Video-guided speech inpainting transformer

Juan F. Montesinos
Universitat Pompeu Fabra

Daniel Michelsanti
Aalborg University, Oticon A/S

Gloria Haro
Universitat Pompeu Fabra

Zheng-Hua Tan
Aalborg University

Jesper Jensen
Aalborg University, Oticon A/S

Abstract

Audio and visual modalities are inherently connected in speech signals: lip movements and facial expressions are correlated with speech sounds. This motivates studies that incorporate the visual modality to enhance an acoustic speech signal or even restore missing audio information. Specifically, this paper focuses on the problem of audio-visual speech inpainting, which is the task of synthesizing the speech in a corrupted audio segment in a way that it is consistent with the corresponding visual content and the uncorrupted audio context. We present an audio-visual transformer-based deep learning model that leverages visual cues that provide information about the content of the corrupted audio. It outperforms the previous state-of-the-art audio-visual model and audio-only baselines. We also show how visual features extracted with AV-HuBERT, a large audio-visual transformer for speech recognition, are suitable for synthesizing speech.

1. Introduction

Speech is one of the most common multimodal events in our daily life. Thanks to the expansion of the Internet, we are exposed to a lot of speech signals from digital content as well: news, social networks, virtual meetings and video calls. Sometimes, the audio stream is corrupted due to, e.g., muted microphones, external noises or transmission losses. One solution is to estimate the lost audio information, saving content creators the time to re-make their videos or avoiding a speaker to repeat a sentence. The process of restoring the corrupted audio signal is known as *audio inpainting* [1]. Carrying out such a restoration for long segments of corrupted audio (>200ms) is not a simple task, as there is no prior information about the missing content. There are several ways to address the problem. From an audio-only (AO) perspective, the work in [7] relies on a generative adversarial network approach to generate realistic speech content for a gap size up to 500 ms. In [3], an encoder-decoder architecture is used to inpaint the audio in both time-frequency and time domain for segments up to 250 ms. The works in [12] and [10] propose a similar idea operating only in the time-frequency domain for gaps up to 64 ms and 400 ms, respectively.

Some works use additional modalities, that are not affected by the acoustic noise, as cues to guide the inpainting process. This allows to inpaint larger gaps. For example, [2] uses text to guide the inpainting process of audio gaps up to 1000 ms, relying on transformers and contrastive learning. In [15], video information is extracted from face landmarks to inpaint gaps up to 1600 ms using Bi-directional Long-Short Term Memory (Bi-LSTM) units. The task of restoring a missing audio segment by leveraging the visual information of the speaker is known as audio-visual speech inpainting (AVSI). We present in this work an AVSI deep learning model which can restore long gaps of speech. In contrast to [15], we use high-level visual features useful for speech recognition and a multi-modal transformer that allows to establish long range interactions across the audio and visual modalities, while being robust to potential misalignments among them. Moreover, the work in [15] was limited to a constrained dataset of non natural speech [6], while we train and test our model in a large-scale dataset of natural and unconstrained speech [5] (in addition to [6]).

The contribution of this paper is two-fold: First, we propose a transformer architecture that analyzes a time-frequency representation of the corrupted audio signal and the corresponding uncorrupted visual information to synthesize intelligible speech even for a long corrupted audio segment, obtaining state-of-the-art results. Secondly, we show that speech inpainting can benefit from using high-level visual features extracted with the Audio-Visual HuBERT Network (AV-HuBERT) [19], whose effectiveness for related tasks has previously been reported.

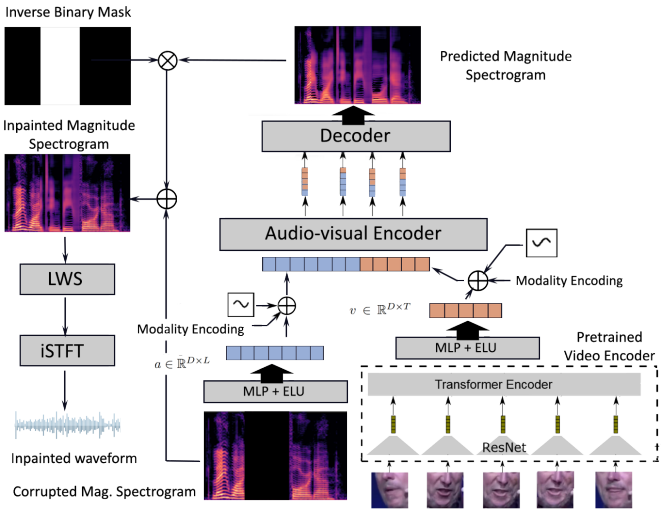

Figure 1. Proposed audio-visual model. The pre-trained video encoder corresponds to [19].

