A Generative Approach for Face Mask Removal Using Audio and Appearance

Luiz E. L. Coelho, Raphael Prates, William Robson Schwartz
Smart Sense Laboratory, Department of Computer Science, Universidade Federal de Minas Gerais, Brazil
luizduducoelho@ufmg.br, prates@dcc.ufmg.br, william@dcc.ufmg.br

***Abstract*—Since the COVID-19 pandemic, the use of facial masks in public spaces or during people gatherings has become common. Therefore, journalists, reporters, and interviewees frequently use a mask, following the public health measures to contain the pandemic. However, using a mask while speaking or conducting a presentation can be uncomfortable for viewers. Furthermore, the usage of a mask prevents lip reading, which can harm the speech comprehension of people with hearing impairment. Thus, this work aims at artificially removing masks in videos while recovering the lip movements using the audio and uncovered face features. We use the audio to infer the lip movement in a way it matches with the uttered phrase. From the audio, we estimate landmarks representing the mouth structure. Finally, the landmarks (i.e. uncovered and estimated) are the input in a generative adversarial network (GAN) that reconstructs the full face image with the mouth in a correct shape. We present quantitative results in the form of evaluation metrics and qualitative results in the form of visual examples.**

I. Introduction

Since the COVID-19 pandemic, the use of facial masks in public and crowded spaces has become very common and necessary. For instance, public health international agencies such as World Health Organization (WHO) advises the use of masks, as studies like Chu *et al.* [1] show that it significantly reduces the risk of infection from the disease. In this scenario, journalists, reporters and interviewees frequently must wear a mask to follow public health norms. Nonetheless, mask usage while speaking or giving a presentation can be uncomfortable for the viewers. Moreover, the presence of a facial mask blocks lip reading, which is a relevant issue as the lip reading is an important communication strategy for those with hearing impairment, and can assist the speech comprehension according to Dell'Aringa *et al*. [2]. As an example, Trecca *et al*. [3] showed that impaired patients had mild to severe communication difficulties when nursed by health personnel wearing a mask due to the impossibility of lip reading.

Some works perform mask removal from images, like Din *et al.* [4] that use a Generative Adversarial Network (GAN) based model to complete the full face image. Li *et al.* [5] propose a de-occlusion module to reconstruct the masked face aimed at face recognition. But both approaches are suited to single images and ignore the mouth movement during speaking.

Addressing the impossibility of lip reading, we aim at designing a method that processes a video in which a subject speaks while wearing a mask and uses the audio and uncovered facial features to reconstruct the whole face image as if there was no mask. The audio is used to infer the lip movement so that they are coherent with the sentence being spoken. We use landmarks as a representation of the mouth to be estimated from the audio. To make the reconstruction realistic, we use deep learning methods, in special GAN, with the landmarks guiding the lip movements.

To remove masks from videos, the problem is divided into three steps: mask segmentation; prediction of the mouth landmarks using the audio; and the face image reconstruction guided by the face landmarks. The segmentation step uses a color based approach, the mouth landmarks prediction uses a recurrent model and the face image reconstruction is achieved using an image-to-image GAN guided by the landmarks.

The major contributions of this work are: (1) the proposed method for mask removal in videos and (2) the use of more robust metrics for lip sync accuracy assessment. Regarding the first, there are similar works like [4] that remove masks from still images. Nonetheless, they do not leverage the audio information and can not be used to assist in lip reading. Moreover, the better evaluation of generated images is a seldom addressed issue in literature. In fact, works in face inpainting [6] and talking faces generation [7] employ simple mouth landmarks distance (MLD), which is very sensitive to small variations, or user study metrics, which can be subjective and hard to obtain.

We show experiments from both the mouth landmarks prediction step and face reconstruction step. The former achieved a value of 2.666 for the mean absolute error (MAE) of the mouth landmarks. The latter achieved the value of 0.4536 for the intersection over union (IoU) of the parsing of the mouth regions. We report further metrics for both models and show ablation studies analysing the impact of different implementation decisions.

II. Related Works

In this section we will briefly review some important works regarding generative adversarial networks followed by an overview of image reconstruction and then we will discuss works regarding face reconstruction specifically.

