# Show Me Your Face, And I'll Tell You How You Speak

Christen Millerdurai[1*], Lotfy Abdel Khaliq[1*], and Timon Ulrich[1*]

Saarland University

Abstract. When we speak, the prosody and content of the speech can be inferred from the movement of our lips. In this work, we explore the task of lip to speech synthesis, i.e., learning to generate speech given only the lip movements of a speaker where we focus on learning accurate lip to speech mappings for multiple speakers in unconstrained, large vocabulary settings. We capture the speaker's voice identity through their facial characteristics, i.e., age, gender, ethnicity and condition them along with the lip movements to generate speaker identity aware speech. To this end, we present a novel method "Lip2Speech", with key design choices to achieve accurate lip to speech synthesis in unconstrained scenarios. We also perform various experiments and extensive evaluation using quantitative, qualitative metrics and human evaluation.

Keywords: Lip Reading in the Wild · Speech synthesis · Multi-speaker speech generation.
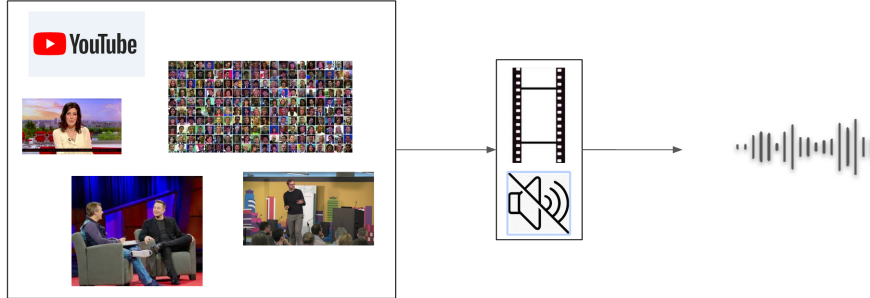
Fig. 1. Our objective is taking a silent video of a person speaking and generate speech from their lip movements.

---

* All authors contributed equally(Listed alphabetically).

1 Introduction

Lip reading is a technique of understanding speech by visually interpreting the movements of the lips. Although lip reading is used most extensively by deaf and hard-of-hearing people, most people with normal hearing process some speech information from sight of the moving mouth [47]. Also, understanding linguistic cues from readings the lips can improve the clarity of the conversation in environments with noisy backgrounds. One possible approach for lip reading is to generate text for the corresponding lip movements. Although this method works, humans also emote through their lip movements which is lost in the textual representation. Also, they cannot represent non-linguistic vocal repertoire[2]. Moreover, several use-cases like voice inpainting[53], speech recovery from background noise[10], generating a voice for people who cannot produce voiced sounds (aphonia) are not possible. While Lip-to-text-to-speech might solve certain problems, the intermediate textual information distills the prosody, tempo, non-verbal cues of the lip movements which are essential to solve most of the problems mentioned above.Hence, both Lip-to-text, Lip-to-text-to-speech are at a disadvantage and a direct lip-to-speech method is needed. Moreover, text-based systems require transcribed datasets for training, which are hard to obtain because they require laborious manual annotation. And training of Lip-to-speech methods can be done in a self-supervised manner since most video comes paired with the corresponding audio. While, we have established Lip-to-speech methods are superior to compared to other methods, generating voice requires different vocal identity for each person to bolster realism. Hence, the Lip-to-speech methods should be conditioned on the speaker's vocal identity while generating the speech content. While methods like [13, 23, 3, 30, 17] sample the voice from the target speaker which are then conditioned to generate the content, we assume that a voice to be "cloned" is not available at test time. Hence, we rely on facial features such as the speaker's gender, age, ethnicity to generate a voice which is then conditioned along with the lip movements to generate the speech.

While several methods[22, 43, 40] have tried to solve lip to speech synthesis, most methods focus on clinical settings, i.e GRID corpus[1] but our work mainly focuses on lip reading in the wild. We draw parallel to two works, Lip2Wav[31] and Facetron[40], while Lip2Wav also accomplishes our objective, its multi-speaker pipeline is conditioned on the speaker's voice, which if not available will not work. While Facetron does not have this issue, its GAN[14] based decoder limits its capability to generate very long sequences. Moreover, even though Facetron uses the face to condition the speaker's voice, it is not demonstrated on Lip reading in the wild(LRW)[6] dataset. In this work, we propose "Lip2Speech", an end-to-end model that is capable of directly converting silent video to melspectogram, which can converted to raw audio either by Griffin-Lim Algorithm[15] or neural vocoders[27, 26].

Our method is based on encoder-decoder architecture, where we encode the lip movements along with the face features and decode the melspectorgram of the generated speech. The generated speech is evaluated using standard audio reconstruction and speech synthesis metrics. Additionally, we propose YouTube

Lip Data(YLD), a pipeline that can create training data from any YouTube video.

