Generating Talking Face Landmarks from Speech

Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan ⋆

University of Rochester, 500 Joseph C. Wilson Blvd., Rochester, NY 14627, USA
`{eeskimez}@ur.rochester.edu,`
`{ross.maddox,chenliang.xu,zhiyao.duan}@rochester.edu`

Abstract. The presence of a corresponding talking face has been shown to significantly improve speech intelligibility in noisy conditions and for hearing impaired population. In this paper, we present a system that can generate landmark points of a talking face from an acoustic speech in real time. The system uses a long short-term memory (LSTM) network and is trained on frontal videos of 27 different speakers with automatically extracted face landmarks. After training, it can produce talking face landmarks from the acoustic speech of unseen speakers and utterances. The training phase contains three key steps. We first transform landmarks of the first video frame to pin the two eye points into two predefined locations and apply the same transformation on all of the following video frames. We then remove the identity information by transforming the landmarks into a mean face shape across the entire training dataset. Finally, we train an LSTM network that takes the first- and second-order temporal differences of the log-mel spectrogram as input to predict face landmarks in each frame. We evaluate our system using the mean-squared error (MSE) loss of landmarks of lips between predicted and ground-truth landmarks as well as their first- and second-order temporal differences. We further evaluate our system by conducting subjective tests, where the subjects try to distinguish the real and fake videos of talking face landmarks. Both tests show promising results.

Keywords: Visual generation, face landmarks, audio-visual models, LSTM

1 Introduction

Speech is a natural way of communication, and understanding speech is essential in daily life. The auditory system, however, is not the only sensory system involved in understanding speech. The visual cues from a talker's face and articulators (lips, teeth, tongue) are also important for speech comprehension. Trained professionals are able to understand what is being said by purely looking at lip movements (lip reading) [9]. For ordinary people and the hearing impaired

---

population, the presence of visual signals of speech has been shown to significantly improve speech comprehension, even if the visual signals are synthetic [14]. The benefits of adding the visual speech signals are more pronounced when the acoustic signal is degraded, due to background noise, communication channel distortion, and reverberation.

In many scenarios such as telephony, however, speech communication is still acoustical. The absence of the visual modality can be due to the lack of cameras, the limited bandwidth of communication channels, or privacy concerns. One way to improve speech comprehension in these scenarios is to synthesize a talking face from the acoustic speech in real time at the receiver's side. A key challenge of this approach is to make sure that the generated visual signals, especially the lip movements, well coordinate with the acoustic signals, as otherwise more confusions will be introduced.

In this paper, we propose to use a long short-term memory (LSTM) network to generate landmarks of a talking face from acoustic speech. This network is trained on frontal videos of 27 different speakers of the Grid audio-visual corpus [6], with the face landmarks extracted using the Dlib toolkit [13]. The network takes the first- and second-order temporal differences of the log-mel spectra as the input, and outputs the x and y coordinates of 68 landmark points. To help the network capture the audio-visual coordination instead of the variation of face shapes across different people, we transform all training landmarks to those of a mean face across all talkers in the training set. After training, the network is able to generate face landmarks from an unseen utterance of an unseen talker. Objective evaluations of the generation quality are conducted on the LDC Audiovisual Database of Spoken American English dataset [18], which will be referred as the LDC dataset in the remaining of the paper. Subjective evaluation is also conducted to ask evaluators to distinguish speech videos with ground-truth and generated landmarks. Both the objective and subjective evaluations achieve promising results. The code and pre-trained talking face models are released to the community[1]

The remaining of the paper is structured as follows: Section 2 describes the related work. Section 3 describes the data and pre-processing steps. The architecture of the network is described in Section 4. Objective and Subjective evaluations are presented in Section 5. Finally, Section 6 concludes the paper.

2   Related Work

Generating a talking head automatically has been a great interest in the research community. Some researchers focused on text-driven generation [23, 22, 10, 3]. These methods map phonemes to talking face images. Compared to text, voice signals are surface-level signals that are more difficult to parse. Besides, voices of the same text show large variations across speakers, accents, emotions, and the recording environments. On the other hand, speech signals provide richer

---

[1] http://www.ece.rochester.edu/projects/air/projects/talkingface.html

cues for generating natural talking faces. For text, any plausible face image sequence is sufficient to establish natural communication. For speech, it must be a plausible sequence that matches the speech audio. Therefore, text-driven generation and speech-driven generation are different problems and may require different approaches.