A. Generative Adversarial Networks

Since Goodfelow proposed the GAN [8], many improvements on the initial model came up, such as the Deep Convolutional Generative Adversarial Networks (DCGAN) [9] and conditional GAN [10]. And more elaborate models

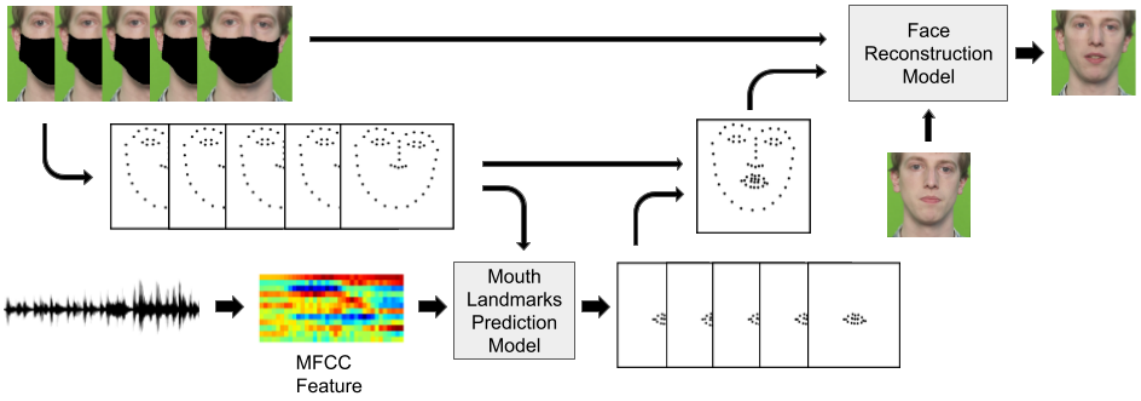

Fig. 1. Face reconstruction diagram. This diagram shows the process of face reconstruction of a video by using the audio information. First, we consider synchronized segments of audio and frames to extract the MFCC features and the facial landmarks (except for the mouth), respectively. These two inputs are arranged in temporal sequences and used to estimate the mouth landmarks with a recurrent model. Then, we employ the predicted facial landmarks to guide the mouth reconstruction with a generative model that also receives a reference image (the same for the whole video) to preserve the subject's appearance.

emerged, performing image-to-image translation as [11], [12], [13] where they use GANs to somehow modify an image.

B. Image Reconstruction

Image reconstruction or image inpainting are methods that attempt to complete missing regions in an image. An initial work [14] employed a traditional DCGAN to try to generate a similar image as the input, and then applied blending and superposition operations to complete the image, but this approach is limited. Most studies use image-to-image translation for image reconstruction, being able to tackle the problem more robustly. Iizuka *et al.* [15] proposed the use of both a global discriminator and a local discriminator for image completion. Yu *et al.* [16] have an architecture similar to [15], but using a two step approach, with a "coarse" generator and a "fine" generator to improve the completion quality.

C. Face Reconstruction

Some reconstruction methods have been specialized in face reconstruction. Yang *et al.* [17] present a face reconstruction method guided by facial landmarks. They propose a face landmark detector trained in degraded images, and a conditional DCGAN that uses the landmarks as a guide for the facial reconstruction. Koumparoulis *et al.* [6] use the audio information to reconstruct talking faces keeping the lip movement coherent. They process the degraded image and the audio Mel-Frequency Cepstral Coefficients (MFCC) features through separate encoders and then concatenate the embeddings and pass it through a decoder to recover the image dimensions. They also use a binary classification discriminator to assist the completion of the face.

In contrast with other existing methods, we leverage the audio to estimate the mouth landmarks in a video, and then use the estimated landmarks to model the mouth movements.

III. Methodology

For a given video of a subject talking while wearing a face mask, the desired output of the method is the same face region reconstructed as if the subject was not wearing the mask. Furthermore, the lip movement should match the uttered sentence, to enable lip reading. A diagram showing the overview of the method in shown in Figure 1. To accomplish our goal, we can split the task into three steps: *Mask Segmentation*, *Prediction of Mouth Landmarks* and *Face Reconstruction*. We present them in the following subsections.

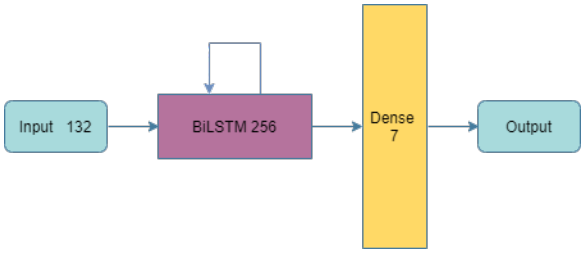

Fig. 2. Mouth Landmarks Prediction Model architecture. The numbers in each layer show the output dimension of the layer. Inputs and outputs are shown in blue, the bidirectional LSTM in purple and the dense layer in yellow.

A. Mask Segmentation

The mask segmentation step consists in finding the pixels corresponding to the mask in the image. To accomplish that, we employ a simple color-based approach that works as follows. First, we apply a face detector on the image to localize the face region and then a facial landmark detector to spotlight keypoints in the face. Then, we select landmarks that are positioned inside and outside (i.e. image border) the mask region. Finally, we consider these two sets of landmarks as different seeds in a watershed-based color segmentation algorithm that discriminates between mask and non-mask regions. To make the segmentation easier, we employ the YC_bC_r color space in this step, which separates the brightness (Y) component from the color components.