2. Approach

Let $x[t]$ be a discrete-time acoustic speech signal and $X = \{X(k,l); k = 0,\dots,K-1; l = 0,\dots,L-1\}$ be the corresponding short-time Fourier transform (STFT), where k and l indicate frequency and time indices, respectively. Furthermore, let $\mathcal{A} \in \mathbb{R}^{K\times L}$ denote a magnitude spectrogram matrix defined from the element-wise absolute values of the elements in X and $\mathcal{M} \in \mathbb{R}^{K\times L}$ a binary mask that provides the position of the corrupted region of the spectrogram [15, 17]). Then, the inpainted magnitude spectogram, $\mathcal{Q} \in \mathbb{R}^{K\times L}$, can be defined as $\mathcal{Q} = \mathcal{M}\odot\mathcal{A} + (\mathbf{1}-\mathcal{M})\odot\hat{\mathcal{A}}$, where $\odot$ indicates the element-wise product and $\hat{A} \in \mathbb{R}^{K\times L}$ denotes an estimated speech STFT magnitude matrix. For the binary mask matrix $\mathcal{M}$, we assume that the i-th column consists of ones if the i-th column of A is uncorrupted and zeros otherwise.

AVSI leverages the video stream to improve speech inpainting, by providing information about the acoustic speech content within the corrupted region. Our processing pipeline is divided into four different stages: feature extraction, multi-modal fusion, inpainting process and waveform reconstruction. The whole process is depicted in Fig. 1.

In the feature extraction stage, we extract high-level visual features using the AV-HuBERT's [19] video encoder, which processes the sequence of video frames using a ResNet [8] followed by a transformer encoder to model the temporal dependencies. In addition, we use a simple multi-layer perceptron (MLP) with exponential linear unit (ELU) activation on top, leading to a signal $v \in \mathbb{R}^{D\times T}$, where D is the dimensionality of our embeddings and T the amount of frames. In order to extract learned acoustic features, we use a similar MLP that takes as input the masked spectrogram $X \odot \mathcal{M}$, resulting in a signal $a \in \mathbb{R}^{D\times L}$ (see Fig. 1). The aim behind this design is to process each audio frame independently, as we assume many audio frames can be corrupted.

In the multi-modal fusion stage, the goal is to fuse the acoustic and visual features, learning the relationship between both. To do so, we rely on a six-block transformer encoder that ingests an audio-visual (AV) embedding. We construct the AV embedding by concatenating both modalities temporally, as in [19, 16, 9, 4]. Since the transformer is unaware of the position or the modality type of each element in the sequence, we sum a positional encoding (pe) that reflects the temporal sorting of the elements in the sequence [14] and a modality encoding (me) that transmits whether each element is an acoustic or a visual feature [4], obtaining: $(a\,;v) = (pe_a + me_a + a\,; pe_v + me_v + v)$, where $(\cdot\,;\cdot)$ denotes the concatenation of two sequences, resulting in an AV sequence $(a\,;v) \in \mathbb{R}^{D\times(T+L)}$. Alternatively, channel-wise concatenated AV embeddings can be used. Compared to the temporal concatenation, channel-wise concatenation would require an extra hyperparameter related to the number of features devoted to the visual and acoustic signals. Besides, due to the difference in the sampling rate, visual features should be upsampled to the temporal size of the audio ones. On the other hand, we empirically found that, in case of an out-of-sync AV stream, temporal concatenation results in predictions which are shifted in time, whereas in the channel-wise case, the system collapses and generates mumbling. An out-of-sync AV stream may occur due to software or hardware issues: codecs, latency, missing frames and it is frequent in low-quality videos. The robustness of the temporal concatenation to out-of-sync (i.e. misaligned) audio-visual pairs was also noticed in [16]. The downside effect is that the final sequence is larger, thus implying more computational cost.

