Visual Self-Supervision by Facial Reconstruction for Speech Representation Learning

Abhinav Shukla
Imperial College London
a.shukla@imperial.ac.uk

Stavros Petridis
Imperial College London
stavros.petridis04@imperial.ac.uk

Maja Pantic
Imperial College London
m.pantic@imperial.ac.uk

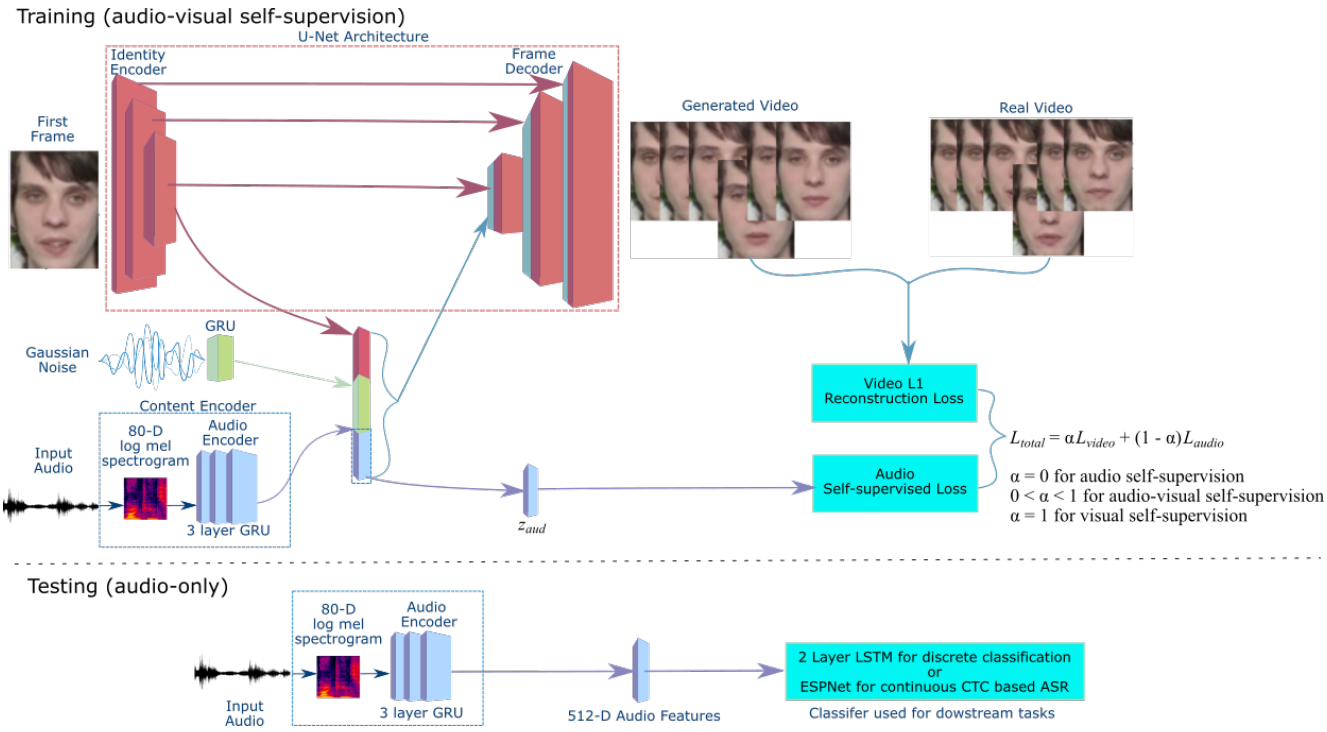

Figure 1: An overview of our proposed model for visually guided self-supervised audio representation learning. During training, we generate a video from a still face image and the corresponding audio and optimize the reconstruction loss. An optional audio self-supervised loss can be added to the total to enable multi-modal self-supervision. During testing, we use the audio encoder to extract features for (or finetune on) downstream audio-only tasks.

1. Introduction

Self-supervised learning has attracted plenty of recent research interest. However, most works are typically unimodal and there has been limited work that studies the interaction between audio and visual modalities for self-supervised learning. This work [1] (1) investigates visual self-supervision via face reconstruction to guide the learning of audio representations; (2) proposes two audio-only self-supervision approaches for speech representation learning; (3) shows that a multi-task combination of the proposed visual and audio self-supervision is beneficial for learning richer features that are more robust in noisy conditions; (4) shows that self-supervised pretraining leads to a superior weight initialization, which is especially useful to prevent overfitting and lead to faster model convergence on smaller sized datasets. We evaluate our audio representations for emotion and speech recognition, achieving state of the art performance for both problems. Our results demonstrate the potential of visual self-supervision for audio feature learning and suggest that joint visual and audio self-supervision leads to more informative speech representations.