B. Prediction of Mouth Landmarks

In this section, we describe the proposed method for predicting the landmarks for the mask-covered mouth region.

It is an important stage as these landmarks guide the face reconstruction process and, therefore, must be in sync with the audio. To achieve that, we employ the audio and the uncovered facial landmarks as input in a recurrent architecture that models the temporal correlation between inputs and outputs. A good choice of architecture for this type of problems are recurrent neural networks such as Long Short-Term Memory (LSTM) [18]. An improvement on the vanilla LSTM is the bidirectional LSTM [19] that considers both directions in time (i.e. direct and reverse) as input.

The first step in the proposed mouth landmarks prediction consists in computing the audio and uncovered face features using a temporal sliding window. In this work, we represent the audio using the MFCC - a widely used feature in speech recognition and talking faces generation [20] - and the uncovered face features using the 48 landmarks outside the mouth region. Finally, we compute low-dimensional latent representations that decorrelates both features using Principal Component Analysis (PCA).

Specifically, we consider audio segments of 350ms, temporally centered according to each frame from the video. Then, we compute 12 MFCC coefficients in windows of 20ms with a 10ms overlap, resulting in an audio feature with dimension 12x35. Similarly, we compute the 96-dimensional representation using the both coordinates for landmarks in the corresponding video frame. Then, we obtain 127 and five PCA coefficients from MFCC and landmarks, respectively. These features are employed as input in the following model.

To predict the mouth landmarks, we devise a network with two layers where the first is the bidirectional LSTM with 128 units and the second is a fully connected layer with linear activation that maps to the seven PCA coefficients of the mouth landmarks. The model architecture is shown in Figure 2. The model operates on temporal sequences of n sequential frames in a video. The fully connected layer shares the same weights for all the positions in the sequence. We adopt a stride of one frame for the sequences in a video. The loss function used is the MSE for the mouth landmarks PCA coefficients. Finally, we obtain the mouth landmarks using the PCA inverse transformation of the predicted PCA coefficients.

C. Face Reconstruction

The Face Reconstruction model receives an image with the mask region extracted along with the face landmarks and a reference image to produce the output of a reconstructed face with no mask. The mouth generation is guided by the landmarks, since we want to emulate the lips movement. Therefore, the quality of the reconstructed mouth is dependent on the quality of the estimated mouth landmarks, in a way that they represent a good approximation of the lips movement while speaking.

1) Architecture: The architecture used is the conditional DCGAN proposed by [17]. The main difference is that Yang *et al.* [17] train a landmark detector for degraded images, while we use a recurrent model to estimate the landmarks of a sequence of frames using the audio. Besides, we propose a new input to the model, the reference image, an image with the subject without mask in order to help the model to reconstruct the mask region while preserving the facial appearance. We use the procrustes [21] transformation in the facial landmarks (excluding the ones from the mouth) to align the reference image to the image to be reconstructed. The generator architecture is based on the U-Net [22], which was designed to image to image translation networks. The network is made of three downsampling blocks followed by seven residual Resnet [23] blocks, followed by upsampling blocks, to recover the original dimensions of the image. The discriminator has a PatchGan [11] architecture with five convolutional layers. The final image is obtained by replacing the pixels from the mask region for the equivalent pixels in the image generated by the face reconstruction model. Doing so, pixels outside the mask region are preserved.

2) Training: To train the model, besides the adversarial losses we employ four other loss functions, defined in Equations 1, 2, 3 and 4.

$$L_{pixel} := \frac{1}{N_m}||\hat{I} - I||_1 \tag{1}$$

The pixel difference loss, defined in Equation 1, is a straightforward similarity metric between two images. Here, $\| \cdot \|_1$ is the *L1-norm*. I is the ground truth image, $\hat{I}$ is the output reconstructed image. N_m is the mask region size in pixels. The loss is adjusted by the mask size because we expect a smaller occlusion to be easier to reconstruct.

$$L_{perc} := \sum_p \frac{||\phi_p(\hat{I}) - \phi_p(I)||_1}{N_p H_p W_p} \tag{2}$$

The perceptual loss [24], defined in Equation 2, compares feature maps extracted from pre-trained deep neural networks. We used VGG-19 [25] trained in the ImageNet [26] dataset. $\phi(\cdot)$ are the N_p feature maps from layer p. H_p and W_p are respectively the height and width of the feature maps from layer p.

$$L_{style} := \sum_p \frac{1}{N_p N_p}||\frac{G_p(\hat{I} \circ M) - G_p(I \circ M)}{N_p H_p W_p}||_1 \tag{3}$$

