46th International Conference on Acoustics, Speech, and Signal Processing
Toronto, June 6-11, 2021

Audio-Visual Speech Inpainting with Deep Learning

Giovanni Morrone, Daniel Michelsanti, Zheng-Hua Tan, Jesper Jensen

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AALBORG UNIVERSITY
DENMARK

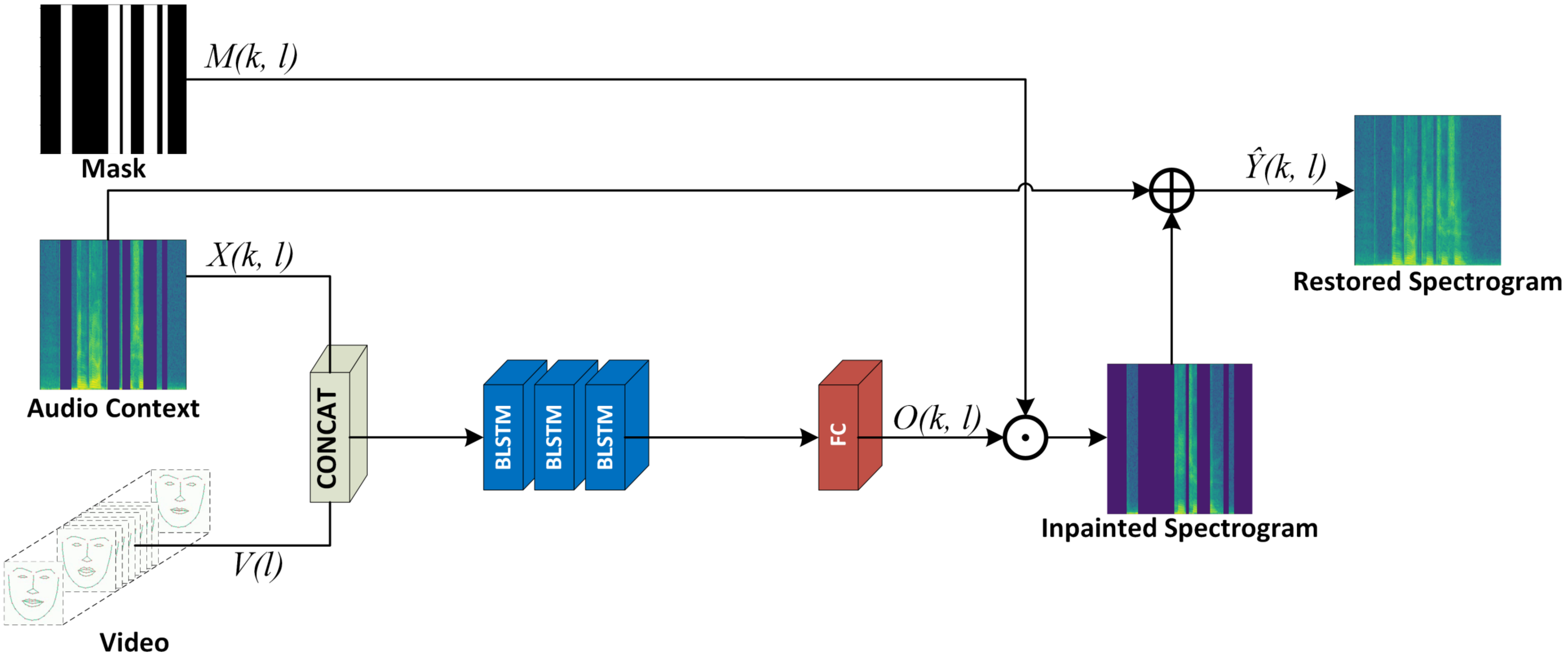
Motivation

- In real life applications, audio signals are often corrupted by accidental distortions, such as impulsive noises, clicks and transmission errors.
- **Speech Inpainting:** the process of restoring the lost speech information from the audio context.
- In our paper, we address the problem of **Audio-Visual Speech Inpainting**: in addition to reliable audio-context, uncorrupted visual information is exploited.
- This approach is beneficial especially when the time gaps are large (> 400 ms).
- Visual information was successfully used in many speech-related tasks (e.g., speech recognition, speech enhancement, speech separation, etc.), but it has not adopted for speech inpainting yet.