In this work, we investigate self-supervised learning for audio. Audio representations are a cornerstone of speech and affect recognition. Self-supervised learning may offer better representations for these applications, especially in

[1] Work first published at ICASSP 2020, May 4 - 8, 2020, extended version submitted to IEEE Transactions on Affective Computing in May 2020, website: https://sites.google.com/view/visually-guided-speech

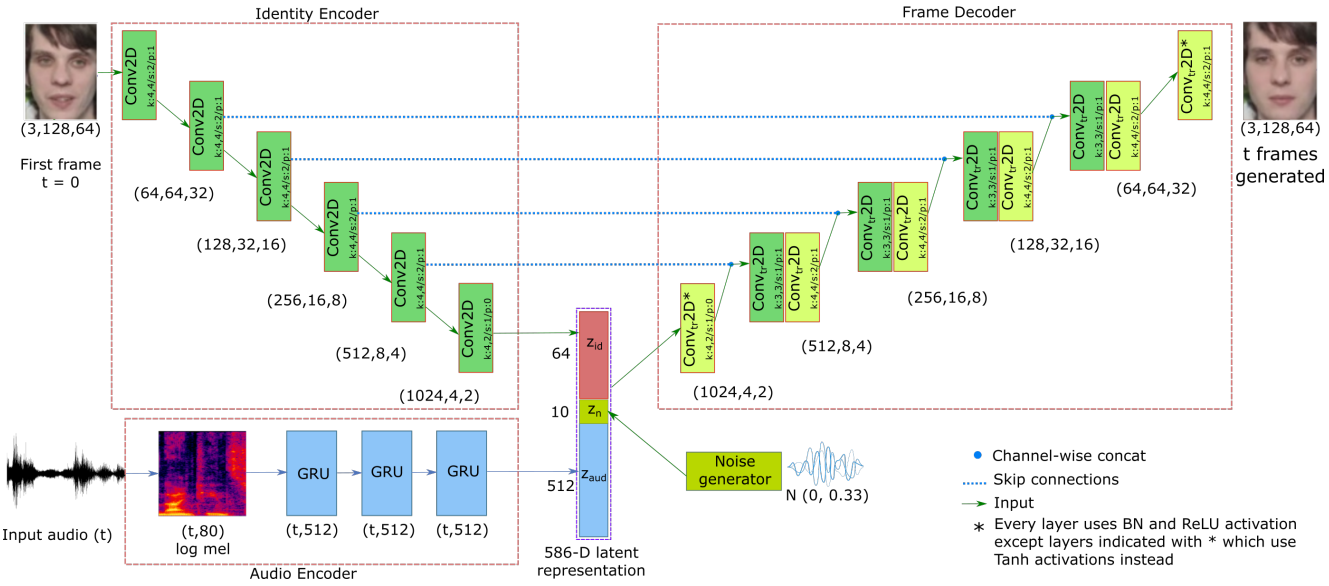

Figure 2: A detailed illustration of our the encoder-decoder model we use for video reconstruction. From an unlabeled sample of audiovisual speech, we use the audio and the first frame of the video ($t = 0$) to generate a video with t frames. The model contains: (1) an identity encoder which produces a 64-D identity embedding; (2) an audio encoder which converts the input audio (t frames of 80 dimensional log mel spectrograms) into a 512-D audio embedding; (3) a frame decoder which generates video from the concatenated latent representation using transposed convolutions.

cases where labeled data is hard to come by and unlabeled audio data is readily available. Most existing approaches are unimodal (audio-only). The few cross-modal approaches [4, 1] typically have some interaction between the modalities in the latent space by pretext tasks like clustering but they do not produce an intuitive interaction between the two modalities (especially in the context of audiovisual speech by using facial information). By contrast, our work proposes audio features that are explicitly guided by lip movements and facial expressions' reconstruction (see Fig. 1). We impicitly capture visual information related to lip movements and facial expressions in the audio features. The visual modality is needed only during training and our audio features can be evaluated on audio-only datasets.

2. Methods

2.1. Visual-only self-supervision via facial reconstruction (L1)

The proposed method is illustrated in Fig. 1 and is based on prior work on visually guided speech representation learning through speech-driven facial animation [9, 8, 7]. The model is a temporal encoder-decoder which takes a still image of a face (frame from a 25 fps video) and an audio signal as inputs and generates video frames from these. The model itself can be conceptually divided into three subnetworks (see Fig. 1 and Fig. 2), namely the content/audio encoder (3 layer GRU), the identity encoder (6 layer 2D CNN) and the frame decoder (with skip connections from the identity encoder).

The architecture of the content encoder is a 3 layer GRU with log mel spectrograms as input (closely following [2]), as shown in Fig. 2. The log mel spectrogram is computed with 80 frequency bins, a window width of 25ms and a stride of 10ms. It is converted into a latent representation with dimensionality (t, 512) z_{aud}. Similarly, the identity encoder (see Fig. 2 top-left), which is made of 6 (Conv2D - BatchNorm - ReLU) blocks, reduces a 64x128 input image (which is the first video frame of the audiovisual speech segment) to a 64x1 feature vector z_{id}.

We also use a noise generator (see Fig. 1) capable of producing noise that is temporally coherent. A 10 dimensional vector is sampled from a Gaussian distribution with mean 0 and variance of 0.33 and passed through a single-layer GRU to produce the noise sequence. This latent representation z_n accounts for randomness in the face synthesis process (such as the generation of random sequential behaviour like blinks [10]), which leads to a more realistic facial reconstruction.

