Realistic Lip Animation from Speech for Unseen Subjects using Few-shot Cross-modal Learning

1st Swapna Agarwal
Research & Innovation
Tata Consultancy Services
Kolkata, India
agarwal.swapna@tcs.com

2nd Dipanjan Das
Research & Innovation
Tata Consultancy Services
Kolkata, India
dipanjan.da@tcs.com

3rd Brojeshwar Bhowmick
Research & Innovation
Tata Consultancy Services
Kolkata, India
b.bhowmick@tcs.com

***Abstract*—Recent advances in Convolutional Neural Network (CNN) based approaches have been able to generate convincing talking heads. Personalization of such talking heads requires training of the model with a large number of examples of the target person. This is also time consuming. In this paper, we propose a meta-learning based few-shot approach for generating personalized 2D talking heads where the lip animation is driven by a given audio. The idea is that the model is meta-trained with a dataset consisting of a large variety of subjects' ethnicity and vocabulary. We show that our meta-trained model is then capable of generating realistic animation for previously unseen face and unseen audio when finetuned with only a few-shot examples for a very short time (180 seconds). Considering the fact that facial expressions driven by audio are mainly expressed through motion around lips, we restrict ourselves to animating lip only. We have done the experiments on two publicly available datasets: GRID and TCD-TIMIT and our own captured data of Asian people. Both qualitative and quantitative analysis show that animations generated by such meta-learned model surpasses the state-of-the-art methods both in terms of realism and time taken.**

***Index Terms*—MAML, lip animation, meta-learning**

I. Introduction

Talking face model, which consists of speech-driven human facial animation lends itself to a variety of applications involving human-computer interaction like gaming, telepresence etc. In telepresence applications with a constraint network bandwidth we can render a talking $2D$ face at the receiver end by sending only the speech signal. The current State-Of-the-Art (SOA) methods [2], [2], [3], [7], [10], [17] try to model speech-driven talking heads by one-time training of deep neural network with a huge training dataset comprising of multiple subjects. With such trainings it is difficult to produce a faithful rendering of lip movement on a target unseen face image. On the other hand, there are methods that train the network by a huge set of examples of the target face only [7], [14]. Training such models is time consuming and requires a huge training set of each of the target faces.

In this paper, we address these issues by proposing a meta learning based approach. In this approach, a global model is first trained (meta-learned) using a wide variety of data consisting of multiple faces. For adapting this global model to an unseen face, we finetune the global model with a handful of examples of the target face. Within a short time as less as 180 seconds of finetuning, our model can generate realistic animation that beats the existing SOA both quantitatively and qualitatively.

Since face animation when driven by audio is mainly perceived by lip movement, we animate only lip rather than the whole face to keep things simple. One could also have used transfer learning for this purpose. However, transfer learning requires a large training set of the unseen subject and several hours of training which may be infeasible. Inspired by the work proposed by [1], [5], [9], [13]. The main contributions of this paper are:

1) To the best of our knowledge, we are the first one to apply MAML in cross-modality problem where the facial animation is driven by audio. Such an approach is more principled and produces better rendering on arbitrary faces after being finetuned for a very short time (180 seconds) using only a few examples of the unseen subject.
2) Our method preserves intrinsic subject features such as skin color, lip shape etc and retains the sharpness. Thus the rendering is more realistic compared to present SOA.

II. Related Work

The SOA methods for audio-driven 2D talking face animation can be broadly classfied into multiple ways: those which train a personalized model for each individual face [12], [14] vs. those which use a pre-trained generalized model to animate any unseen face [2], those which animate only the lip [2] vs those which animate the whole face [15]. Some train the model as one end-to-end network [2] while some others divide into multiple parts [14]. We present a brief study of these methods and place our work with respect to them.

a) Personalized models: Suwajanakorn *et al.* [14] , use recurrent neural network to predict the mouth shape for every frame given the audio and use a separate network which generates the texture from the produced mouth shape. In a recent work, Prajwal *et al.* [12] have proposed a pipeline to automatically dub a face video from one language to another with lip synchronization. For reslistic lip shape generation, in addition to reconstruction loss, they use L_2 distance beteen input face and audio embeddings in a GAN based framewrok. The major limitation of these methods is the lack of generalization as these methods require to re-train the entire model with

