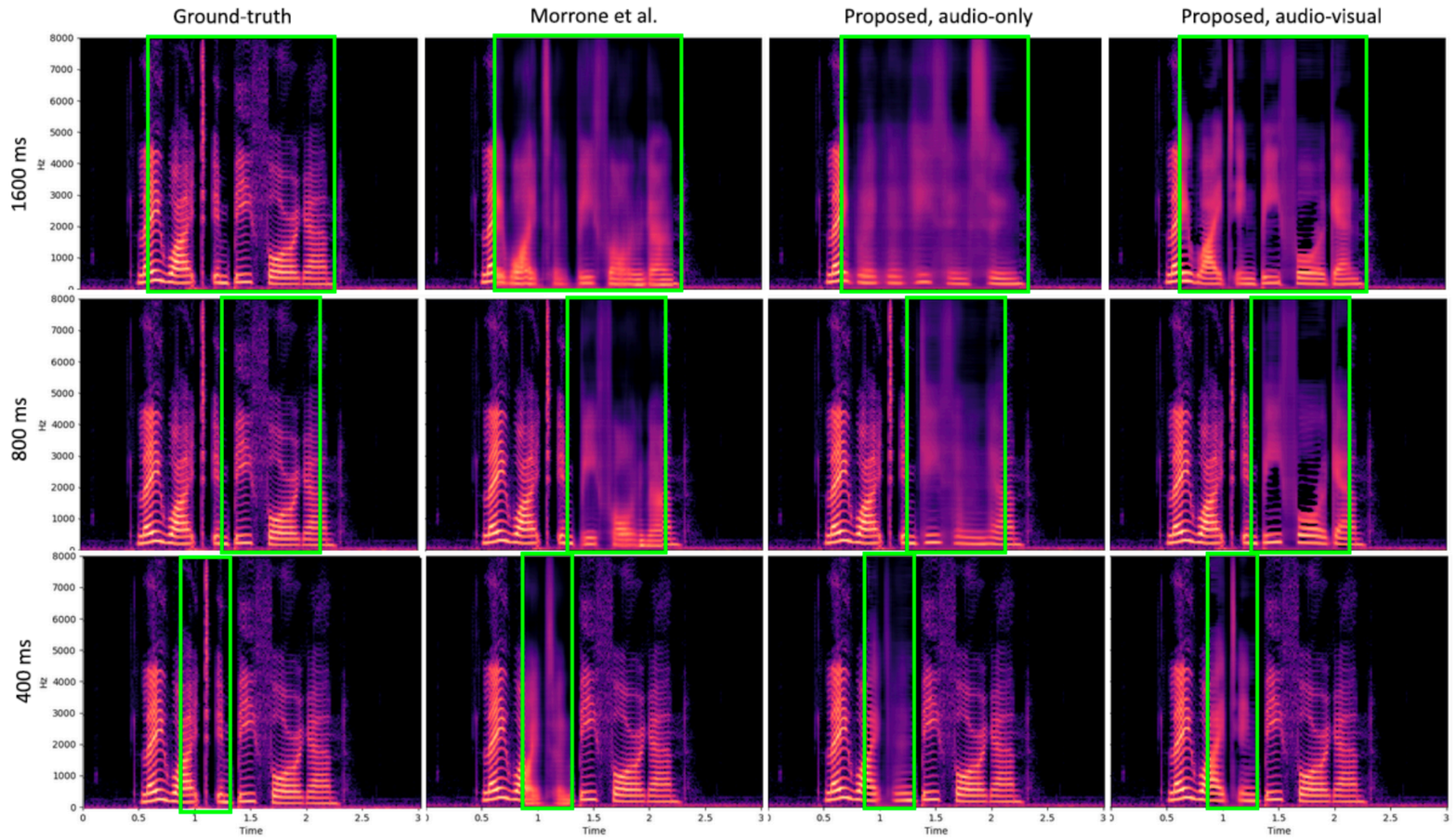
[7] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen, “Audio-visual speech inpainting with deep learning,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6653–6657.

GRID - Performance vs Segment duration



[7] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen, “Audio-visual speech inpainting with deep learning,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6653–6657.

GRID - Inpainting Example



[7] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen, “Audio-visual speech inpainting with deep learning,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6653–6657.

Voxceleb2

- The results show how the proposed AV model is capable of generating meaningful speech on in-the-wild scenarios with unconstrained vocabulary, unlike the baseline and the AO model.
- This reflects the capabilities of AV models to synthesize speech in complex scenarios.

	PESQ ↑	STOI ↑	MAE ↓
Corrupted input	1.37	0.43	0.56
Proposed, audio-visual	**1.95**	**0.70**	**0.37**

Demo

You are welcome to listen to some inpainted examples on the project website:

https://ipcv.github.io/avsi/

Conclusions

Conclusions

- We presented a **new state-of-the-art AVSI model** that can inpaint long gaps, up to 1600 ms, for unseen-unheard speakers.
- We tested our model in the Grid Corpus and showed that it outperforms its audio-only counterpart for gaps larger than 160 ms, and the previous state-of-the-art approach.
- We showed that the **visual features** extracted **from** the **AV-HuBERT** network **encode enough information** to guide the inpainting process.
- We showed that **our model can inpaint natural speech** in in-the-wild scenarios (Voxceleb2 dataset).

Thanks!