We release the code, trained models publicly for future research along with a demonstration video here[1]. The rest of the paper is organized as follows: In Section 2, we survey the recent developments in this space. Following this, we describe our novel Lip2Speech network in Section 3. Our datasets, training details and implementation details are explained in Section 4. And in Section 5, we perform various experiments with our method and introduce the YLD pipeline. And finally, we conclude our work in Section 8.

2   Related work

2.1   Face reconstruction from Audio

Several methods have been proposed to infer visual information from speech signals. There are two research directions for this problem: one more towards graphic-based generation[36, 38] and the other towards pixel-level generation. However, graphic-based methods typically parametrize the reconstructed face using a priori face model and texture. In pixel-level generation, Sadoughi et al.[33] use speech to reconstruct lip motions. Duarte et al.[8] reconstruct the face including the expression and pose from speech using GANs. While several methods have explored inferring separate face traits[51, 12], William et al[25]. did not restrict the model to learn pre-defined facial traits but rather a more general visual representation of how the person speaking looks like.

2.2   Lip reading

Many end-to-end approaches have been proposed to tackle this problem and can be divided into two categories. In the first category, an MLP is used to extract features and LSTM layers model the temporal dynamics of the sequence[29, 45]. In the second category, a 3D conv layer is used followed by standard conv layers combined with LSTMs. However, these methods do not extract features from audio directly, and they rely on MFCCs as a feature extractor and an attention mechanism is applied to lip crops and MFCCs.[7, 48]. Current state of the art methods combine all three, i.e., applies spatio-temporal convolution to the lip ROIs followed by a spatial feature extractor coupled with LSTMs.

2.3   Lip to Text Generation

Although multiple methods[45, 49] were proposed to generate text from lip movements, most of them are limited to small vocabulary space. Also, there has been multiple approaches that tackle the problem in an in-the-wild manner by using sequence-to-sequence transformer[41] models to generate sentences given a silent lip movements sequence.

---

[1] https://github.com/Chris10M/Lip2Speech

2.4 Speech Synthesis

Despite the fact that a lot of research has been investigated in generating natural speech from text, it still remains a challenging task. Existing statistical parametric speech synthesis[4, 39] use a vocoder to generate speech from trajectories of speech features. However, they produce an unnatural sounding voices. Recent works such as WaveNet[26] produces audio quality that is very close to the real human speech but the inputs to the model (linguistic features, predicted log fundamental frequency(F0), and phoneme duration) are hard to produce and requires significant domain expertise. Tacotron[46], a sequence-to-sequence architecture for producing melspectrograms from a sequence of characters, mitigates this issue by replacing the production of these linguistic features with a single neural network trained from data alone. Other approaches strive to generate speech with the voice of different speakers. Jia et al.[17] proposed a model that is composed of three independently trained neural networks, (i) a speaker encoder[44], (ii) a sequence-to-sequence synthesizer[34] (iii) a neural vocoder[26]

2.5 Lip to Speech Generation

End-to-end methods as Vid2Speech[11] and Lipper[20] extract LPC (Linear Predictive Coding) features for k subsequent frames followed by applying a 2-D CNN. The drawback of such methods is that head motion is not taken into consideration so it does not model a real-world scenario. Another approach was taken to improved speech quality by generating waveforms using GANs but they do not make use of of sequence-to-sequence paradigm that is used for text-to-speech generation which can greatly improve the speech quality. Furthermore, these works provide results on a narrow vocabulary and very minimal head motion which might be ill-suited for real-world scenarios. Recent works such as [31] address these issues by using spatio-temporal encoder coupled with attention based sequence-to-sequence decoder for high-quality speech generation.

3 Lip2Speech Network

In this section, we introduce our proposed network, Lip2Speech. We start by first introducing the problem statement, then explain the speaker identity module, followed by explaining the Lip-reading module and the speech synthesis module.

3.1 Problem Statement

Given an input video of a speaking face, our objective is to synthesize the speech of the speaking face with the identity of the speaker estimated through their face attributes, i.e., gender, age, ethnicity. The video contains a sequence of image frames, $F = (F_1, F_2, F_3, ..., F_N)$, and while the synthesized speech contains speech frames of the melspectogram, $S = (S_1, S_2, S_3, ..., S_T)$. And this can be formulated as a sequence-to-sequence learning task[35], where the correspondences between the image frames and speech frames are learnt. These
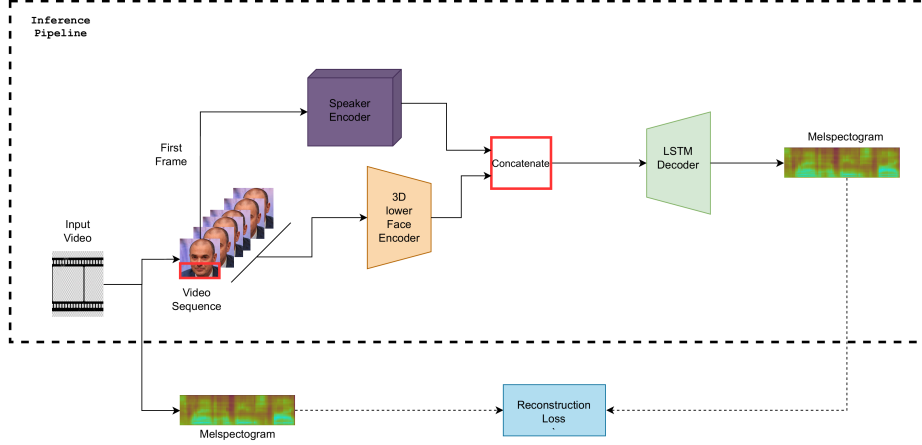
Fig. 2. The general overview of our architecture.