In the inpainting stage, we use a seven-block transformer that processes the high-level features generated by the encoder to provide an estimate $\hat{\mathcal{A}}$ of the underlying uncorrupted speech magnitude spectogram. At this stage, the transformer's role is two-fold: It has to act as an auto-encoder, i.e. reconstruct the uncorrupted segment of the audio, and it has to inpaint the corrupted segment.

In the waveform reconstruction stage, we estimate the phase of the underlying uncorrupted speech spectrogram using Local Weighted Sums (LWS) [11] and then compute the inverse STFT to recover the waveform, as done in [15] (for a fair comparison).

Training Loss. We use a weighted mean absolute error (*MAE*):

$$\mathcal{L}(\mathcal{A},\hat{\mathcal{A}}) = \alpha\, MAE(\hat{\mathcal{A}}^c, \mathcal{A}^c) + \beta\, MAE(\hat{\mathcal{A}}^u, \mathcal{A}^u), \quad (1)$$

where $\alpha, \beta \geq 0$, and the superindices c and u denote corrupted and uncorrupted parts, respectively. We set $\alpha > \beta$, so that the network is forced to focus on the inpainting task,

as it is much harder than the auto-encoding task (we use $\alpha = 10$ and $\beta = 1$). While training, we use the loss (1), so the network predicts the whole spectrogram (both the corrupted and uncorrupted parts). At inference, once the spectrogram is predicted, we replace the known parts of the spectrogram via masking as shown in Figure 1.

3. Experiments

Datasets. We train our model and the baselines using two different datasets: the *Grid Corpus* [6] and the *Voxceleb2* dataset [5].

The *Grid Corpus* is a 30-hour AV dataset consisting of 33 speakers recorded in a controlled environment with a small vocabulary. We split the dataset into training, validation and testing as in [15].

The *Voxceleb2* dataset is a large-scale dataset consisting of in-the-wild recordings, which contains unconstrained natural vocabulary. We select only those videos that are in English, the predominant language in the dataset, in order to discard biases in the results due to the language distribution.

We corrupt the speech data with fullband temporal gaps of a duration between 160 and 1600 ms. During training, the corrupted segments are distributed randomly along each sample in a batch. During validation we apply the same logic so that the distribution of the validation set is as close as possible to the one of training. During testing, we run the system in 5 different setups: a random distribution of the gaps, as described before; corrupted segments with a gap of size 160 ms, 400 ms, 800 ms and 1600 ms. The *Grid* sentences typically include initial and trailing silence regions. When corrupting the speech signals, we ensure that the entire corrupted segment is located in the speech active parts of the *Grid* sentences (unlike [15], where corrupted segments were randomly chosen).

Baselines. We compare the proposed AVSI model against the previous state-of-the-art AV model, proposed in [15], and against the AO version of our model. In the AV baseline [15], the authors propose a framework whose core is a stack of three Bi-LSTM layers fed with an AV signal. As acoustic features, they use normalized log magnitude spectrograms, while the visual features are landmark-based motion vectors. In order to fuse the acoustic and the visual features via concatenation, they upsample the visual features to the sampling rate of the spectrogram. They minimize the mean squared error of the predicted log magnitude spectrogram with respect to the ground-truth one in the corrupted segment. Note that this is different from our setup, as we apply the loss on the whole predicted signal, not only in the corrupted segment.

To explore the benefits of using the visual modality, we also train our model in an AO setup (without visual information as input).

4. Results

We evaluate our model using three metrics: the *MAE* between the magnitude spectrogram and the ground-truth within the corrupted speech region; *STOI* [20], a speech intelligibility estimator; and *PESQ* [18], a speech quality estimator. *STOI* and *PESQ* scores lie between -1 and 1, and -0.5 and 4.5, respectively. While lower *MAE* scores corresponds to a lower reconstruction loss, for *PESQ* and *STOI*, the higher the better.

