LEARNING TO PREDICT SPEECH IN SILENT VIDEOS VIA AUDIOVISUAL ANALOGY

Ravindra Yadav[⋆] *Ashish Sardana*[†] *Vinay P Namboodiri*[‡] *Rajesh M Hegde*[⋆]

[⋆] Indian Institute of Technology Kanpur, India
[†] NVIDIA
[‡] University of Bath, UK

ABSTRACT

Lipreading is a difficult task, even for humans. And synthesizing the original speech waveform from lipreading makes it even a more challenging problem. Towards this end, we present a deep learning framework that can be trained end-to-end to learn the mapping between the auditory and visual signals. In particular, in this paper, our interest is to design a model that can efficiently predict the speech signal in a given silent talking-face video. The proposed framework generates a speech signal by mapping the video frames in a sequence of feature vectors. However, unlike some recent methods that adopt a sequence-to-sequence approach for translation from the frame stream to the audio stream, we posit it as an analogy learning problem between the two modalities. In which each frame is mapped to the corresponding speech segment via a deep audio-visual analogy framework. We predict plausible audio stream by training adversarially against a discriminator network. Our experiments, both qualitative and quantitative, on the publicly available GRID dataset show that the proposed method outperforms prior work on existing evaluation benchmarks. Our user studies confirm that our generated samples are more natural and closely match the ground truth speech signal.

Index Terms— Speech prediction, AudioVisual analogy, self-supervised learning.

1. INTRODUCTION

Human learning in the real world often involves understanding the correspondence between multiple co-occurring modalities. Thus, for an AI system to operate efficiently in a real-world environment, a key component that is needed is its ability to reason about the multimodal signals together. In this paper, we investigate the task of speech prediction, a particular instantiation of self-supervision [1, 2] where a model learns to predict the speech in a silent talking-face video. Given the lack of annotated data, the model needs to capture a notion of the complex dynamics of the human speech signal and its dependence on the speaker's facial expressions. The problem of speech prediction can be seen as one of a learning a Phoneme-to-Viseme mapping. However, predicting the audio from a visual-only signal is an exceptionally challenging task, given the audio signal's high dimensionality (typically 16,000 samples/sec) and the complex dynamics of the human facial expressions. This is a nontrivial problem due to the wide variability of speech signals, thus approaches based on handcrafted rules may fail to generalize properly [3, 4].

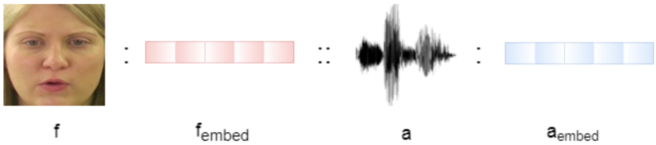


Fig. 1. Illustration of our Audio-Visual analogy approach for speech prediction in silent talking-face videos.

Even though the problem of speech-driven facial animation has been well studied, we believe the inverse problem, that is, facial-to-speech mapping remains less explored due to the reasons mentioned above. A line of work in speech prediction rely on the sequence-to-sequence models [5]. However, in the seq-to-seq models, the model needs to encode all the information needed to predict the whole audio stream into a single latent code; this makes it difficult for the model to deal with long video sequences. In [6], Ariel et al. proposed a seq-to-seq model that predicts the Linear Predictive Coding (LPC) features given a silent video. Although it is easier to generate the LPC features, the resulting voice is often unrealistic and robotic. Later in follow-up work, in [7] Ariel et al. proposed a deep learning-based model that uses Mel-spectrogram features, instead of the LPC features, for audio representation. The model also uses optical flow maps computed between each pair of consecutive video frames to obtain a low-dimensional representation of the video along with the frame sequence input. However, the absence of any recurrent units for modeling the speech signal's temporal evolution induces temporal inconsistencies. To overcome this issue, in [8] author used recurrent neural network units (RNN), specifically LSTMs (Long Short-Term Memory) [9], that are capable of learning long-term dependencies in sequential data.