correspondences capture both the speaker identity and speech content. And, after learning the correspondences, speech frames can be synthesized for the image frames in an auto-regressive manner. It is to be noted that the number of image frames, N and number of speech frames, T does not need to be equal. Concretely, each output speech frame, S_t is modelled as a conditional distribution of the previous speech frame and image frames, F.

$$P(S|F) = \prod_{t=1}^{t=T}(S_t|S < t, F)$$

3.2 Architecture Overview

For an input frame sequence F $= (F_1, F_2, F_3, ..., F_N)$, the first frame is passed to the speaker encoder to get the "speaker embeddings". Also for every frame F_n, the lip ROIs are cropped and passed to the 3D lower face encoder to get the face features. The speaker embeddings are tiled and then concatenated to the face features along the frame sequence N, which we call as "visual features". These visual features are passed to the LSTM decoder which generates the melspectogram of the speech in an auto-regressive manner. As shown in Figure 2, during training time we use the melspectogram from the input video and minimize the mean square error to the generated melspectogram.

3.3 Speaker Identity Module

The speaker identity module, as shown in Figure 3 is used to learn the correlations between a speaker's face and their voice. We know from [18, 24], humans have the ability to "put a face to a voice" and [25] tries to generate a face using a given voice sample. In our work, we try to "produce a voice to a face",

an inverse of previous work. And we do this by generating encodings which are tied to every individual speaker. While [13, 23, 3, 30, 17] also use encodings to map the speaker identity, they condition them using the speaker's voice. On the contrary, our work conditions the encodings from the face of the speaker.

We use a Inception-ResNet[37] that is pre-trained on CASIA-WebFace[50] as the face recognition encoder. We freeze the first 3 blocks of the network and fine-tune the reset, i.e., 2 blocks. This network taken in a face and produces a face encoding. Our objective is to, learn the correspondences of the face that correlates to the speakers voice, hence we use a another network, speech encoder[17]. The speech encoder takes in a melspectogram and generates a speaker encoding that correlates to the speaker's identity. And the speech encoder is frozen throughout the training.

We learn a cross-modal mapping between the face encodings and the speaker encodings, which we define as "speaker embeddings" through instance based contrasting learning.
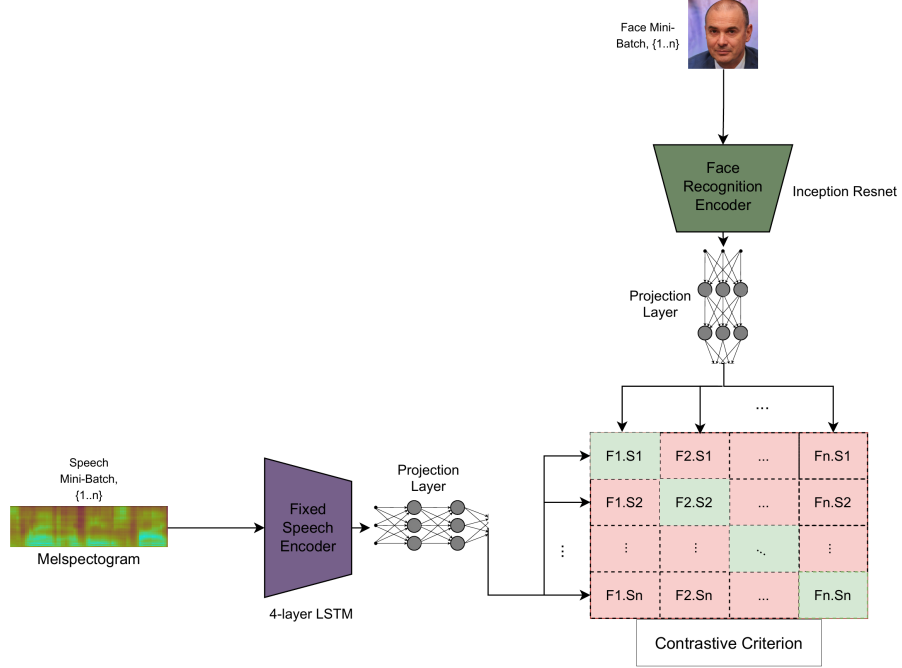