Since it is not possible to use *STOI* and *PESQ* for signals shorter than a few hundreds ms [20, 18], we cannot use them only on the corrupted part. Therefore, we compute the scores for the whole signal. This lowers the sensitivity of the metrics, especially when inpainting short segments.

Constrained Vocabulary Performance. As our goal is to develop a system capable of dealing with corrupted segments of any duration, rather than training a system specifically for each gap length, in Table 1, we report the overall performance of the model trained and evaluated in the *Grid Corpus* for a distribution of segment durations that matches that of the training stage. As it can be seen, the proposed AV model is not only better than its AO counterpart, but it also outperforms the previous state-of-the-art AV model [15].

Table 1. Performance scores averaged across the *Grid* test set. Corrupted segment lengths sampled from a uniform distribution. The symbol ↑ (↓) indicates higher (lower) the better.

	PESQ ↑	STOI ↑	MAE ↓
Corrupted input	1.78	0.58	0.43
Morrone et al. [15]	1.98	0.79	0.39
Proposed, audio-only	2.07	0.79	0.34
Proposed, audio-visual	**2.21**	**0.84**	**0.31**

Performance vs Segment Duration From Table 1, we can notice that the performance of the AV baseline, [15], is worse than the proposed AO model. As AO methods are good at inpainting short gaps, we carried out an analysis of the performance of each model against the corrupted segment duration. The results are shown in Fig. 2.

Considering the *MAE* values, we can see that they do not change significantly for segments larger than 800 ms. We hypothesize that the uncorrupted audio is used to determine the voice characteristics and the speech continuity in the boundary of the corrupted segment, while the rest is purely generated from the visuals. That is why the AO model works well for short gaps, where the missing information can be inferred from the audio context, and fails to inpaint larger gaps. Besides, the relative *MAE* between the reconstructed segments of 1600 ms and 160 ms (27% for the AO model and around 10% for the AV models) shows the effectiveness of the AV methods, as the *MAE* degradation of the AO model is much higher.

Analysing the results for each segment duration, the per-

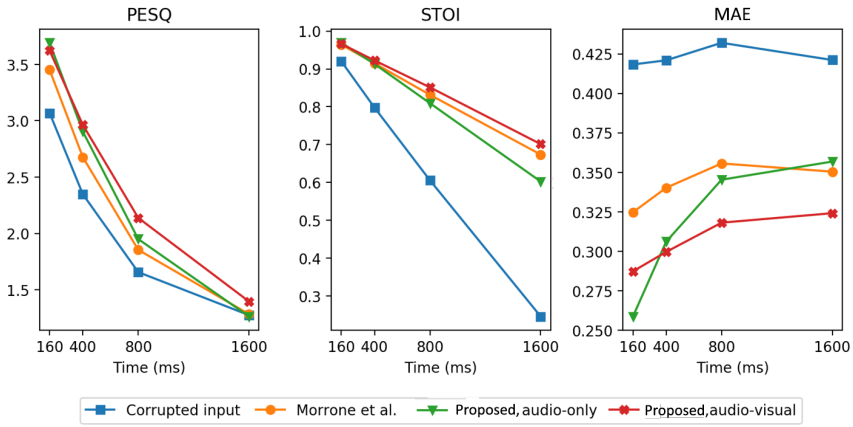

Figure 2. Comparison of performance vs corruption duration evaluated in the *Grid Corpus* test set (see Sec. 3).

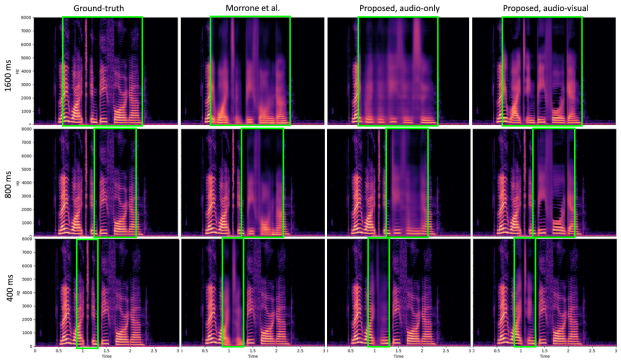

Figure 3. Sentence *lwib4a* for speaker 34 in the *Grid corpus* test set. Transcription: *"lay white in b four again"*. The region within the green square indicates the corrupted area. In practice, that region is set to zero as input to the network.

formance of the proposed AV and AO models, according to the estimated intelligibility (*STOI*), is roughly similar when considering corrupted segments of 160 ms.