The style loss function [27], defined in Equation 3, encourages the texture of output and ground images to be similar. As well as the perceptual loss, the style loss compares intermediate activations from a pre-trained network. We also use VGG-19 here. M is the mask region, $\circ$ the Hadamard product and G_p the Gram matrix for the feature maps. The Gram matrix is defined as: $G_p(x) = \phi_p(x)^T \phi_p(x)$.

$$L_{tv} := \frac{1}{N_I}||\nabla \hat{I}||_1 \tag{4}$$

The total variation loss [28], defined in Equation 4, is based on the homonym regularization, used for image denoising. The loss helps to smooth the checkerboard patterns in images, that can appear in convolutional networks generated images. N_I

is the number of pixels in image I and ∇ is the first order derivative, containing the vertical and horizontal derivatives.

$$L_{advG} := E[(D(G_P(I^M, L), L_{gt}) - 1)^2] \tag{5}$$

$$L_{advD} := E[D(\hat{I}, L_{gt})^2] + E[(D(I, L_{gt}) - 1)^2] \tag{6}$$

The adversarial loss functions in Equations 5 e 6 are proposed in LSGAN [29]. They showed better stability during training and better visual quality for the generated images. L are the landmarks estimated by the *mouth landmarks prediction* model, L_{gt} are the ground truth landmarks, extracted by the detector. $G_P(\cdot)$ is the generator output and $D(\cdot)$ the discriminator output. We define that $I^M := I \circ M$ and $E[\cdot]$ is the expected value operator.

According to LSGAN [29], during the training we want to minimize the discriminator loss function, L_{advD} and the generator loss function, that is a weighted sum as:

$$\begin{aligned} L_G := L_{pixel} + \lambda_{perc} L_{perc} + \lambda_{style} L_{style} + \\ \lambda_{tv} L_{tv} + \lambda_{adv} L_{advG} \end{aligned} \tag{7}$$

The total loss function for the generator is presented in Equation 7. The values used are $\lambda_{perc} = 0.1$, $\lambda_{style} = 250$, $\lambda_{tv} = 0.1$ and $\lambda_{adv} = 0.01$, as suggested in [17].

3) Mask Projection: To train the face reconstruction model, we need videos of subjects uttering a phrase while wearing a mask and the same videos without the mask, to be the ground truth. To obtain this data and enable the training of the model, we can project masks in the videos of a dataset. The mask projection applies the region from a segmented mask in another image. First we need to segment a mask from some source image, like in samples of a masked people dataset [30]. Then we extract the facial landmarks (except the ones from mouth) from both the source image and the target image in which we want to project the mask.

We use the procrustes [21] transformation in the landmarks to align the source image with the target image, since the procrustes finds the translation, rotation and scale factor that best fit a set of points (such as landmarks). We do not use the mouth landmarks because they do not influence the alignment of the mask to the face, and could introduce noise in the superposition. With the mask projected in the target images, they can be used as input to the model during training.

IV. Evaluation Metrics

In this section, we present the metrics employed to evaluate the obtained experimental results for the landmarks prediction (Section IV-A) and the face reconstruction (Section IV-B).

Works in talking face inpainting like [6] usually adopt Mouth Landmarks Distance (MLD) as a metric for assessing the quality of the mouth movements. But using the mouth landmarks distance alone is not ideal, because it is very sensitive to translations and can be very noisy. Therefore, we propose two novel metrics, disparity and parsing IoU (PIoU). Since disparity is robust in relation to translation, rotation and scale factor, it is better suited to assess the mouth shape and relative positions of the landmarks. On the other hand, parsing IoU captures the mouth regions instead of only the landmarks, which is a richer information and a better way to evaluate position and scale factor of the mouth. In that sense, disparity and IoU parsing work well together.

A. Metrics to Evaluate Mouth Landmarks

In the following paragraphs, we describe the different metrics used in the *mouth landmarks prediction model* (MLPM). For each frame in the test partition videos, the model estimates the mouth landmarks. The predicted landmarks can be compared with the ground truth landmarks, which were obtained from the landmark detector.

In this work, we define $P_q^{(t)}$ as the mouth landmark *q*, predicted from the frame *t*. On the other hand $G_q^{(t)}$ stands for the ground truth landmark *q*, from the frame *t*. Furthermore, we set N_{flm} as the number of mouth landmarks in each face, which is 20 for the landmark detector used in the experiments, and N_t as the total number of frames in the test dataset.