There exist a few approaches to speech-driven talking face generation. Early work in this field mostly used Hidden Markov Models (HMM) to model the correspondence between speech and facial movements [2, 4, 8, 7, 24, 20, 25]. One of the notable early work, Voice Puppetry [2], proposed an HMM-based talking face generation that is driven by only speech signal. In another work, Cosker et al. [8, 7] proposed a hierarchical model that animates sub-areas of the face independently from speech and merges them into a full talking face video. Xie et al. [24] proposed coupled HMMs (cHMMs) to model audio-visual asynchrony. Choi et al. [4] and Terissi et. al [20] used HMM inversion (HMMI) to estimate the visual parameters from speech. Zhang et al. [25] used a DNN to map speech features into HMM states, which further maps to generated faces.

In recent years, a few DNN-based approaches have also been proposed. Suwajanakorn et al. [19] designed an LSTM network to generate photo-realistic talking face videos of a target identity directly from speech. Their system requires several hours of face videos of the specific target identity, which greatly limits its application in many practical scenarios. Chung et al. [5] proposed a convolutional neural network (CNN) system to generate a photo-realistic talking face video from speech and a single face image of the target identity. Compared to [19], the reduction from several hours of face videos to a single face image for learning the target identity is a great advance.

While end-to-end speech-to-face-video generation is very useful in many scenarios, the main limitation of this approach is the lack of freedom for further manipulation of the generated face video. For example, within a generated video, one may want to vary the gestures, facial expressions, and lighting conditions, all of which can be relatively independent of the content of the speech. These end-to-end systems cannot accommodate such manipulations unless they can take these factors as additional inputs. However, that would significantly increase the amount and diversity of data required for training the systems.

A modular design that separates the generation of key parameters and the fine details of generated face images is more flexible for such manipulations. Ideally, the key parameters should just respond to the speech content, while the fine details should incorporate all other non-speech-content related factors. Pham et al. [16] adopted a modular design: the system first maps speech features to 3D deformable shape and rotation parameters using an LSTM network, and then generates a 3D animated face in real-time from the predicted parameters. In [17], they further improved this approach by replacing speech features with raw waveforms as the input and replacing the LSTM network with a convolutional architecture. However, compared to face landmarks used in our proposed approach, these shape and rotation parameters are less intuitive, and the mapping from these parameters to a certain gesture or facial expression is less clear. In
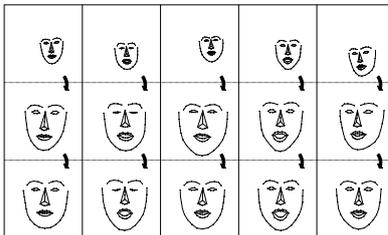
Fig. 1. Examples of extracted face landmarks from the training talking face videos. Certain landmarks are connected to make the shape of the face easier to recognize. The first row shows unprocessed landmarks of five unique talkers. The second row shows their landmarks after outer-eye-corner alignment. The third row shows their landmarks after alignment and the removal of identity information.

addition, the landmarks generated by our system are for a normalized mean face instead of a certain target identity. This also helps remove factors that are not directly related to the voice.

3   Proposed Method

In this section, we describe our method to generate talking face landmarks. First, we extract face landmarks and align them across different speakers and transform their shapes into the mean shape to remove the identity information. We extract the first and second order temporal difference of the log-mel spectrogram and use them as the input to our system. Finally, we train an LSTM network to generate the face landmarks from the speech features.

3.1   Training Data & Feature Extraction

We employ the audio-visual GRID dataset [6] to train our system. There are in total 16 female and 18 male native English speakers, each of which has 1000 utterances that are 3 seconds long. The sentences are structured to contain a command, a color, a preposition, a letter, a digit, and an adverb, for example, *"set blue at C5 please"*.

The videos are provided in two resolutions, low (360x288) and high (720x576). In this work, we use the high-resolution videos. The videos use a frame rate of 25 frames per second (FPS), resulting in 75 frames for each video. The speech audio signal is extracted from the video with a sampling rate of 44.1 kHz.