On contrary, for corrupted segments of 400 ms, the proposed AV model is better at intelligibility and perceived quality. However, the AO model is still very effective. In Fig. 3 we can observe how the spectrogram predicted by the AO model is similar to that of the AV model, even the harmonics are better-defined than in the AV baseline.

For corrupted segments of 800 ms and 1600 ms, the proposed AV model is the best. For such long gaps, the AO model is no longer capable of estimating the content of the sentence. It just generates a kind of mumbling, either as a consequence of inpainting the sample with certain energy bands that match the harmonics of the voice or as an attempt to mimic sentences learned from the dataset. If we consider *PESQ*, we can see that, for segments of 1600 ms, the scores for the models tend to collapse to a single point. Our hypothesis is that, for such a long gap, the speech context is almost non-existent (see the first row of Fig. 3), therefore the task becomes close to speech reconstruction from silent videos [13], for which speech characteristics of unknown speakers, that are important for *PESQ*, cannot be easily estimated using only the video information.

Natural Vocabulary Performance. In the *Grid Corpus* the vocabulary is limited and unnatural. Since it is of our interest to study the performance of the model with real-world, in-the-wild data, in Table 2 we evaluate the model trained in the English subset of *Voxceleb2*. Both, the baseline model from [15] as well as the AO model did not converge when training in *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.

Table 2. Performance scores averaged across *Voxceleb2* test set. Corrupted segment lengths sampled from a uniform distribution. The symbol ↑ (↓) indicates higher (lower) the better.

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

Some demos of reconstructed audio signals (both in *Grid* and *Voxceleb2* test samples) are available at `https://ipcv.github.io/avsi/`.

5. Conclusions

This paper 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* [6] and showed that it outperforms its audio-only counterpart for gaps larger than 160 ms, and the previous state-of-the-art approach. In addition, we showed that the visual features extracted from the AV-HuBERT network encode enough information to guide the inpainting process. Besides, we showed our model can inpaint natural speech in in-the-wild scenarios (*Voxceleb2* dataset [5]).

References

[1] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley. Audio inpainting. *IEEE Trans. on Audio, Speech, and Language Processing*, 20(3):922–932, 2012.

[2] Z. Borsos, M. Sharifi, and M. Tagliasacchi. Speechpainter: Text-conditioned speech inpainting. *Interspeech*, 2022.

[3] Y.-L. Chang, K.-Y. Lee, P.-Y. Wu, H.-y. Lee, and W. Hsu. Deep long audio inpainting. *arXiv*, 2019.

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

[5] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, 2018.

[6] 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*, 120(5):2421–2424, 2006.

[7] P. P. Ebner and A. Eltelt. Audio inpainting with generative adversarial network. *arXiv preprint:2003.07704*, 2020.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] V. S. Kadandale, J. F. Montesinos, and G. Haro. Vocalist: An audio-visual synchronisation model for lips and voices. *Interspeech*, 2022.

[10] M. Kegler, P. Beckmann, and M. Cernak. Deep speech inpainting of time-frequency masks. *Interspeech*, 2020.

[11] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama. Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency. In *Proc. DAFx*, volume 10, pages 397–403, 2010.

[12] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak. A context encoder for audio inpainting. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 27(12):2362–2372, 2019.

[13] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 29:1368–1396, 2021.

[14] Juan F. Montesinos, Venkatesh S. Kadandale, and Gloria Haro. Vovit: Low latency graph-based audio-visual voice sseparation transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

[15] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen. Audio-visual speech inpainting with deep learning. In *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2021.

[16] A. Rahimi, T. Afouras, and A. Zisserman. Reading to listen at the cocktail party: Multi-modal speech separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10493–10502, 2022.

[17] I. Ramírez Paulino and I. Hounie. Paco and paco-dct: Patch consensus and its application to inpainting. In *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2001.

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

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.