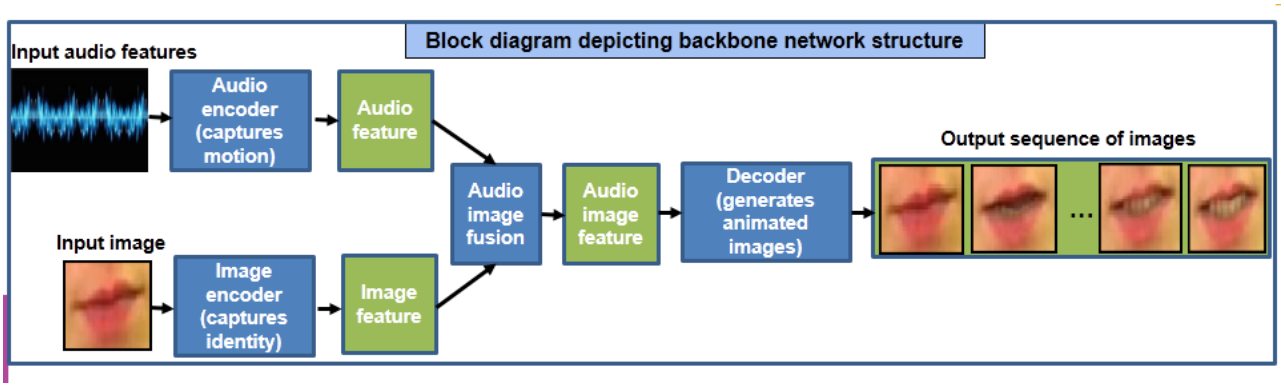


Fig. 1. Block diagram depicting the proposed network structure.

hours/minutes of video for any other person. On the contrary, our work requires only a few examples of the new target face for finetuning the meta-trained network.

b) Generalized models: There are several works [2], [3], [17] which claim to produce realistic animation of the unseen face without retrainining the main model. The model by Chung *et al.* [3] can be applied to an arbitrary subject, but the generated face sequences are sometimes blurry and not temporally meaningful. Sometimes the generated lips still move even when there is no utterance. Chen *et al.* [2] proposed a new correlation loss which helps in producing temporally more coherent result. Still they are not able to preserve the intrinsic subject traits e.g., lip shape, skin color for the unseen subject. Zhou *et al.* [17] disentangle the subject and speech-related information from input video. The major issue of this method is considerable zoom-in and zoom-out effect in the generated output. It requires post processing like subspace video stabilization to achieve smoothness. While most of the SOA methods [12] produce animation around the lips only, Vougioukas et al. [15] claim to produce realistic facial animation with movement in upper face (e.g., eye blink). When tested on random new faces, they produce blurred output where facial identity is compromised.

These SOA methods are not able to preserve the intrinsic traits of the unseen subject well. This is because limited number of subjects are available during training and it is a difficult problem to predict intrinsic subject traits without observing some samples of that subject. To overcome this problem we use the concept of Model Agnostic Meta Learning (MAML) proposed by [5]. While Zakharov et. al. [16] have also used meta-learning for training talking-head model, the main disadvantage of their work is the requirement of having sets of landmark points that drive the animation. Meta-learning helps our model for fast adaptation, by learning internal representations which are more transferable to the unknown subject using a very few samples of the unknown subject.

Rest of our paper is organized as follow. The working principle of meta-learning is presented first. The proposed methodology in the MAML framework and the experimental setup including the description of the dataset are given next. Then we report the results, draw conclusions and discuss future direction.

III. Model Agnostic Meta-learning

We use the Model Agnostic Meta Learning (MAML) method similar to [5]. The objective of MAML is that, a meta-trained model is capable of quickly learning a new task from a small number of examples. tasks. The meta-training process of different tasks makes the model capable of learning internal features that are broadly applicable to all the tasks, instated of a single task. In our case, an unseen task is defined to be able to animate an unknown lip image driven by an unknown audio. Next, we present the proposed methodology for accomplishing our task.