Fig. 3. The speaker encoder is trained to discriminate

3.4 Video feature extractor

The Video feature extractor, as shown in Figure 4 is used to extract the semantic and temporal lip features from the video frame sequence, F. We pass F =

$(F_1, F_2, F_3, ..., F_N)$ to a 3-D convolution block to induce temporal awareness of the frame features and then apply a 2-D feature extractor, ShuffleNet V2[21] to extract the semantic features. We preserve the temporal length, N by padding appropriately and reduce the spatial dimensions to 1 by performing global average pooling to the extracted frame features.
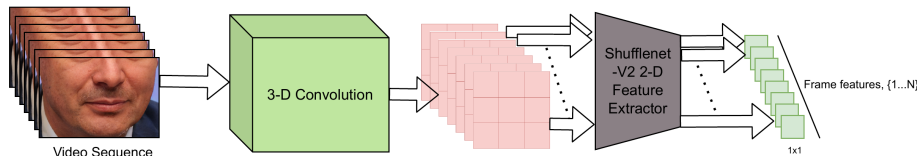
Fig. 4. The video feature extractor encodes the frame sequence by applying spatio-temporal convolutions.

3.5   Melspectogram Decoder

The speaker embeddings are tiled and concatenated to the frame features according to the frame sequence length N.

These "visual features" are taken in by the melspectogram decoder, as shown in Figure 5 which decodes them to melspectogram frames(melspec-frame) in a auto-regressive manner. Since the frame sequence length N, and the number of time steps T does not need to be the same, we first encode the visual features into a latent representation using the 2-layer BiLSTM's hidden and cell state. We use a 4-layer LSTM[16] initialized with the BiLSTM's hidden and cell state to perform the decoding. The initial melspec-frame input to the decoder is trainable and jointly optimized along with the network. And, the regression of each time-step is done by passing the previous melspec-frame to an bottleneck network with heavy dropouts, called as "Prenet", which reduces the ground-truth bleeding during teacher forcing. It also forces the LSTM not to ignore the attended queries, as we restrict the information flow from the previous time-step. And to know when to stop the regression i.e., "stop token", we use a linear projection of the hidden state of each time-step concatenated with the hidden state of latent representation. The linear projection is activated using the sigmoid function, which acts as the gating value. During inference, we stop if the gating value is above a certain fixed threshold, with a default being 0.5.

We use localized attention mechanism to improve the contextual information from the frame sequences, as the condensed latent encoding of the frame sequences will not be able to completely represent the temporal semantic flow. We generate the key, value by convolving 1-D convolution filters across the BiLSTM's outputs. And the query is generated for each time-step by using the previous melspec-frame and current hidden state. We add sinusoidal positional encodings[42] to key, query and value, to induce position-aware querying, thereby

preventing previous queries to be concatenated for the query of the current time-step.
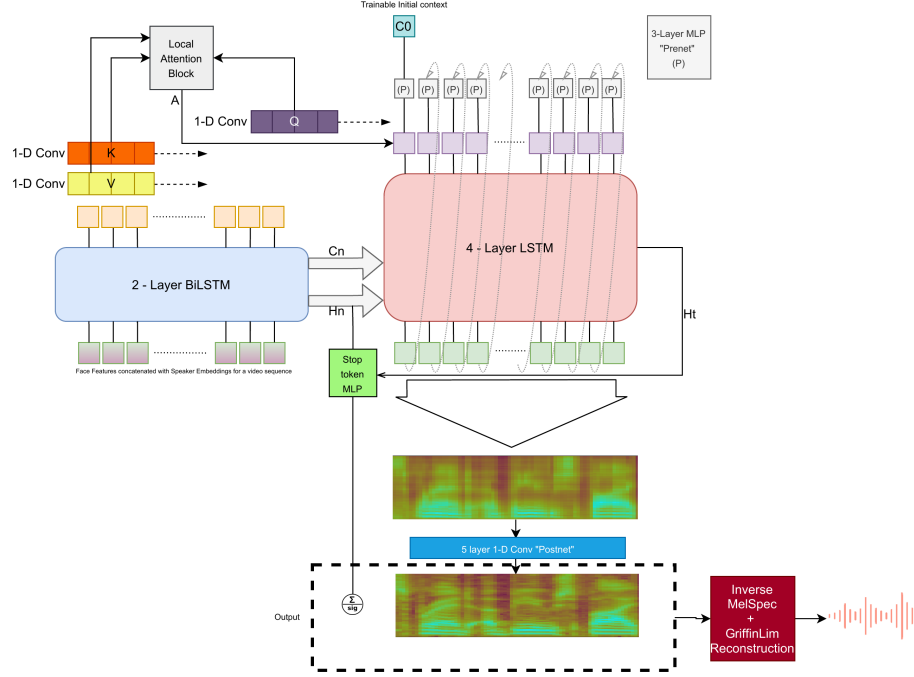
Fig. 5. The melspec decoder encodes the frame sequence into a condensed latent repretation and decoded melspec-frame auto-regressively. And attention mechanism is used to improve the temporal semantic flow.