We extract 68 face landmark points (x and y coordinates) using the DLIB library [13] from each frame for each video in the dataset. Examples are shown

in the first row of Figure 1. We calculate 64 bin log-mel spectra of the speech signal covering the entire frequency range using a 40 ms hanning window without any overlap to match the video frame rate. We then calculate the first- and second-order temporal differences of the log-mel spectra and use them as the input (128-d feature sequence) to our network. We experimented using log-mel spectrogram with and without its first- and second-order derivatives as input to our network. The generated mouth for many speech utterances in these two setups, however, were almost always open even in silent segments, and the lip movements were less prominent than the current system. The first- and second-order temporal differences of the log-mel spectrogram may show less variations on the same syllable uttered by different speakers, and the mismatch problem is less pronounced.

3.2   Face Landmark Alignment

Since the talking face may appear in different regions with different sizes in different videos, we need to align them to reduce the complexity of training data. To do so, we follow the procedure described in [15] to simply pin the two outer corners of the eyes in the first frame of each video to two fixed locations, (180, 200) and (420, 200) in the image coordinate system, through an 6 DOF affine transformation. We then transform all of the landmarks in all video frames with the same transformation. Note that we do not align each video frame using their own affine transformation separately because we find that the eye-corner-based alignment is sensitive to eye blinks, which often results in zoom in/out effects of the transformed face shape. Also note that our approach assumes that the head does not move significantly within a video, as otherwise, the same affine transformation would not be able to align faces in different frames. The second row of Figure 1 shows several examples of the aligned face landmarks.

3.3   Removing Identity Information from Landmarks

After alignment, faces of different speakers are of a similar size and general location; however, their shapes are still different as well as their mouth locations. This identity-related variation may pose challenges to the network for capturing the relation between speech and lip movement, especially when the amount and diversity of training data are small. Therefore, we propose to remove the identity information from the landmarks before training the network.

To do so, we apply the following steps. First, we calculate the mean face shape by averaging all aligned landmark locations across the entire training set. Second, for each face landmark sequence, we calculate the affine transform between the mean shape and the first frame of the sequence. Third, we calculate the difference between the current frame and the first frame and multiply with the scaling coefficients obtained from the second step with the result obtained in the third step. Finally, we add the mean shape to results obtained in fourth step to obtain the face landmark sequence that has no identity. The third row of Figure 1 shows several examples of landmarks with the identity removed.
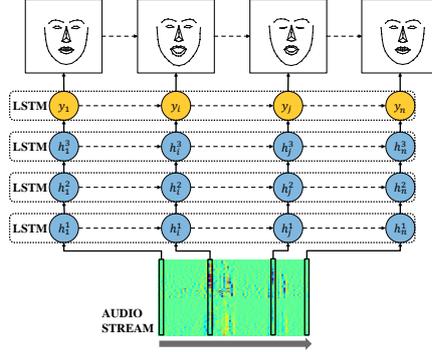
Fig. 2. The LSTM network architecture for generating landmarks of a talking face from the first and second order temporal differences of the log-mel spectrogram. h_t^l are the hidden layers, where t is the time step and l is the hidden layer index. y_t are the output face landmarks for the time step t.

3.4   LSTM Network

Our proposed network, as shown in Figure 2, uses four long short-term memory (LSTM) [12] layers with a sigmoid activation function. At each time step, the input to the network is the first and second order temporal differences of the log-mel spectra of the current and the previous N frames. This provides short-term contextual information. The output is the predicted the x and y coordinates of face landmarks of the current frame (if no delay is added) or a previous frame (if a delay is added as described below). The reason for adding delay is because lips often move before the sound is produced. With a little delay, the network is able to "hear into the future" and can better prepare for those lip movements. The generated lip movements tend to be smoother. The amount of delay we introduce is between 1 (40 ms) and 5 frames (200 ms). This turns out to be enough for good generation results and is still tolerable in real-time speech communication.

During training, we use dropout between each layer and between recurrent connections, with a rate of 0.2. We use Adam optimizer to train our network. The training sequences are all 75 frames long. We set the batch size to 128 sequences and the learning rate to 0.001. Our network minimizes the following mean squared error (MSE) objective function J_{MSE},

$$J_{MSE} = \frac{1}{N} \sum_t^N \| \mathbf{s}_t - \widehat{\mathbf{s}}_t \|^2 \,, \tag{1}$$

where $\mathbf{s}$ and $\widehat{\mathbf{s}}$ are the x and y coordinates of ground-truth (GT) and predicted (PD) face landmarks sequences, respectively. N is the number of samples.