IV. Proposed Methodology

Given an audio signal represented by a set of n audio windows $A_a = \{A_a^0 \cdots A_a^{n-1}\}$ and target lip image A_i, our goal is to predict a set of 2D lip images $\hat{A} = \{\hat{A}_i^0 \cdots \hat{A}_i^{n-1}\}$ synchronized with A_a. Our work is inspired by [2]. While [2] does nothing specific to ensure that realisitc lip movement is generated for unseen faces, we use meta learning based neural network for this.

A. Network Topology

Figure 1 shows our network architecture. It has two encoders to extract feature representation from two modalities (image and audio) and uses a generator network which takes the fused representation of input audio and identity image to produce animated lips. We need to transform the audio signal A_a into features that are capable of representing the contents of the audio excluding subject identity, emotion and other features. Deep-speech (DS) [8] is shown to have such capability. We first transform the raw audio waveform, denoted as A_a, to Log-amplitude Mel-frequency Spectrum (LMS) features denoted as A_{lms}. Then, we pass A_{lms} through a pre-trained DS network. We use the output of the one but the last layer of the DS network. These features (denoted as A_{DS}) are used as input to our audio encoder. A convolutional audio encoder (Figure 1) carries out convolution on these features to encode the lip animation related information. The audio encoder θ_a generates the output feature f_a from the A_{DS} For visual stream, an input identity image, denoted as A_i,

is encoded by an identity encoder θ_i to output image features f_i. Lip generator takes f_a and f_i and fuse these two into f_v. It uses several residual blocks and 3D deconvolution operations to generate synthesized video $\hat{v} = \theta_g(f_v)$ where θ_g is the generator model.

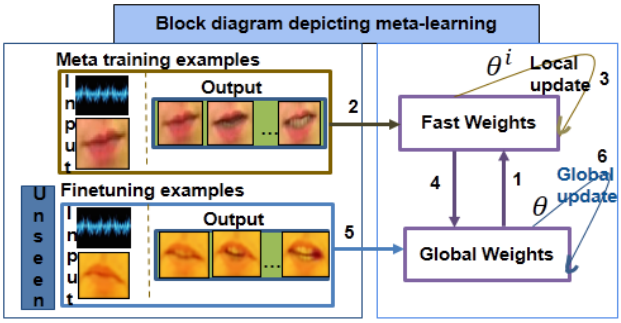


Fig. 2. Block diagram depicting a sample MAML training (meta-training and finetuing) method. The sequence of actions are depicted by giving indices from 1 to 6. For details see Section IV-C.

B. Objective Function

We use pixel-level reconstruction loss (denoted by L_{pix}) between the generated video $\hat{v}$ and the original video v given by the equation 1.

$$L_{pix}(v, \hat{v}) = ||v - \hat{v}||_2^2. \tag{1}$$

This loss tries to retain the texture of the identity image. However, we find that using it alone reduces the sharpness of the synthesized video frames. So, we use perceptual loss proposed by [11]. To get the perceptual features, we train an auto-encoder that reconstructs the video clips v. Perceptual loss (denoted by L_{per}) is defined as.

$$L_{per}(v, \hat{v}) = ||\tau(v) - \tau(\hat{v})||_2^2 \tag{2}$$

where τ is the encoder part of the pre-trained auto-encoder. Our objective is to minimize the composite loss given by 3