1) Mean Absolute Error: A simple way to measure their proximity is to calculate the MAE as:

$$MAE = \frac{1}{N_t} \frac{1}{N_{flm}} \sum_{t=1}^{N_t} \sum_{q=1}^{N_{flm}} |P_q^{(t)} - G_q^{(t)}| \tag{8}$$

2) Procrustes projection residual error (Disparity): To calculate this metric, we apply the procrustes transformation between the predicted facial landmarks of a given frame and its respective ground truth landmarks. The disparity measures the error of the transformation. It is the mean quadratic pointwise difference of the fiducial landmarks from the test set. The benefit of using this metric is due to the robustness to translation and scale factor. MAE is very sensitive to such variations. However, disparity, defined in Equation 9, is a good alternative to assessing similarity in the shape and arrangement of the facial landmarks, since it leverages the procrustes transformation.

$$Disparity = \frac{1}{N_t} \sum_{t=1}^{N_t} \sum_{q=1}^{N_{flm}} (PP_q^{(t)} - G_q^{(t)})^2, \tag{9}$$

where $PP_q^{(t)}$ stands for the mouth landmark *q*, predicted for the frame *t*, after the procrustes transformation.

B. Metrics to Evaluate Face Reconstruction

We define the following metrics for the face reconstruction model evaluation.

1) Face Parsing IoU: Face parsing is the segmentation of the face parts, such as eyes, nose, mouth and even accessories such as glasses and hat. With this information one can use the mouth region to estimate the quality of a reconstructed face based on how similar it is to the ground truth face. We used an implementation of face parsing based on the architecture of Yu et al. [31]. For the metric, we consider only three of the segmented region, all derived from the mouth. The regions are upper lip, interior mouth and lower lip. The parsing is extracted for both the generated image and ground truth image and the IoU is computed for the three regions. The final metric is the mean IoU for the three regions from all the test set frames.

$$ParsingIoU = \frac{1}{N_t}\frac{1}{N_r}\sum_{t=1}^{N_t}\sum_{r=1}^{N_r}\frac{BP_r^{(t)} \cap BG_r^{(t)}}{BP_r^{(t)} \cup BG_r^{(t)}} \qquad (10)$$

Equation 10 is the definition of the parsing IoU metric. BP is the parsing of the reconstructed image, BG is the parsing of the ground truth image, r is the parsing region in the frame t and N_r is the number of parsing regions used, which is equal to three in this case.

2) Mouth Landmarks Distance: This metric is a simple method of measuring how similar the reconstructed mouth is with the ground truth image. The metric is the mean distance from the mouth landmarks of both images.

$$MLD = \frac{1}{N_t}\frac{1}{N_{flm}}\sum_{t=1}^{N_t}\sum_{q=1}^{N_{flm}}|IP_q^{(t)} - IG_q^{(t)}|, \qquad (11)$$

in Equation 11, IP are the mouth landmarks from the generated image and IG the landmarks of the ground truth image.

3) Disparity: As in the method of Section IV-A, we extract the facial landmarks from the images and then apply the disparity metric.

$$Disparity = \frac{1}{N_t}\sum_{t=1}^{N_t}\sum_{q=1}^{N_{flm}}(IPP_q^{(t)} - IG_q^{(t)})^2, \qquad (12)$$

Equation 12 shows the metric definition. IPP are the mouth landmarks from the generated image, after being projected by the procrustes transformation.

4) Structural Similarity Index Measure: SSIM [32] is a well known metric of similarity between two images. We use it as a way of assessing the quality of the generated image as a whole, instead of just looking at the mouth region. It considers the structural information of the image and the inter-dependency of adjacent pixels. The metric is a number between zero and one. Values near one mean a high similarity between the images.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \qquad (13)$$

in Equation 13, x and y are windows of size 11x11 of the image generated by the *face reconstruction* and ground truth image, respectively. Also μ_x is the average of x, μ_y is the average of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy} is the covariance of x and y, C_1 is equal to $(k_1 L)^2$, C_2 is equal to $(k_2 L)^2$, L is the dynamic range of the pixel values, $k1 = 0.01$ and $k2 = 0.03$.

5) Peak Signal-to-Noise Ratio: PSNR is also a metric of similarity between two images. But it takes in consideration the absolute difference between pixels.

$$PSNR = \frac{1}{N_t}\sum_{t=1}^{N_t} 10\log_{10}(\frac{MAX^2}{MSE}), \qquad (14)$$

Equation 14 shows the PSNR definition in which MAX is the maximum possible value of the pixels (255 in this case). MSE is the mean squared error between the two images.

6) L1 Distance: L1 is a straightforward distance measurement that can be applied in images. Its equation is shown in 15.

$$L1 = \frac{1}{N_t}\frac{1}{N_p}\sum_{t=1}^{N_t}\sum_{p=1}^{N_p}|VP_p^{(t)} - VG_p^{(t)}|, \qquad (15)$$

where VP_p is the pixel p of image VP generated by the *face reconstruction model*. VG_p is the pixel p of the ground truth image. N_p is the total count of pixels in the mask region.