Finally, the predicted landmarks are further processed in order to fix the eye corner points to fixed points as described in Section 3.2, which produces more stable talking face landmarks.
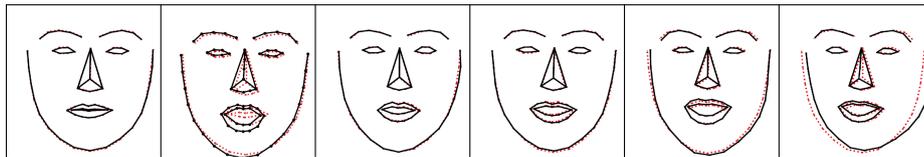
Fig. 3. Pair-wise comparison between ground-truth landmarks (black solid lines) and generated landmarks (red dotted lines) on unseen talkers and sentences. The second image shows a failure case for "oh" sound.

Table 1. Objective evaluation results for different system configurations. The models are named according to the amount of delay and contextual information. For example, "D40-C5" describes a model trained with 40 ms delay and 5 frames of context. The lower value means better results, where the ideal result is zero.

	RMSE	RMSE First Diff	RMSE Second Diff
D0-C3	0.0954	0.0045	0.0073
D0-C5	0.0945	0.0042	0.0071
D40-C3	0.0932	0.0039	0.0068
D40-C5	**0.0921**	**0.0032**	**0.0065**
D80-C3	0.0946	0.0044	0.0072
D80-C5	0.0944	0.0043	0.0069

Due to causality constraints, the bidirectional LSTM network is not considered in our experiments. We have also experimented with fully connected architecture instead of LSTM. However, the resulting face landmarks often show sudden jumps between frames, which looks unnatural. This is due to not having temporal connections in the architecture.

4   Experiments

We conduct our objective and subjective evaluations on a totally different audio-visual dataset, the LDC dataset [18]. It contains 10 female and 4 male speakers, where each speaker provides 94 samples, totaling to 1316 utterances. The duration of the videos is arbitrary, and the resolution of the samples are 720x480. Since the frame rate of the videos is higher than the Grid dataset used to train our system, we resampled the videos to the same frame rate of 25 FPS. The vocabulary of the LDC dataset is much larger than that of the Grid dataset. There are various words and sentences from TIMIT sentences [11], Northwestern University Auditory Test No. 6 [21], and Central Institute for the Deaf (CID) Everyday Sentences [1]. The audio stream is provided at 48 kHz sampling rate, which we down-sampled to 44.1 kHz. Figure 3 shows examples of ground-truth and generated face landmarks in the first and second row, respectively. Examples of generated videos are publicly accessible[2].

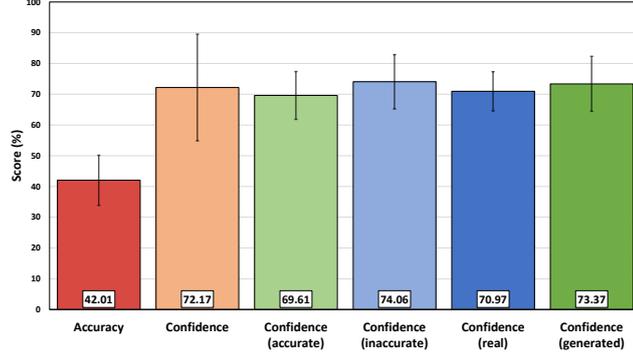[2] http://www.ece.rochester.edu/projects/air/projects/talkingface.html

Fig. 4. Subjective evaluation results. The mean accuracy score and its standard deviation are averaged over all subjects. The mean confidence scores and their standard deviations are averaged over all subjects and videos.

4.1 Objective Evaluation

We report the root-mean-squared error (RMSE) results between the ground-truth (GT) and predicted (PD) face landmarks according to Equation 1. The landmarks scale are between 0 and 1, therefore RMSE value of 0.01 approximately equivalent to 1% error. We also report the RMSE of the first and second order temporal differences of the GT and PD face landmarks to assess the movement. We report the results in Table 1. These results serve as a way of model selection. The best model according to these results is the model that has 40 ms delay and 5 frames of context information (D40-C5). We selected this model to conduct the subjective evaluations, which are described in the next section.