4 Methodology

4.1 Datasets

AVSpeech is a large-scale audio-visual dataset[9] . It consists of 3 to 15 second clips from YouTube videos. We sample 33 thousand clips from this dataset and use it to train the Speaker Encoder.

Lip Reading in the Wild consists of up to 1000 utterances of 500 different words, spoken by hundreds of different speakers[6]. We sampled 70 clips from the dataset for MOS and Face encoder evaluation. We also sample another 153 clips for Lip2Speech evaluation.

UTKFace [52] dataset is a large-scale face dataset with over 20,000 face images of people with ages ranging from 0 to 116 years old with different gender and ethnicity. We use UTKFace for training our face decoder. We randomly sample 3000 images from the dataset, then pass them to our speaker encoder to get embeddings and save them.

4.2   Implementation Details

Training settings All experiments were implemented with PyTorch[28]. We use the adam optimizer[19] with initial learning rate set as 0.001 for training all experiments. We have developed a prototyping framework for models, where any change to the model, will generate appropriate model/validation/training logs automatically. We extensively use this to log our experiments.

To train the speaker encoder on AVSpeech, for every video sequence, a random frame is taken and we crop the face and its corresponding audio window, where the window length is chosen random from $1 \sim 3$ seconds. We train the network with a batch size of 64 and train until the contrastive criterion does not improve.

The Lip2Speech network is trained for 300 epochs on both LRW and YLD, with early stopping when the validation accuracy does not improve for 10 epochs. We use a batch size of 84 and teacher force the decoder inputs. The teacher forcing is annealed for every 10 epochs, which slowly shifts the decoder to test sequence generation. We also augment the train data with random horizontal flips for the lip sequence.

Inference settings During inference time, the video frame sequence and a single frame of the sequence is given as the input to the network. We decode the melspec-frames auto-regressively until the stop-token threshold is greater than 0.5. After generating the melspectogram, we run Griffin-Lim Algorithm[15] to reconstruct the raw audio waveform. And to reconstruct long sequences in the YLD dataset, we split the video into 2-second chunks and generate melspectogram for each chunk. We concatenate these chunks and apply Griffin-Lim to get the raw audio waveform. We found this method to be effective as the melspectogram inversion implicitly overlaps across melspec-frames to reconstruct the raw audio waveform, preventing abrupt discontinuities between adjoining audio chunks.

5   Experiments

5.1   Speaker Encoder

Mean Opinion Score The evaluation is done via the pre-trained Real-Time Voice Cloning repository which is based on SV2TTS[17]. The network takes a text and a sample audio of a speaker to generate the given text in the given voice. The pipeline consists of three separately trained parts. First off a speaker

encoder which takes an audio sample of the speaker as input and outputs a speaker embedding which the network will use to generate the correct voice. The next part is the synthesizer. Its input is the given text and the speaker encoding to output a melspectogram. The last part of the pipeline is a vocoder which generates raw audio from the melspectogram. The relevant design choice here is that the speaker encoder is separately trained and thus replaceable. We use this to give the model our speaker embeddings generated from the faces of people instead. With this, the pipeline produces an audio which we can compare to the ground truth and also to the audio generated by the original pipeline using an audio sample to encode the voice.

For the evaluation, 70 samples from the Lip Reading in the Wild dataset[6] are passed through the pipeline. Once with the pre-trained speaker encoder and once with our speaker encoder. Then these 2 samples as well as the ground truth are graded with two metrics, their voice quality and the correlation between the voice and the image of the speaker. As shown in table1, our model provides an overall better quality but the speaker audio embeddings produce voices with higher correlation. It is noticeable that the generated audio, no matter the given embeddings, is highly biased towards middle-aged white people. The network properly discerns between male and female with only a few outliners and different age groups have different amounts of energy to their voice. For both genders the voices do not deviate much from the mean voice. It is also perceptible that the voices generated for young students sound like adults. Also for different ethnicity's than white the generated audios sound distorted instead of producing a proper accent.

Table 1. Voices generated with different embeddings and their respective MOS score compared to the ground truth.

Voice	Quality		Voice	Correlation
ground truth	**4.56**		ground truth	**4.44**
speaker audio embedding	3.37		speaker audio embedding	3.12
speaker face embedding	3.55		**speaker face embedding**	3.03

Speaker Encoding Decoder In order to evaluate our speaker encoder qualitatively besides the voice quality, we need to visualise what the model has learned, i.e what traits of the face are captured in the speaker embedding. To achieve this goal, we train a face decoder that projects the 256 dimensional speaker embedding into a $64 \times 64 \times 3$ image. Our decoder is based on DCGAN[32].