V. Experimental Results

In this section, we evaluate both the *mouth landmarks prediction model* and the *face reconstruction model*. We also show how much the error of the former influences the performance of the latter. Ablation experiments show the impact of different architecture aspects in the models. For the experiments, the face detector employed is the one available in Dlib toolkit [33], that is based on the Histogram of Oriented Gradients (HOG) [34] combined with a linear classifier and a sliding window detection scheme. The landmark detector used was also from Dlib tookit, with an implementation based on the work of Sullivan and Kazemi [35]. The segmentation used was the OpenCV implementation of watershed [36] algorithm.

A. Dataset

The dataset used in the experimental evaluation is TCD-TIMIT [37], which is widely used in the field of talking face generation [7] and talking face image inpainting [6]. It was designed to address automatic speech recognition using audio-visual approaches. It is composed of 62 speakers saying a total of 6913 short sentences in English. The videos are high quality, with 1920x1080 pixels, and audio frequency of 48KHz. The videos were recorded simultaneously in a frontal view and in side view of 30º, but only the frontal videos were used in our experiments.

B. Training Protocol

The dataset TCD-TIMIT suggests an experimental protocol for partitioning training and test sets by subject. The same protocol is used by [6] and many related works. The partition

TABLE I
COMPARISON ON THE EFFECT OF SEQUENCE SIZE IN THE EVALUATION METRICS.

Model Name	Metrics	
	MAE↓	Disp $*10^2$ ↓
MLPM seq=3	2.689	1.589
MLPM seq=7	2.677	1.566
MLPM seq=11	**2.666**	**1.533**
MLPM seq=15	2.667	1.557
MLPM seq=19	2.667	1.544

TABLE II
COMPARING METRICS WITH A TWO LAYERS FULLY CONNECTED BASELINE AND A UNIDIRECTIONAL AND BIDERECTIONAL LSTM.

Model Name	Metrics	
	MAE↓	Disp $*10^2$ ↓
Bi-LSTM	**2.666**	**1.533**
LSTM	2.724	1.650
Fully connected	2.772	1.616

splits 39 subjects for training and 17 for test, without subject overlap between partitions.

The aforementioned division was used for the training and evaluation of both *mouth landmarks prediction model* and *face reconstruction model*.

C. Mouth Landmarks Experimental Results

In this section, we present the experimental results on the MLPM using the metrics defined previously. We show an experiment evaluating the effect of the sequence size in the metrics, one evaluating the recurrent model used and an ablation study with the model inputs.

1) Sequence Size: The recurrent models used have the sequence size as a hyperparameter. Table I shows the results regarding the sequence size. A range with the following values was chosen: 3, 7, 11, 15 e 19. The experiments show that the sequence size has a minor impact in the metrics, with small variations. The chosen size was 11, which presented the best metrics. This size was fixed for the next experiments.

2) Recurrent Model: Table II shows the evaluation of the recurrent model. We compare it with a simpler architecture, a two layers fully connected network, with the hidden layer of the same size as the output of the LSTM, 128. We also compare the same architecture, replacing the bidirectional LSTM with a unidirectional one. According to the results, the bidirectional recurrent model performs better than the fully connected and the simple LSTM, which is evidence that the data is temporally dependent in both directions. Therefore, frames that are near in time help in the estimation of the current mouth landmarks.

TABLE III
METRICS OF ORIGINAL MODEL AND OF THE MODEL USING ONLY ONE INPUT AT A TIME.

Model Name	Metrics	
	MAE↓	Disp $*10^2$ ↓
MLPM	**2.666**	**1.533**
MLPM w/o landmarks	6.191	1.692
MLPM w/o audio	2.831	2.146

TABLE IV
METRICS OF THE FACE GENERATIVE MODEL WITH GROUND TRUTH MOUTH LANDMARKS AND AUDIO ESTIMATED MOUTH LANDMARKS.

Model Name	Metrics					
	PIoU↑	Disp ↓	MLD↓	SSIM↑	PSNR↑	L1↓
GTFL	**0.7592**	**0.00351**	**1.39**	**0.9289**	**31.18**	**14.01**
EFL	0.4536	0.01147	10.99	0.9102	29.77	16.12

TABLE V
ABLATION STUDY WITH MODEL LOSSES.