Given the success of the Generative adversarial networks [10] in vision-related applications, in [11] author proposed an adversarial approach for training a generator

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

network for speech signal. More recently, in [12] Prajwal et al. proposed a variant of the Tacotron2 model [13] that makes Text-to-Speech prediction for the video-to-speech prediction task. However, the model tends to generate an overly smooth waveform resulting in a lower quality speech signal. Alternative approaches are based on autoregressive models, where each generated speech segment is fed back to the model to produce the next speech segment [14]. However, as the prediction error accumulates, these models' performance quickly deteriorates after a few time-steps, thus limiting the application of the model to the short-term horizon.

To overcome the issues mentioned earlier, single latent representation and accumulation of errors, we propose a Audio-Visual analogy (AVN) framework where each speech segment is independently decoded from the latent space. Figure 1 provides an illustration of our AVN framework that seek to learn the transformation $f : f_{embed} :: a : a_{embed}$, read as ”$f$ is to f_{embed} as a is to a_{embed}” .

2. PROPOSED MODEL

Given a video sequence, we first preprocess the audio and frame streams so that both streams consist of equal number of elements N. The frame stream is processed, using a deep learning based face alignment method [15], to only contain the (full) facial region of the speaker. Similar to prior work [14], each frame input is accompanied with few neighboring frames, that acts as a context, when given as an input to the model. Thus, in the later discussions, a frame f_i implies both the frame and the context frames. The audio stream is transformed into a sequence of Mel-spectrogram features, which makes training more computationally efficient.

We tackle the task of speech generation from a cascade perspective. Given the input frame sequence $f_{1:N}$, our model generates Mel-spectrogram features $m_{1:N}$ and subsequently uses a discriminator network to distinguish the generated sample and the ground truth samples. Our complete model for generating the Mel-spectrogram features given the video frames input is shown in Figure 2.

The model consists of a frame encoder module that embeds the video frames into a fixed-length feature vector e^t. The frame encoder module is a stack of five 3D-convolution layers with each layer followed by batch normalization and a LeakyReLU non-linearity. The obtained embedding vectors are then fed to an LSTM to learn the temporal dependencies between the obtained feature representations. Formally,

$$e^t = \phi(x_t; W_{ee}) \rightarrow h_e^t = LSTM(e^t; W_{encoder}) \qquad (1)$$

where W_{ee} and $W_{encoder}$ are the training parameters of the frame encoder module ϕ and LSTM, respectively.

To predict the speech segment, we design an AudioVisual analogy (AVN) framework for audio and visual data, based on a deep analogy-making network [16]. AVN learns the mapping from the frame input to the audio in a higher-level feature space that supports reasoning about analogies using a simple vector addition operation. The analogy based transformation can be expressed as $f_1 : e_1 :: a_t : h_e^t$ (Figure 1), that applies the transformation $f_1 \rightarrow e_1$ to h_e^t to obtain a_t, all that is done in the embedding space. In Figure 1, frame f_1 and its feature representation e_1 are used as a fixed reference input to the AVN, for all the future time steps. However, it is possible to have different reference input also. We show these results based on different reference inputs in our ablation study (Sec. 3.4). The output of the AVN network is computed as,

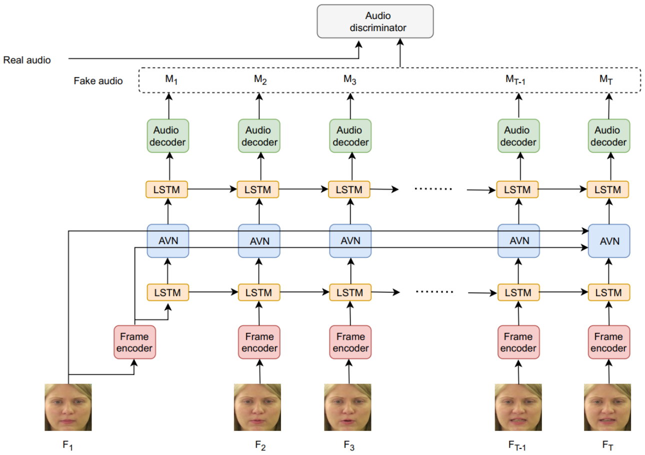


Fig. 2. Overall framework of proposed speech prediction model.

$$\begin{aligned} g^t &= AVN(f_1, e^1, h_e^t) \\ &= f_{dec}\Big(f_{enc}(h_e^t) + T(f_{img}(f_1), f_{enc}(e^1), f_{enc}(h_e^t))\Big) \end{aligned} \qquad (2)$$

where,

$$T(x, y, z) = f_{analogy}([f_{diff}(x - y), z]) \qquad (3)$$

We parameterize the functions f_{img}, f_{enc}, f_{dec}, f_{diff}, and $f_{analogy}$ using deep neural networks (see Figure 3), and $[\cdot]$ denotes concatenation along the channel dimension.