AV Speech Inpainting

- We use a deep learning model based on Bi-directional Long-Short Term Memories (LSTM).
- The model works in the spectrogram domain and uses facial landmarks motion (Morrone et al., 2019) as visual features.
- As done in previous work, we assume to know a priori the location of uncorrupted and lost data. This information is used in the signal reconstruction stage.

System Architecture



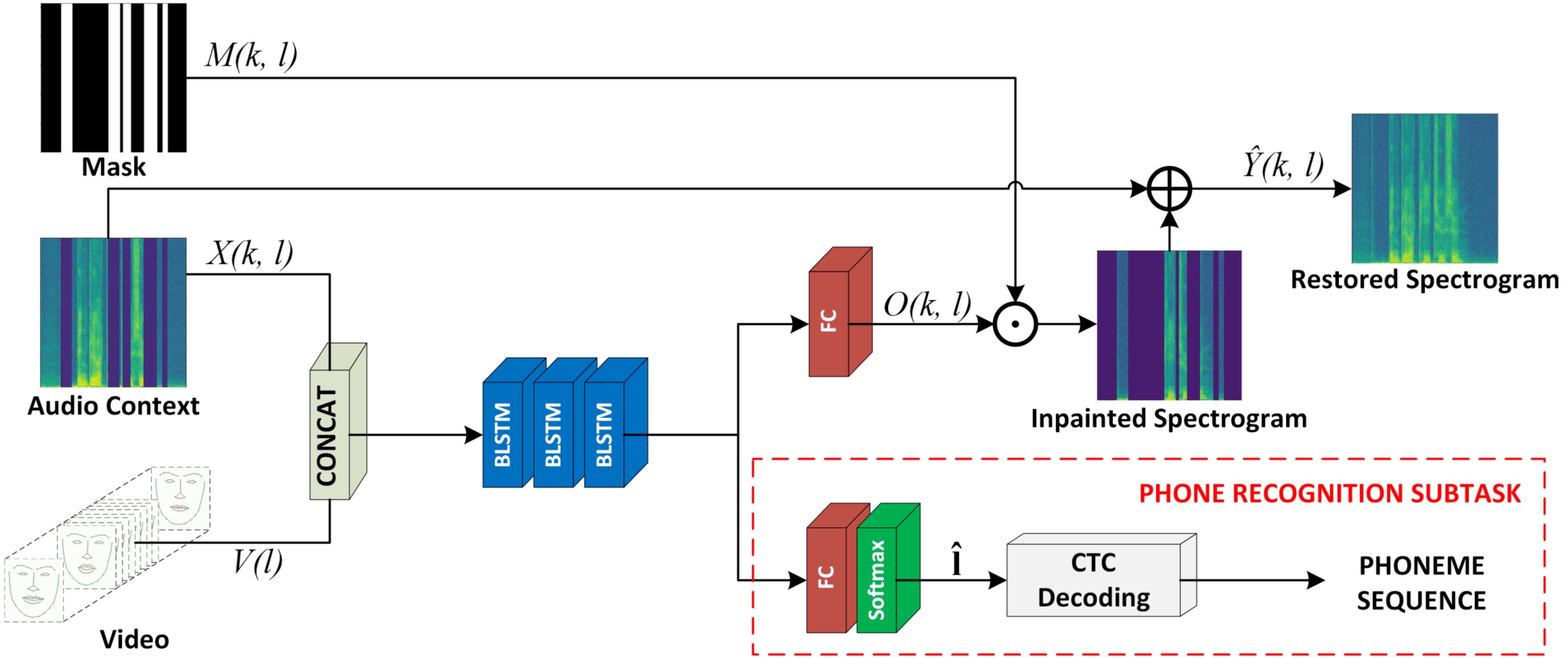
Mask: uncorrupted/lost time-frequency bins $\oplus$: element-wise sum $\odot$: element-wise product