Model Name	Metrics					
	PIoU↑	Disp ↓	MLD↓	SSIM↑	PSNR↑	L1↓
EFL	**0.4536**	**0.01147**	**10.99**	**0.9102**	**29.77**	**16.12**
EFL w/o Discriminator	0.4499	0.01248	11.51	0.9129	29.67	16.34
EFL w/o Style Loss	0.4495	0.01258	11.82	0.9159	29.65	16.61
EFL w/o Perceptual Loss	0.4503	0.01261	11.90	0.9122	29.71	**16.12**
EFL w/o TV Loss	0.4508	0.01231	11.75	0.9118	28.69	17.74

3) Ablation Study: The audio to sequence model receives two inputs, the facial landmarks other than the mouth and the audio MFCC features. To assess the influence of each of them, the model was trained with only one of the inputs to obtain the metrics. According to Table III, the model trained only with the audio information had a high MAE but a reasonable disparity, which indicates that the format and arrangement of the mouth landmarks were close to the expected but the coordinates were far from the ground truth. This occurs because without the remaining facial landmarks it is hard to position and scale the mouth landmarks accordingly.

Using only the facial landmarks as input, we see the opposite behaviour. The MAE was reasonable, meaning that the predicted mouth landmarks coordinates were close to the ground truth. But the disparity was high, showing difficulty in adjusting the relative position of the landmarks and its movements. This suggests that both inputs are complementary, since they provide different information to the model.

D. Face Reconstruction Experimental Results

In this section, we present the experimental results regarding the face reconstruction generative model in a quantitative manner, using the metrics defined and in a qualitative manner through the example figures. We use the ground truth mouth landmarks for training and the mouth landmarks predicted by the MLPM for testing.

1) Quantitative Results: Table IV, shows metrics comparing the model trained with the ground truth facial landmarks (GTFL) as input, with the model trained with the estimated facial landmarks (EFL) obtained from the MLPM. While the first three metrics, Parsing IoU, Disparity and Mouth Landarks Distance evaluate the quality of the generated mouth, the

TABLE VI
MODELS TRAINED WITH AND WITHOUT THE REFERENCE IMAGE AS AN INPUT.

Model Name	Metrics					
	PIoU↑	Disp ↓	MLD↓	SSIM↑	PSNR↑	L1↓
EFL	**0.4536**	**0.01147**	**10.99**	**0.9102**	**29.77**	**16.12**
EFL w/o ref image	0.4348	0.01438	14.96	0.8927	28.06	20.53

other metrics, SSIM, PSNR and L1 evaluate the image as a whole. According to results, the GTFL model shows mouth reconstruction metrics far superior than the EFL model, which hints that the face generative model could benefit from more accurate mouth landmarks and shape estimators. It is still an open problem for research.

Table V shows the ablation study regarding the loss functions used during training. Among the studied configurations, we can see that the best performance is the model with all the losses. There is a slight drop in performance generally speaking, when removing one of the loss functions. Nevertheless, the absence of one of them does not seem to significantly harm the model quality, showing some robustness with respect to it.

One of the EFL inputs is the reference image, but the model can also be trained without it. The compromise is that without the reference image, the model reconstructs the image face generically, while the reference image provides prior information about the subject face. In Table VI we compare the model trained with and without the reference image (using it accordingly in the test set). From the metrics, we can see that both the mouth quality and image quality metrics drop when removing the reference image. The MLD increases from 10.99 to 14.96 while parsing IoU only drops from 0.4526 to 0.4348. We can see that classical metrics such as landmark distance suffer more without the mouth appearance information. Moreover, there are scenarios where it is not practicable to collect a reference image from the subject, hence the importance of having a model trained without this kind of input.

2) Qualitative Results: Figure 3 shows a visual result from the method for some frames of a random subject from the test partition. Each column refers to a different frame. The first two rows are the model inputs, the reference image and the image to be reconstructed guided by the facial landmarks. Third row shows the result of the face reconstruction. And the fourth row shows the ground truth image. According to the figures, the mouth shape resembles the ground truth image, even though they are not exactly identical. Since the model has no previous information regarding the subject teeth (usually not visible in the reference image), it tries to fill in the space generically.

Figure 4 compares the model trained with a reference image and the model trained without the reference image. First row shows the reference image, second row the input image to be reconstructed, third row the image reconstructed by the model using the reference image, fourth row the image reconstructed by the model without the reference image and finally the fifth row exhibits the ground truth image. According to the images, one can see that without the reference image some facial features are reconstructed in a general way, which alters the subject appearance. Since in Figure 4, the generated image without the reference changed the nose and beard from the subject, which did not happen to the reconstructed image using the reference.

Fig. 3. Model results. Column index the example frame and rows the kind of image. The row represent, respectively, the reference image, input image with masked projected, image generated by the face generative model and ground truth. This subject belongs to the test partition.