UTKFace dataset [52] was chosen to train the face decoder. Then the decoder training process starts by passing the saved embeddings of 153 samples of the dataset and optimizing L2 distance between the reconstructed face and the GT as shown in Fig 6.
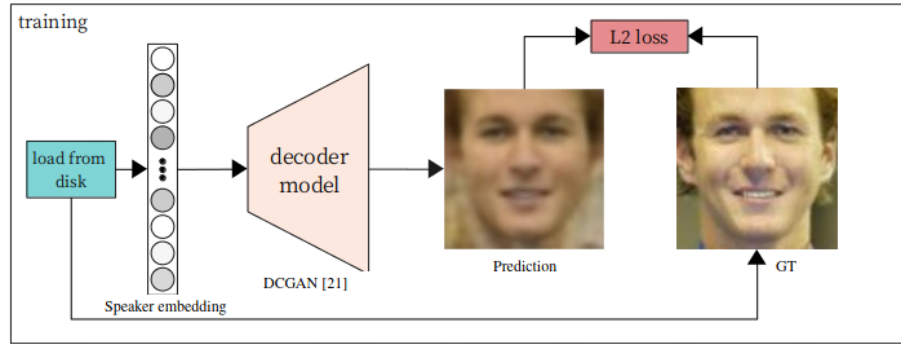
Fig. 6. Face decoder training process. Our decoder is based on [32] and adapted to our face encoder's embedding output shape.

Mean-Face Analysis The results of the decoder are visualized in Figure 7. As one can see the key features of the original face are reconstructed. The wrinkles of old people are still present which indicates that the voice of older people is successfully detected as different and thus encoded as a feature. The same can be said for gender features as well as skin color. Non-important features for the voice like glasses are removed. In terms of age we start to see some limitations, because children start to look like young adults. We also notice some unexpected results. For some reason some people smile while others do not. We were not able to link these smiles to any voice feature like a certain pitch and further testing would be required. Even more surprising is how beards are treated. One would assume that beards do not affect a persons voice and yet beards remain clearly visible. They often get reduced but never completely vanish. Related to this is the last image in Figure 7 where the brown skin color was lost and replaced with white skin while the beard remained very visibly. Again we were not able to link a beard to a voice feature.

Fig. 7. Left is the ground truth and right is the reconstructed face.

5.2 Lip2Speech

We use LRW dataset for evaluating our model. We sample 153 videos from the whole dataset with people of different age, gender, ethnicity and accent. In fig 8, sample faces and the corresponding generated melspectograms are shown. We show Short-time Objective Intelligibility (STOI), extended STOI (ESTOI), Perceptual Evaluation of Speech Quality (PESQ), and Word Error Rate (WER) metrics results for the generated speech in table 2. STOI, ESTOI, and PESQ metrics measure the speech intelligibility which is the the degree to which speech sounds (whether conversational or communication-system output) can be correctly identified and understood by listeners.
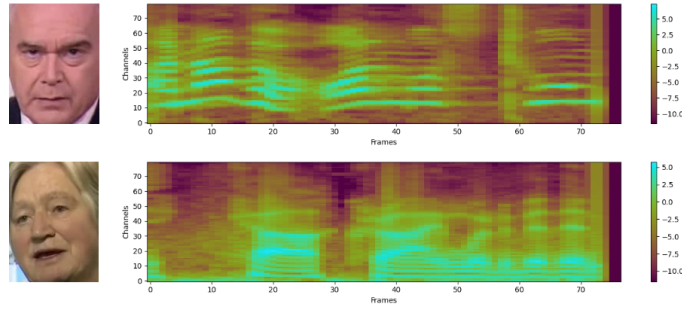

Fig. 8. Faces and the melspectograms generated with our Lip2Speech network.

Table 2. Quantitative results of our model on the 153 test samples

	STOI ↑	ESTOI ↑	PESQ ↑	WER ↓
Lip2Speech	1.38	0.66	0.42	26.1%

5.3   YouTube Lip Data (YLD)

The YouTube lip data(YLD) can transform any YouTube video into a Lip2Speech dataset. Given a YouTube video link and a target face that needs to be captured, the video is split into short segments of $1 \sim 3$ seconds. And only segments which contain the target face are taken by by performing face recognition[5] on every frame.
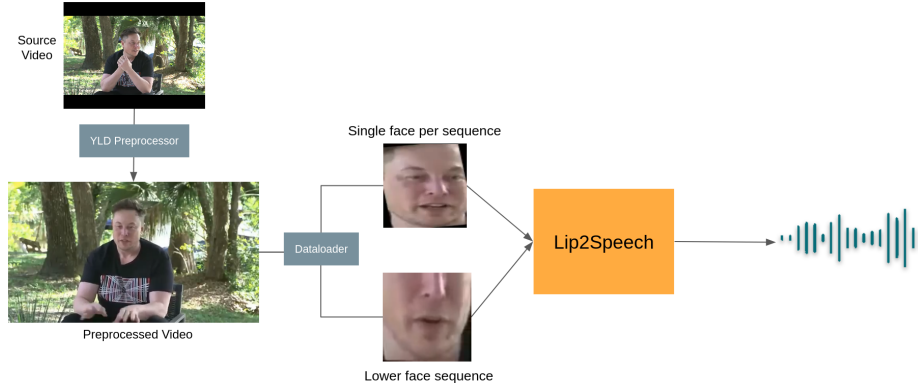
Fig. 9. The YLD pipeline can transform any Youtube video into training/testing data.

