AUDIO-VISUAL SPEECH INPAINTING WITH DEEP LEARNING

Giovanni Morrone[1], Daniel Michelsanti[2], Zheng-Hua Tan[2], Jesper Jensen[2,3]
[1]*University of Modena and Reggio Emilia,* [2]*Aalborg University,* [3]*Oticon A/S*

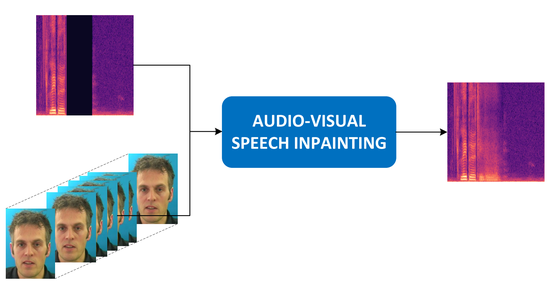
PAPER #1217

PROBLEM DESCRIPTION

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 reliable audio context.

Audio-Visual Speech Inpainting: in addition to reliable audio-context, uncorrupted visual information is exploited.



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.

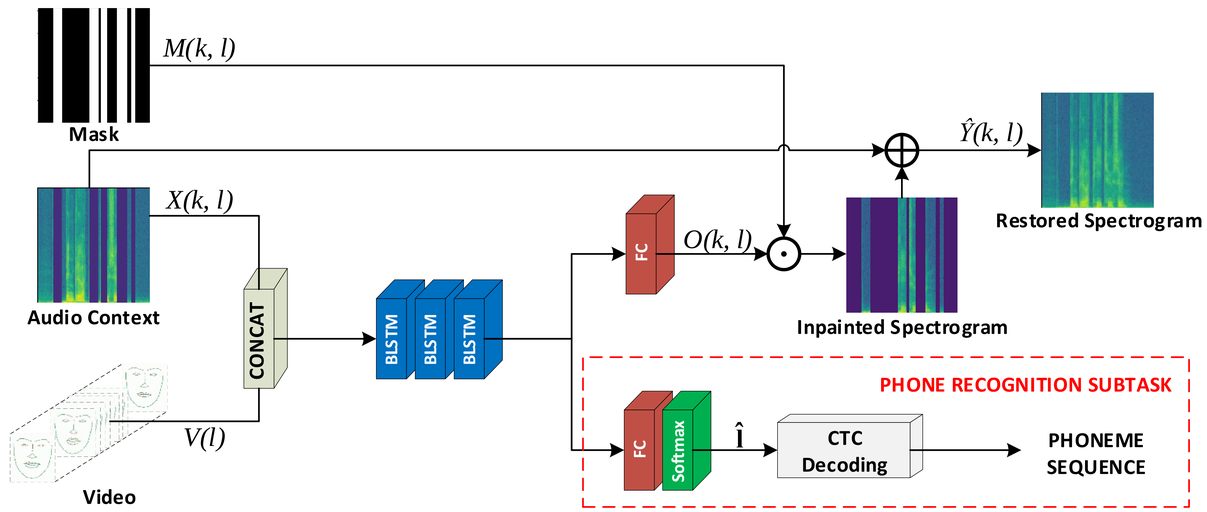
We propose the use of vision to solve the speech inpainting task.

AUDIO-VISUAL SPEECH INPAINTING MODELS

- We employ 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.

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, by using the Connectionist Temporal Classification (CTC) algorithm.
- The MTL loss consists of a weighted sum between the inpainting loss and the CTC loss.



Legenda

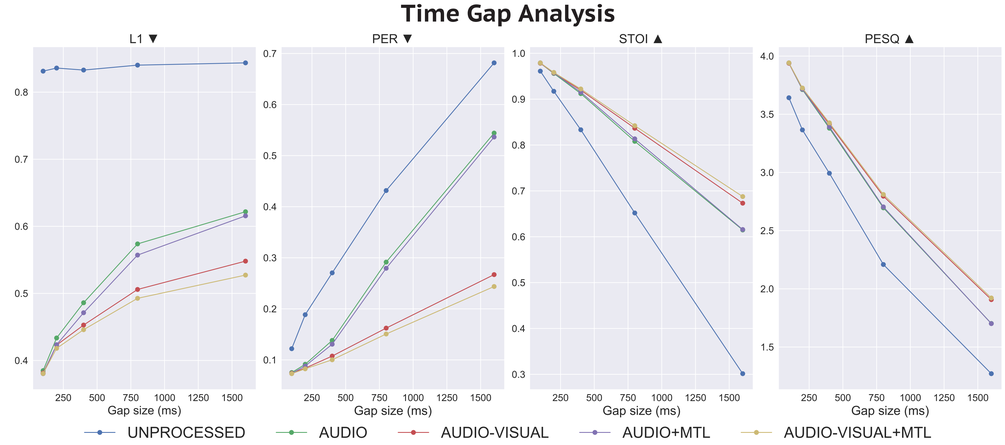
$M(k,l)$: reliable/corrupted indicator mask $X(k,l)$: audio context input $V(p,l)$: video input
$O(k,l)$: missing audio inpainted $\hat{Y}(k,l)$: reconstructed audio
$\hat{l}$: phone posteriors probabilities

EXPERIMENTAL RESULTS

- Dataset: **GRID** corpus (Cooke et al., 2006). Speaker-independent setting.
- We generate a corrupted version of the dataset by introducing random missing time gaps in speech signals.
- Audio-only baseline: we remove the video input, leaving the rest unchanged.

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**

Time Gap Analysis



DISCUSSION AND 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 on all metrics.
- Audio-only approach degrades rapidly when missing time gaps get large.
- In general, audio-only models inpaint the gap with a stationary signal, whose energy is concentrated in the low frequencies.
- Audio-visual approach is still able to plausibly restore missing information for extremely long time gaps (> 400 ms).
- The PER scores are way better for audio-visual models, and demonstrate that vision improves a lot speech intelligibility.
- Learning a phone recognition task together with the inpainting task leads to better results, although its contribution to performance is lower compared to vision.

Project page: `https://dr-pato.github.io/audio-visual-speech-inpainting/`

Contacts:
giovanni.morrone@unimore.it
danmi@es.aau.dk
zt@es.aau.dk
jje@es.aau.dk

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