$$L = \alpha_1 L_{pix} + \alpha_2 L_{per}, \alpha_1 + \alpha_2 = 1. \tag{3}$$

C. Meta-learning stage

Fig. 2 shows the overall process for training a model in MAML way. The overall process can be broadly divided into three parts: (a) meta-training, (b) finetuning and (c) synthesizing. Meta-training considers a dataset as a set of episodes. Each episode consists of say, t number of tasks which split into train and test sets consisting of trn and qry samples respectively for each task. Each update iteration during meta-training is done with one episode of data. We define a task set as $T^{=}(A_i^1, A_{lms}^1) \cdots (A_i^{trn+qry}, A_{lms}^{trn+qry})$ where T refers to a task corresponding to a particular subject and $trn+qry$ is the number of samples for each subject (task) per episode. MAML is executed with the help of fast weights and global weights. Fast weights are used for learning each individual task whereas, global weights are updated based on the fast weights for global representation of all the training tasks. Initially, global weights are copied into the fast weights (operation 1 in Fig. 2). These fast weights are updated using gradient descent technique for one randomly chosen task using the trn training samples (operation 2 in Fig. 2). Then, qry samples of that task are used to measure the loss with respect to (wrt.) the fast weights. Let us denote this loss as l^1. Again global weights are copied to fast weights and the same process is repeated for all the t tasks (operation 3 in Fig. 2). Thus, we get $L = l^1 + l^2 \cdots + l^t$. This loss L is used to update the global weight (operation 4 in Fig. 2). Thus, the direction of global weight updating is such that it learns all the t tasks.

After the meta-learning converges, our model can learn to generate talking lip sequences for a new person, unseen during meta-learning stage. First, we need to update the network parameters θ, with few samples of the unseen subject so that our model can understand the subject's intrinsic traits. Finetuning consists of the same steps as described above (steps 5 and 6 in Fig. 2). Only here the number of tasks is one. The finetuned model can be used to synthesize animated lip image sequences given only one lip image of the target face and an audio signal.

V. Experiments and Results

A. Datasets

We have used two publicly available datasets: GRID [4] and TCD-TIMIT [6] and our own captured dataset. Most State-Of-the-Art (SOA) methods use these public datasets. Therefore, these datasets are best suited to compare the proposed method's performance with SOA. GRID consists of 33 subjects each having 1000 videos. In each video, a subject utters a small sentence. To evaluate the proposed framework on harder vocabulary, we use TCD-TIMIT [6]. TCD-TIMIT contains all white subjects. To increase the variety of the subjects' ethnicity, we create our own dataset and name it 'LVoc'. LVoc consists of videos where each of 28 subjects utters a subset of sentences taken from TIMIT dataset [6]. In LVoc, 300 sentences are randomly chosen from 6300 sentences of the TIMIT dataset. The audio of the GRID,TCD-TIMIT and LVoc datasets are sampled at 44.1KHz, 48KHz and 48KHz respectively and have frame rate of 25, 30 and 30 respectively. All the audio signals are transformed into 16KHz before using as input to the DS network.

B. Experimental Setting

We meta-train two models: one with randomly chosen 900 videos each from randomly chosen subjects from GRID dataset and one with the train set of TCD-TIMIT dataset. For finding the minimum number of samples that are enough for finetuning and generating realistic looking animation, we do a grid search. We observe finetuning with 10 samples is enough. For meta-training and finetuning a training sample consists of an audio segment, a randomly chosen image of the target and the sequence of images corresponding to the audio segment. To have a fair comparison with Chen et. al., [2], we keep the experimental setting as close as possible to [2].

From each video, we sample image sequences of window length of 16 images with a stride of 8 images. Considering 25fps, each image consists of 0.04 seconds and each sample of image sequence and the corresponding audio segment is of length 0.64 seconds. We use dlib for extraction of lips from face images. Each lip image is resized to 64×64 pixels. We encode audio into Mel-Frequency Cepstral Coefficients (MFCC) to use as input to the DS network. The number of Mel bands, the FFT window length and the number of samples between successive frames are 128, 1024 and 512. We use Adam optimizer with learning rate for global weight and local weight being 0.001, 0.01 respectively. We have done all the experiments on Quadro P500/PCIe/SSE2.

C. Results

a) Ablation Study: Here we establish the importance of different parts of our proposed network.

Deep-Speech features: Since the lip animation is driven

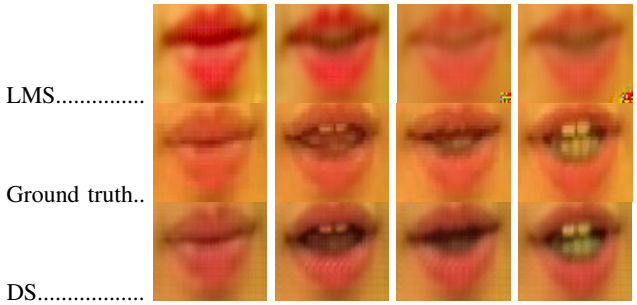


Fig. 3. Comparing lip animations generated from models trained with Deep Speech (DS) features and LMS features. Model trained with DS features seems to produce better lip animation in less time.