Multi-Task Learning Approach

- In addition, we propose a Multi-Task Learning (MTL) approach, which attempt to perform speech inpainting and phone recognition simultaneously.
- This strategy allows the distillation of phonetic information during training.
- The MTL training makes use of a Connectionist Temporal Classification (CTC) loss to compute the error between the phone posteriors and the ground-truth phone labels.
- The MTL loss, J_{MTL}, consists of a weighted sum between the inpainting loss, J_{MSE}, and the CTC loss, J_{CTC} :

$$J_{MTL} = J_{MSE} + \lambda \cdot J_{CTC}, \lambda \in \mathbb{R}$$

MTL System Architecture



Mask: reliable/unreliable time-frequency bins
CTC: Connectionist Temporal Classification

$\oplus$: element-wise sum
$\odot$: element-wise product

Experimental Setup

- Dataset: GRID corpus (Cooke et al., 2006). Speaker-independent setting:
  - Training set: 25 speakers, 1000 utterances per speaker.
  - Validation set: 4 speakers, 1000 utterances per speaker.
  - Test set: 4 speakers, 1000 utterances per speaker.
- We generate a corrupted version of the GRID corpus where random missing time gaps with different durations are introduced in audio speech signals.
- To assess the performance of the AV models, we devise an audio-only baseline models by simply removing the video input, leaving the rest unchanged.
- Hyperparameters:
  - BLSTM: 3 layers, 250 hidden units per layer
  - Optimizer: Adam
  - Learning rate: 0.001
  - Mini-batch size: 8
  - λ weight MTL loss: 0.001

Evaluation Results

We evaluate our systems with 4 metrics: L1 loss, PER[1] (Phone Error Rate), and two perceptual metrics, STOI and PESQ.

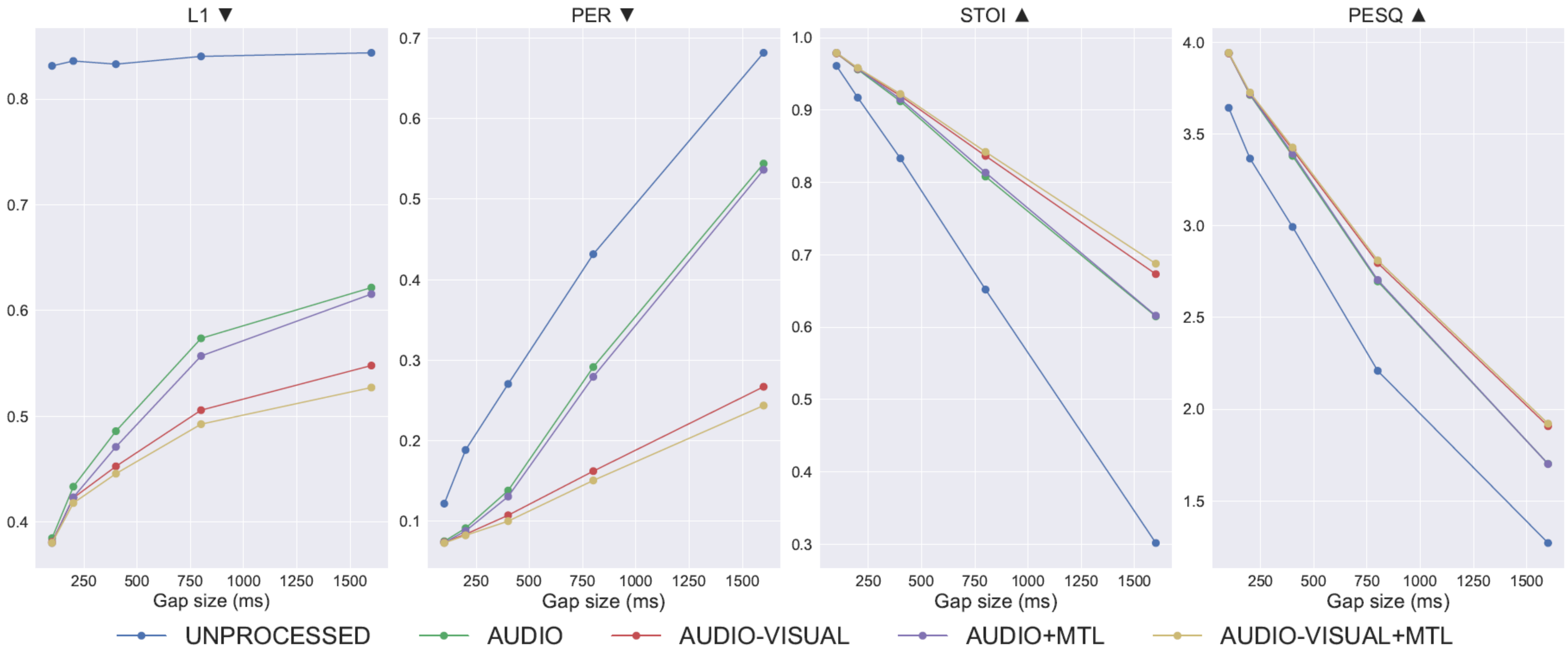
A	V	MTL	L1 ▼	PER ▼	STOI ▲	PESQ ▲
Unprocessed			0.838	0.508	0.480	1.634
✗			0.482	0.228	0.794	2.458
✗	✗		0.452	0.151	0.811	2.506
✗		✗	0.476	0.214	0.799	2.466
✗	✗	✗	**0.445**	**0.137**	**0.817**	**2.525**