The obtained representation vector g^t is further passed through an LSTM network before being passed to a decoder network that predicts the speech waveform's Mel-spectrogram features. Formally,

$$c_e^t = LSTM(g^t; W_{decoder}) \rightarrow M_t = \gamma(c_t; W_{de}) \qquad (4)$$

where $W_{decoder}$ denotes the weights of the LSTM network, γ is the audio decoder module with weights W_{de}. M_t indicates the predicted Mel-spectrogram features at time *t*.

Since element-wise reconstruction errors (such as L_1/L_2 loss) are not appropriate for the reconstruction of audios signals or other signals with invariances, we propose to use a discriminator network, as used in Generative adversarial networks. Thus replacing the lower-level element-wise similarity with the higher-level feature-wise similarity expressed in

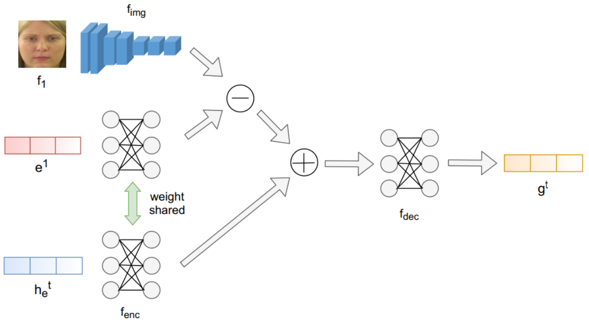


Fig. 3. Architecture of Audio-Visual analogy Network

the discriminator.

$$\mathcal{L}_{GAN} = log(Dis(M)) + log(1 - Dis(\hat{M})) \tag{5}$$

where, M and $\hat{M}$ denotes the ground truth (real) Mel-spectrogram features and the Mel-spectrogram features that are predicted (fake) by the model, respectively.

However, due to adversarial learning, the model may suffer from mode-collapse; therefore, to prevent this from happening, we add L_2 loss as a regularization term to ensure that the predicted Mel-spectrogram features match the ground truth. However, we only use this to train the generator network, not the discriminator network as this would collapse the discriminator to 0.

The overall training loss is the weighted sum of the discriminator loss and the reconstruction loss,

$$\mathcal{L} = \min_{G} \max_{D} \ \mathcal{L}_{GAN} + \lambda ||M - \hat{M}|| \tag{6}$$

where λ is a hyper-parameter that controls the relative importance of different loss terms. We use the Griffin-Lim algorithm to produce the time-domain audio from the obtained Mel-spectrograms features.

3. EXPERIMENTS

Dataset: We evaluate our model on the publicly available GRID dataset [17]. It consists of audio and video (facial) recordings of 1000 sentences spoken by 34 speakers (18 male, 16 female). Similar to prior work [6, 7, 11, 12, 14], we conducted experiments on four speakers (two male and two female) independently, namely, S1, S2, S4, and S29. We split the dataset in ratios 90–5–5% for training, validation, and test sets.

Evaluation Metrics: As in [12, 14], we use the Short-Time Objective Intelligibility (STOI) [18] and Extended Short-Time Objective Intelligibility (ESTOI) [19] and Perceptual Evaluation of Speech Quality (PESQ) [20] metrics for quantitative evaluation. For qualitative comparison, we compare the Mel-spectrogram and the predicted audio waveform with respect to the ground truth speech signal. Besides, we performed an extensive user study to investigate our generated results' audio qualities compared with the state-of-the-art Lip2Wav model [12] and the recently proposed VAE-based model [14]. We provide a thorough ablation study to quantitatively assess the effect of the different design choices of the model.