Fig. 4. Comparing lip animations generated from models trained with and without meta-learning. Model trained with meta-learning seems to produce better output in terms of skin color.

by only audio, the audio features should represent proper animation related features. On the other hand, an audio signal contains multiple characteristics including a person's identity, emotion and the content (what is uttered) of the audio. In our case, DS properly encodes the features necessary for fast convergence of the model. To establish this fact, we train two models, one with DS features and another with Log-amplitude Mel-frequency Spectrum (LMS) features. The model trained with DS seems to converge faster. Fig. 3 compares the results at 84000 iterations.

Meta-learning: To test meta-learning's effect on our network, we implement the same network architecture as shown in Fig. 1 but without meta-learning. The network is trained and tested on GRID data. The test subjects are not included in the training. Then we use transfer learning on the base non-meta model using 10 samples of the target person. The results are shown in Fig. 4. The lip images synthesized following transfer learning show different color than the target lip. The graph in Fig. 5 shows clear advantage of MAML over transfer learning in terms of loss with number of epochs.

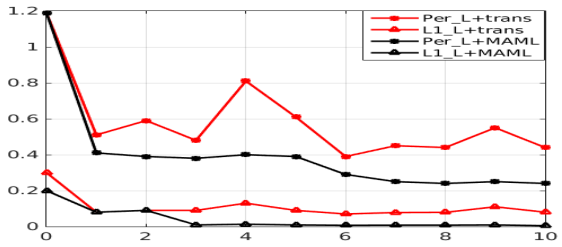
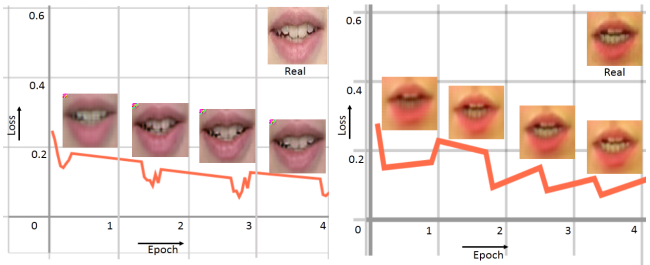


Fig. 5. Comparison of transfer vs. meta learning loss with increasing epochs. Per_L: Perceptual loss, L1_L: L1 loss, trans: transtransfer learning, MAML: Model Agnostic Meta Learning.

Cross Dataset Evaluation: We meta-train our model with one dataset say, GRID (or TCD-TIMIT) and finetune with 10 samples of one subject from the other dataset say TCD-TIMID (or GRID). Fig. 6 shows that for both the models two epochs of fineturing are enough for producing realistic results. The magnitude of loss for both the cross-model evaluations are comparable. These results establish the robustness of our model over dataset.



(a) (b)

Fig. 6. Loss and corresponding improvement of visuals over increasing finetuning epochs. The two models are trained on GRID (a) and TCD-TIMIT (b) datasets respectively and finetuned and tested on TCD-TIMIT and GRID dataset respectively. The groud-truth real image per test subject is shown at the upper-right corner.

b) Comparison with State-Of-the-Art (SOA): We compare our work with [2] and [12]. Out of the SOA, the work of [2] is the closest to that of ours. Both the processes synthesize lip animation driven by audio and given one lip image. Moreover, Chen et al. [2] claims generalization over unseen target face. Table I compares our method with [2] and [12] quantitatively. Our method outperforms [2] on all the measures. Better L1 loss and SSIM shows better reconstruction

TABLE I

QUANTITATIVE COMPARISON OF THE PROPOSED METHOD WITH STATE-OF-THE-ART. THE BEST RESULTS ARE SHOWN IN BOLD. NOTE THAT HIGHER PSNR AND SSIM MEAN BETTER RESULT AND LOWER L1 LOSS AND LMD INDICATE BETTER RESULT. THE MODELS ARE TRAINED AND TESTED WITH GRID [4] DATA. THE TEST SUBJECTS ARE NOT INCLUDED IN THE TRAINING DATA.