A: Audio V: Video MTL: multi-task learning with CTC

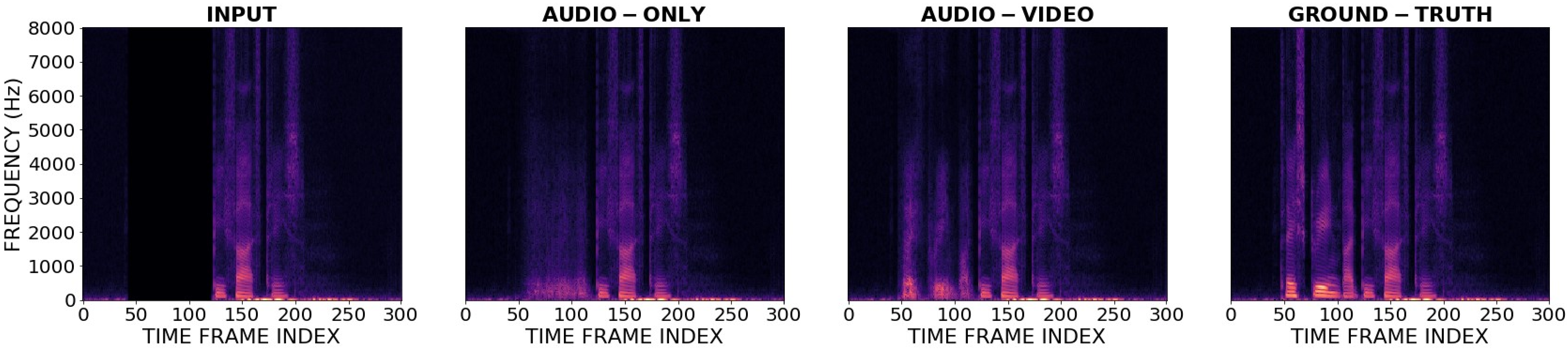
- AV models outperform the audio-only counterparts on all metrics.
- The MTL strategy is beneficial.

[1]PER is obtained with a phone recognizer trained on uncorrupted data. The PER score of uncorrupted speech is 0.069.

Time Gap Analysis



Example - 800 ms Time Gap



Conclusion

- To the best of our knowledge, this is the first work that exploits vision for the speech inpainting task.
- Audio-visual models strongly outperform audio-only models.
- Audio-only approach degrades rapidly when missing time gaps get large.
- Audio-visual approach is still able to plausibly restore missing information for very long time gaps (> 400 ms).
- Learning a phone recognition task together with the inpainting task leads to better results, although its contribution to performance is lower compared to vision.

Thanks for your attention!

Contacts:

Giovanni Morrone (giovanni.morrone@unimore.it)
Daniel Michelsanti (danmi@es.aau.dk)
Zheng-Hua Tan (zt@es.aau.dk)
Jesper Jensen (jje@es.aau.dk)

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AALBORG UNIVERSITY
DENMARK