3.1. Quantitative Evaluation

Table 1 shows the results of our proposed method against the baseline (proposed model without AVN module) & state-of-the-art methods. We see that the models based on Mel-spectrograms features perform comparatively better than the models that operate on raw audio or LPC features. Also, we observed that the generative model [14] outperforms the discriminative models based on the quality of speech generated (as indicated by a higher PESQ score). This may be because generative modeling gives the model the ability to reason about uncertainties in decision-making via probability distributions. Our proposed model performs better in terms of the STOI and ESTOI metrics that measure speech intelligibility than the other baseline models.

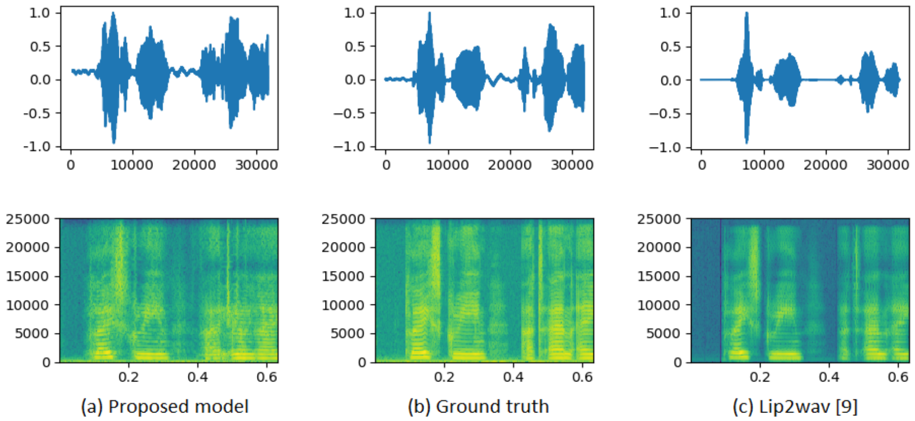
Table 1. Quantitative evaluation: Mean STOI, ESTOI, and PESQ results for all the speakers.

Model	STOI	ESTOI	PESQ
Vid2Speech [6]	0.491	0.335	1.734
Lip2AudSpec [8]	0.513	0.352	1.673
Konstantinos et. al. [11]	0.564	0.361	1.684
Ephrat et al. [7]	0.659	0.376	1.825
Lip2Wav [12]	0.731	0.535	1.772
Yadav et al. [14]	0.724	0.540	**1.932**
Proposed model (w/o AVN)	0.432	0.169	1.332
Proposed model	**0.736**	**0.545**	1.776

As the main contribution of our work is the introduction of the Audio-Visual Analogy (AVN) framework, therefore, we also quantitative evaluate if the higher performance of the proposed model is due to the proposed analogy framework or the model is simply learning the direct mapping from visual to audio inputs ignoring the AVN network completely. To verify that, in Figure 2, we remove the AVN network and train the network without it. As we can see, the quantitative results drop significantly if we do not use the AVN module. In fact, it is lower than any of the compared baseline models.

3.2. Qualitative Evaluation

For qualitative evaluation, we compare the Mel-spectrogram and the audio waveform generated by our model and Lip2Wav model, for given silent video input, in Figure 4. We see that



(a) Proposed model

(b) Ground truth

(c) Lip2wav [9]

Fig. 4. Qualitative comparison of proposed model and the Lip2Wav model based on the generated audio and Mel-spectrogram.

the Lip2Wav model generates overly smooth waveforms, whereas our proposed model generates results that closely match the ground truth. We request reviewers to listen to the generated audio samples shared at url[1]

3.3. User Study

We performed an extensive user study to investigate the quality of the generated speech signal subjectively. The study included samples generated by state-of-the-art Lip2Wav model [12] and the VAE-based model [14]. We showed randomly chosen 30 different samples generated by each model to 20 participants and asked them to rate the samples on a scale of 1 to 5 (a higher score is better), based on the two different criteria: (i) realism (including synchronization between the generated audio and lip movements) and (ii) correct transcription. The subjects were asked to transcribe the generated audio without having access to any ground truth information. We then compute the percentage of correctly transcribed words. The results are shown in Table 2. We see that our proposed model outperforms the other two models on both measures.