6   Limitations and Future Work

First off we note that our subset of AVSpeech[9] to train the speaker encoder has been chosen randomly and thus might contain a bias towards a certain group of people. From the samples created for the MOS with our speaker face embeddings we observe that the generated voice is lacking in variation. Also the mean-face analysis shows that some features are not represented and need further learning. A bigger and more balanced training set should improve our results on this end.

Lip2Speech requires a lot of training data, and the generated voice sounds robotic. A way to alleviate the robotic voice is to use a neural vocoders[27, 26] rather than optimization based Griffin-Lim Algorithm[15]. And to improve the generalization and reduce the training data, the vocabulary of the words need to be clustered. This takes us in a direction where we also learn the visual and linguistic cues present in the context of lip reading.

7   Concluding Remarks

In this work, we proposed a method to synthesize speech from lip movements. Specifically, we focused on generating different vocal identity for individual

speakers based on their age, gender, ethnicity. As our method uses the speaker embedding to incorporate speaker-specific information in the speech synthesis step, which allows the model to be used for all speakers, even ones that were previously unseen during training. We have evaluated our model with extensive quantitative metrics and human studies. All the code and data for our work has been made publicly available[4].

References

[1] Alghamdi, N., Maddock, S., Marxer, R., Barker, J., Brown, G.J.: A corpus of audio-visual lombard speech with frontal and profile views. The Journal of the Acoustical Society of America **143**(6), EL523–EL529 (Jun 2018). https://doi.org/10.1121/1.5042758, https://doi.org/10.1121/1.5042758

[2] Anikin, A., Bååth, R., Persson, T.: Human non-linguistic vocal repertoire: Call types and their meaning. Journal of Nonverbal Behavior **42**(1), 53–80 (Sep 2017). https://doi.org/10.1007/s10919-017-0267-y, https://doi.org/10.1007/s10919-017-0267-y

[3] Arik, S.Ö., Chen, J., Peng, K., Ping, W., Zhou, Y.: Neural voice cloning with a few samples. In: NeurIPS (2018)

[4] Black, A.W., Zen, H., Tokuda, K.: Statistical parametric speech synthesis. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. vol. 4, pp. IV–1229–IV–1232 (2007). https://doi.org/10.1109/ICASSP.2007.367298

[5] Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision (2017)

[6] Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Asian Conference on Computer Vision (2016)

[7] Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jul 2017). https://doi.org/10.1109/cvpr.2017.367, http://dx.doi.org/10.1109/CVPR.2017.367

[8] Duarte, A., Roldan, F., Tubau, M., Escur, J., Pascual, S., Salvador, A., Mohedano, E., McGuinness, K., Torres, J., i Nieto, X.G.: Wav2pix: Speech-conditioned face generation using generative adversarial networks (2019)

[9] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv preprint arXiv:1804.03619 (2018)

[10] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party. ACM Transactions on Graphics **37**(4), 1–11 (Aug 2018). https://doi.org/10.1145/3197517.3201357, http://dx.doi.org/10.1145/3197517.3201357

[11] Ephrat, A., Peleg, S.: Vid2speech: Speech reconstruction from silent video. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5095–5099 (2017). https://doi.org/10.1109/ICASSP.2017.7953127

[12] Feld, M., Burkhardt, F., Müller, C.: Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services. pp. 2834–2837 (01 2010)

[13] Gibiansky, A., Arik, S.Ö., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J., Zhou, Y.: Deep voice 2: Multi-speaker neural text-to-speech. In: NIPS (2017)

[14] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014)

[15] Griffin, D., Lim, J.: Signal estimation from modified short-time fourier transform. In: ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 8, pp. 804–807 (1983). https://doi.org/10.1109/ICASSP.1983.1172092

[16] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[17] Jia, Y., Zhang, Y., Weiss, R.J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I.L., Wu, Y.: Transfer learning from speaker verification to multispeaker text-to-speech synthesis (2019)

[18] Kamachi, M., Hill, H., Lander, K., Vatikiotis-Bateson, E.: 'putting the face to the voice': Matching identity across modality. Current Biology **13**(19), 1709–1714 (2003). https://doi.org/https://doi.org/10.1016/j.cub.2003.09.005, https://www.sciencedirect.com/science/article/pii/S0960982203006638

[19] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)

[20] Kumar, Y., Jain, R., Salik, K.M., Shah, R., Yin, Y., Zimmermann, R.: Lipper: Synthesizing thy speech using multi-view lipreading. In: AAAI (2019)