The latent representation is the concatenation of $z_{aud}; z_{id}$ and z_n (as shown in Fig. 2). This results in a 586 dimensional embedding. This embedding then goes through the frame decoder (see Fig. 2 top-right), which is a CNN that uses strided transposed convolutions to produce the video frames. The skip connections to the identity encoder help in preserving subject identity.

An L1 reconstruction loss between a random frame from

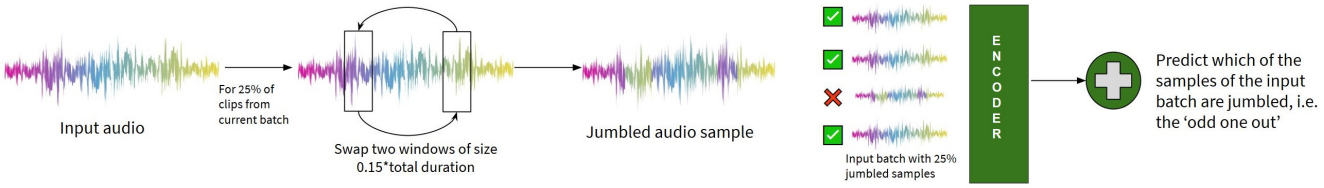

Figure 3: An overview of the proposed Odd One Out networks for audio representation learning. 25% of the input audio batch is jumbled. The audio encoder is then trained on the self supervised task of predicting which clip is the 'odd one out'.

the generated video and the corresponding frame from the real video is used to train the network. The L1 loss on the pixel level is commonly used in facial reconstruction as opposed to the L2 loss which typically produces blurrier reconstructions. We use the Adam optimizer with a learning rate of 0.06 that is decayed by a factor of 0.98 every 10 epochs. Essentially, our model aims to predict the video modality (face reconstruction) given only the audio modality and speaker identity information from the first frame. In this process, the audio encoder is driven to produce useful **speech features that correlate with mouth and facial movements** (because we need to generate these lip and facial movements using only the audio information, so the features z_{aud} must encode this in order to reduce the L1 loss). After this process of visually guided self-supervised pretraining, we simply use the trained audio encoder as a pretrained model for audio-only downstream tasks. The features extracted from this model are especially interesting to evaluate on tasks like speech recognition and emotion recognition. This is because these features are explicitly trained (guided by the visual modality) to contain information related to lip movements (highly correlated with speech) and facial expressions (highly correlated with emotion).

2.2. Audio-only self-supervision (Odd One Out)

Odd One Out networks for video [3] are based on predicting which one out of multiple sets of ordered sequences of frames is in jumbled order (temporally incorrect order). The intuition behind such a method being able to learn useful features is that while learning to predict the task, the encoder learns useful audio features that differentiate between certain phonemes. Being able to predict temporal order should drive the encoder to learn generic useful features about the data. We adapt this idea to the audio modality in a straightforward way as well. For a given input batch of audio clips, we jumble 25% of the clips. The jumbling is performed by selecting at random two windows of a length of 15% of the total audio duration and swapping them. The encoder is then tasked with predicting which element in the input batch is the 'Odd One Out', and is optimized using cross entropy loss. Fig. 3 illustrates the training procedure for Odd One Out networks for audio representation learning. We use the same audio encoder architecture as before (Figure 2).

2.3. Audio-visual self-supervision (L1 + Odd)

We combine the proposed audio and visual self-supervision methods by making the encoder jointly predict the visual self-supervision task and the audio self-supervision task. Since we used the same encoder architecture for both the visual and audio tasks, this is straightforward to accomplish. In the pipeline shown in Fig. 1 for visual self-supervision, we also use the optional prediction for the audio-only self-supervised task (Odd). This leads to two losses being calculated, one for visual and one for audio self-supervision. The total loss L_{total} is the weighted sum of the L1 reconstruction loss from visual self-supervision L_{video} and the cross entropy loss from the audio-only self supervision L_{audio}. is the weight factor which controls how much of the loss term comes from which type of supervision (optimal value is 0.6 [7]). The total loss is given by the equation:

$$L_{total} = \quad L_{video} + (1 \qquad) L_{audio} \tag{1}$$

3. Results and Conclusion

We evaluate all features on emotion and speech recognition. For a classification task (emotion recognition on CREMA, Ravdess, IEMOCAP or word classification on SPC), we use a 2 layer LSTM with 256 hidden units as the classifier. For ASR on GRID with continuous text labels, we use the ESPNet library with hybrid CTC/Attention. Our results can be seen in Table 1. Additionally, we also compare the L1, Odd and L1 + Odd methods under various levels of artificially introduced noise on two datasets. The results can be seen in Figures 4 and 5. Our results outperform existing self-supervised baselines on emotion recognition and speech recognition. We thus demonstrate the potential of visual self-supervision by facial reconstruction in audiovisual speech as a way to learn audio features. We also show that joint audio-visual self-supervision is better than either unimodal method. Additional details and a more detailed discussion can be found in [7].