Table 2. **User study**: Mean opinion score

Model Metric	[14]	[12]	Ours
Realism	2.935	2.809	**3.307**
Transcription (%)	75.40	63.74	**76.51**

3.4. Ablation study

In Figure 2, the AVN network takes three inputs, that is, first video frame (f_1), its latent representation vector (e^1), and, embedding vector (h_e^t). Thus the first frame and its latent representation act as a fixed reference for analogy computation at every time step. We experimented with different choices of reference inputs, the quantitative results for all such cases are reported in Table 3.

Table 3. Quantitative Evaluation (Ablation)

Reference Input	STOI	ESTOI	PESQ
Random frame (f_r)	0.450	0.145	1.334
Random video (f_v)	0.542	0.293	1.400
First frame w/o disc (f_1)	0.540	0.282	1.388
First frame w/ disc (f_1)	**0.736**	**0.545**	**1.776**

We observed that with random frame input, the generated speech is often incomprehensible. To provide a longer context to let the model to learn from, we provided whole random video input as a reference. We observed slight improvement, but the generated speech still sounds unrealistic and was often noisy. It shows that the proposed AVN network works efficiently when the reference input is kept fixed during training. We observed best results when the first frame of the training video was used as a fixed reference, along with the adversarial training approach. With simple element-wise L_2 loss, the resulting speech often fades out in between the video frames. The video results for all the above cases are shared on the project page link given below.

4. CONCLUSION

We presented a novel architecture for speech prediction in silent talking-face videos that outperforms current state-of-the-art methods across several widely used benchmarks. We proposed a novel Audio-Visual analogy framework for Phoneme-to-Viseme mapping from the frame input to the audio via higher-level feature space that supports reasoning about analogies using a simple vector addition operation. Our user study demonstrated that our proposed model produces speech signals that are more realistic, accurate, and closely match the ground truth.

[1] https://sites.google.com/view/speech-prediction

5. REFERENCES

[1] Kenton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423*, p. 4171–4186, 2019.

[2] Praveer Singh Spyros Gidaris and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," *In International Conference on Learning Representations*, 2018.

[3] Thomas Le Cornu and Ben Milner, "Reconstructing intelligible audio speech from visual speech features," *sixteenth annual conference of the international speech communication association*, 2015.

[4] Thomas Le Cornu and Ben Milner, "Generating intelligible audio speech from visual speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25.9, pp. 1751–1761, 2017.

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, "Sequence to sequence learning with neural networks," *In Advances in neural information processing systems*, pp. 3104–3112, 2014.

[6] Ariel Ephrat and Shmuel Peleg, "Vid2speech: speech reconstruction from silent video," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*, 2017.

[7] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg, "Improved speech reconstruction from silent video," *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017.

[8] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2516–2520, 2018.

[9] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation 9.8*, pp. 1735–1780, 1997.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[11] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," *Interspeech*, September 2019.

[12] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13796–13805, 2020.

[13] Jonathan Shen et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*, 2018.

[14] Ravindra Yadav, Ashish Sardana, Vinay P. Namboodiri, and Rajesh M. Hegde, "Speech prediction in silent videos using variational autoencoders," *arXiv preprint arXiv:2011.07340*, 2020.

[15] Adrian Bulat and Georgios Tzimiropoulos, "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.

[16] Scott E. Reed, Yi Zhang, Yuting Zhang, and Honglak Lee, "Deep visual analogy-making," *In Advances in neural information processing systems 28*, pp. 1252–1260, 2015.

[17] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America 120, no. 5*, pp. 2421–2424, 2006.

[18] Cees Taal, Richard Hendriks, R. Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *In 2010 IEEE international conference on acoustics, speech and signal processing*, pp. 4214–4217, 2010.

[19] Jesper Jensen and Cees H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *In IEEE/ACM Transactions on Audio, Speech, and Language Processing 24.11*, pp. 2009–2022, 2016.

[20] Antony Rix, John Beerends, Michael Hollier, and Andries Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *In IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221). Vol. 2. IEEE*, p. 749–752, 2001.