[21] Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design (2018)

[22] Mira, R., Vougioukas, K., Ma, P., Petridis, S., Schuller, B.W., Pantic, M.: End-to-end video-to-speech synthesis using generative adversarial networks (2021)

[23] Nachmani, E., Polyak, A., Taigman, Y., Wolf, L.: Fitting new speakers based on a short untranscribed sample. ArXiv **abs/1802.06984** (2018)

[24] Nagrani, A., Albanie, S., Zisserman, A.: Seeing voices and hearing faces: Cross-modal biometric matching (2018)

[25] Oh, T.H., Dekel, T., Kim, C., Mosseri, I., Freeman, W.T., Rubinstein, M., Matusik, W.: Speech2face: Learning the face behind a voice (2019)

[26] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio (2016)

[27] van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., van den Driessche, G., Lockhart, E., Cobo, L.C., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D., Hassabis, D.: Parallel wavenet: Fast high-fidelity speech synthesis (2017)

[28] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H.,

Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[29] Petridis, S., Li, Z., Pantic, M.: End-to-end visual speech recognition with lstms. In: 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017 - Proceedings. pp. 2592–2596. IEEE, United States (Jun 2017). https://doi.org/10.1109/ICASSP.2017.7952625, http://www.ieee-icassp2017.org/, 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, ICASSP ; Conference date: 05-03-2017 Through 09-03-2017

[30] Ping, W., Peng, K., Gibiansky, A., Arik, S.Ö., Kannan, A., Narang, S., Raiman, J., Miller, J.: Deep voice 3: 2000-speaker neural text-to-speech. ArXiv **abs/1710.07654** (2017)

[31] Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V., Jawahar, C.V.: Learning individual speaking styles for accurate lip to speech synthesis (2020)

[32] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2016)

[33] Sadoughi, N., Busso, C.: Speech-driven expressive talking lips with conditional sequential generative adversarial networks (2018)

[34] Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R.A., Agiomyrgiannakis, Y., Wu, Y.: Natural tts synthesis by conditioning wavenet on mel spectrogram predictions (2018)

[35] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks (2014)

[36] Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: Learning lip sync from audio. ACM Trans. Graph. **36**(4) (Jul 2017). https://doi.org/10.1145/3072959.3073640, https://doi.org/10.1145/3072959.3073640

[37] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning (2016)

[38] Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J., Matthews, I.: A deep learning approach for generalized speech animation. ACM Trans. Graph. **36**(4) (Jul 2017). https://doi.org/10.1145/3072959.3073699, https://doi.org/10.1145/3072959.3073699

[39] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech synthesis based on hidden markov models. Proceedings of the IEEE **101**(5), 1234–1252 (2013). https://doi.org/10.1109/JPROC.2013.2251852

[40] Um, S.Y., Kim, J., Lee, J., Oh, S., Byun, K., Kang, H.G.: Facetron: Multi-speaker face-to-speech model based on cross-modal latent representations (2021)

[41] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon,

I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[42] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)

[43] Vougioukas, K., Ma, P., Petridis, S., Pantic, M.: Video-driven speech reconstruction using generative adversarial networks (2019)

[44] Wan, L., Wang, Q., Papir, A., Moreno, I.L.: Generalized end-to-end loss for speaker verification (2020)

[45] Wand, M., Koutník, J., Schmidhuber, J.: Lipreading with long short-term memory (2016)

[46] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q.V., Agiomyrgiannakis, Y., Clark, R., Saurous, R.: Tacotron: Towards end-to-end speech synthesis. In: INTERSPEECH (2017)

[47] Woodhouse, L., Hickson, L., Dodd, B.: Review of visual speech perception by hearing and hearing-impaired people: clinical implications. International Journal of Language & Communication Disorders **44**(3), 253–270 (Jan 2009). https://doi.org/10.1080/13682820802090281, https://doi.org/10.1080/13682820802090281

[48] Xia, L., Chen, G., Xu, X., Cui, J., Gao, Y.: Audiovisual speech recognition: A review and forecast. International Journal of Advanced Robotic Systems **17**(6), 1729881420976082 (2020). https://doi.org/10.1177/1729881420976082, https://doi.org/10.1177/1729881420976082

[49] Xu, K., Li, D., Cassimatis, N., Wang, X.: Lcanet: End-to-end lipreading with cascaded attention-ctc. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) pp. 548–555 (2018)

[50] Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch (2014)

[51] Zazo, R., Sankar Nidadavolu, P., Chen, N., Gonzalez-Rodriguez, J., Dehak, N.: Age estimation in short speech utterances based on lstm recurrent neural networks. IEEE Access **6**, 22524–22530 (2018). https://doi.org/10.1109/ACCESS.2018.2816163

[52] Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder (2017)

[53] Zhou, H., Liu, Z., Xu, X., Luo, P., Wang, X.: Vision-infused deep audio inpainting (2019)