	PSNR	SSIM	L1 loss	LMD
Chen et al. [2]	27.50	0.73	1.58	2.10
KR et al. [12]	**33.40**	0.96	-	**0.6**
Proposed	31.3	**0.98**	**1.24**	1.20

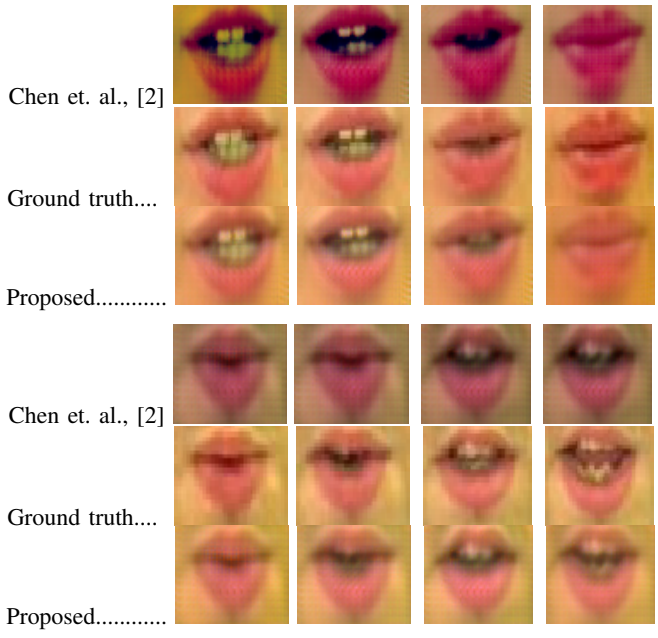


Fig. 7. Comparison of lip images generated with the proposed method with that of Chen et. al., [2]. The models are trained and tested with GRID data [4]. Note that the lips generated by the proposed method better retains the skin color as compared to those generated by [2]. Moreover, shape of the lips generated by the proposed method better resembles the ground truth as compared to [2] (Observe second and fourth column of the second example).

ability of our method over [12]. Better LMD measure signifies better animation capability of our method over [2]. Note that the model by [2] is trained and tested on the data from the same person whereas in our result test faces are not present in the training data. This explains slightly better PSNR by [2] in Table I. Similar results are shown qualitatively in Fig. 7 and 8. The method of [2] could not preserve the subject related intricacies such as skin color or lip shape. Our method due to finetuning, learned these intricacies with few-shot learning. [12] blends generated lip with the rest of the face which results in blurred face and compromised identity. Moreover, generated lips are not properly synced with the audio for unseen faces.

D. Conclusions and future direction

We have proposed a meta-learning based generative model that produces highly realistic lip animation driven by audio. Only a few (ten) samples are enough to finetune the meta-trained model to the unseen face. Our study suggests that deep-speech features help in better encoding of the audio which in turn helps in better animation. On the other hand, meta-learning helps in better and faster learning of person

Fig. 8. Comparison of lip images generated with the proposed method with that of [12]. The models are trained with TCD-TIMIT [6] and tested on our LVoc. Note that lip shapes generated by our method better approaches the ground truth. The generated face has got better contrast for our method.

specific intricacies. Next, we are further extending our work for animating whole faces driven by text.

REFERENCES

[1] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, pages 3981–3989, 2016.

[2] L. Chen, Z. Li, R. K Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018.

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

[4] 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.

[5] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[6] J. S. Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

[7] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019.

[8] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[9] S. Hochreiter, A. S. Younger, and P. R. Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001.

[10] A. Jamaludin, J. S. Chung, and A. Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, pages 1–13, 2019.

[11] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[12] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1428–1436, 2019.

[13] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[14] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017.

[15] K. Vougioukas, S. Petridis, and M. Pantic. Realistic speech-driven facial animation with gans. *arXiv preprint arXiv:1906.06337*, 2019.

[16] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint arXiv:1905.08233*, 2019.

[17] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.