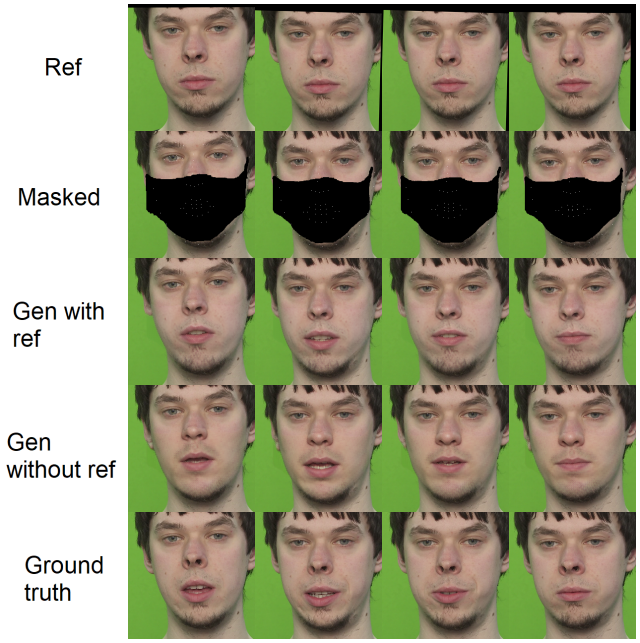

Fig. 4. Comparing the results with the generative model trained with and without the reference image as an input. The rows represent, respectively, the reference image, input image with masked projected, image generated by the face generative model with the reference image, image generated by the model without the reference image and ground truth. This subject belongs to the test partition.

VI. Conclusions

In this paper we proposed a method to remove facial masks in videos of a talking subject. The method explores the temporal correlation of the audio and mouth movements, describing the mouth shape with landmarks that guide the *face reconstruction model* to generate a coherent mouth. By doing so, we intend to enable lip reading. We adopted metrics used in the literature and proposed others more robust. Using them, we explored different aspects of the models' architecture and how they affect the results. The mapping between the audio domain and the landmark coordinates domain showed up as a hard task to obtain precise predictions. A direction for future work is to develop models that improve the accuracy regarding the *mouth landmarks prediction model*.

Acknowledgments

The authors would like to thank the National Council for Scientific and Technological Development – CNPq (Grants 438629/2018-3 and 309953/2019-7), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We also thank the support by the Coordination for the Improvement of Higher Education Personnel (CAPES) under Grant 88887.516264/2020-00 from the Public Security and Forensic Sciences (PROCAD) Notice.

References

[1] D. K. Chu, E. A. Akl, S. Duda, K. Solo, S. Yaacoub, H. J. Schünemann, A. El-Harakeh, A. Bognanni, and et al., "Physical distancing, face masks, and eye protection to prevent person-to-person transmission of sars-cov-2 and covid-19: a systematic review and meta-analysis," *The Lancet*, vol. 395, no. 10242, p. 1973–1987, 2020.

[2] D. A. Dell'Aringa AH, Adachi ES, "Lip reading role in the hearing aid fitting process," *Brazilian Journal of Otorhinolaryngology*, vol. 73, pp. 95–99, 2007.

[3] C. M. Trecca EMC, Gelardi M, "Covid-19 and hearing difficulties," *American Journal of Otolaryngology*, vol. 41, p. 102496, 2020.

[4] N. Ud Din, K. Javed, S. Bae, and J. Yi, "A novel gan-based network for unmasking of masked face," *IEEE Access*, vol. 8, pp. 44 276–44 287, 2020.

[5] C. Li, S. Ge, D. Zhang, and J. Li, "Look through masks: Towards masked face recognition with de-occlusion distillation," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[6] A. Koumparoulis, G. Potamianos, S. Thomas, and E. da Silva Morais, "Audio-assisted image inpainting for talking faces," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7664–7668.

[7] Y. Song, J. Zhu, D. Li, X. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," 2019.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, 06 2014.

[9] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015.

[10] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," 07 2017, pp. 5967–5976.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[13] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," 2018.

[14] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, Jul. 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073659

[16] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling, "Lafin: Generative landmark guided face inpainting," 2019.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[20] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" in *British Machine Vision Conference*, 2017.

[21] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," vol. 9351, 10 2015, pp. 234–241.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[24] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 694–711.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[27] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[28] M. Javanmardi, M. Sajjadi, T. Liu, and T. Tasdizen, "Unsupervised total variation loss for semi-supervised deep learning of semantic segmentation," *arXiv preprint arXiv:1605.01368*, 2016.

[29] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[30] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, H. Chen, Y. Miao, Z. Huang, and J. Liang, "Masked face recognition dataset and application," 2020.

[31] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[32] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[33] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, p. 1755–1758, Dec. 2009.

[34] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 886–893 vol. 1.

[35] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[36] S. Beucher and F. Meyer, "Segmentation: The watershed transformation. mathematical morphology in image processing," *Optical Engineering*, vol. 34, pp. 433–481, 01 1993.

[37] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.