4.2 Subjective Evaluation

We conducted subjective tests to determine if our system can generate realistic face landmarks. 17 naive volunteer evaluators who are graduate students at the University of Rochester participated in the test. The test presented 25 real landmark videos and 25 generated landmark videos in a randomized order to each evaluator and asked the evaluator to label whether each presented video was real or fake. Each video was presented twice in the randomized video sequence. The real landmark videos were created from randomly selected LDC videos. Landmarks were extracted and aligned, and the identity information was removed, according to Section 3. Fake videos were generated from the audio signals of another 25 randomly selected LDC videos. The GT landmarks were noisy; hence we also added Gaussian noise to the PD landmarks to make them look more like the GT landmarks. In addition to a binary decision, the evaluators were asked to report their confidence level of each decision, between 0 and 100 percent.

The mean accuracy score of the evaluators are shown in Figure 4, along with the overall mean confidence score and the mean confidence score for the correctly

and incorrectly predicted samples. The results show that the evaluators struggled to distinguish real and generated samples, as the accuracy is 42.01% which is even below chance (50%). Another interesting observation of this test is that the mean confidence score for accurately determined samples is lower than that for inaccurately determined samples. This suggests that the evaluators had a higher classification accuracy when they were more cautious. Another outcome is that the mean confidence score on answers for generated samples is more than the confidence score on answers for the ground truth samples.

5   Conclusion

In this work, we present a method to generate talking face landmarks from speech. We extract face landmarks from the Grid corpus, align them across different speakers, and transform their shapes into the mean shape to remove the identity information. The LSTM network predicts the face landmarks from the first and second order temporal differences of the log-mel spectrogram from any arbitrary voice. The network can produce face landmarks that look natural for the given speech input. The main limitation of this network is that it cannot produce "oh" and "oo" sounds right. We plan to balance the phonetic content of the dataset to enable the network to produce all phonemes correctly in our future work. We will evaluate the system against noise, and improve it to obtain a noise-resilient system in our future work. We report objective and subjective evaluation results that are promising. We release the code and example videos to the community.

References

1. Blamey, P.J., Pyman, B.C., Clark, G.M., Dowell, R.C., Gordon, M., Brown, A.M., Hollow, R.D.: Factors predicting postoperative sentence scores in postlinguistically deaf adult cochlear implant patients. Annals of Otology, Rhinology & Laryngology 101(4), 342–348 (1992)
2. Brand, M.: Voice puppetry. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 21–28. ACM Press/Addison-Wesley Publishing Co. (1999)
3. Cassidy, S., Stenger, B., Dongen, L.V., Yanagisawa, K., Anderson, R., Wan, V., Baron-Cohen, S., Cipolla, R.: Expressive visual text-to-speech as an assistive technology for individuals with autism spectrum conditions. Computer Vision and Image Understanding 148, 193 – 200 (2016)
4. Choi, K., Luo, Y., Hwang, J.N.: Hidden markov model inversion for audio-to-visual conversion in an mpeg-4 facial animation system. Journal of VLSI signal processing systems for signal, image and video technology 29, 51–61 (2001)
5. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? arXiv preprint arXiv:1705.02966 (2017)
6. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America 120(5), 2421–2424 (2006)

7. Cosker, D., Marshall, D., Rosin, P.L., Hicks, Y.: Speech driven facial animation using a hidden markov coarticulation model. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR). vol. 1, pp. 128–131. IEEE (2004)
8. Cosker, D., Marshall, D., Rosin, P., Hicks, Y.: Video realistic talking heads using hierarchical non-linear speech-appearance models. Mirage, France 147 (2003)
9. Dodd, B.E., Campbell, R.E.: Hearing by eye: The psychology of lip-reading. Lawrence Erlbaum Associates, Inc (1987)
10. Fan, B., Wang, L., Soong, F.K., Xie, L.: Photo-real talking head with deep bidirectional lstm. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4884–4888. IEEE (2015)
11. Garofalo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: The darpa timit acoustic-phonetic continuous speech corpus cdrom. Linguistic Data Consortium (1993)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
13. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research 10, 1755–1758 (2009)
14. Maddox, R.K., Atilgan, H., Bizley, J.K., Lee, A.K.: Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. eLife 4 (2015)
15. Mallick, S.: Face morph using opencv c++ / python (2016), `http://www.learnopencv.com/face-morph-using-opencv-cpp-python/`
16. Pham, H.X., Cheung, S., Pavlovic, V.: Speech-driven 3d facial animation with implicit emotional awareness: a deep learning approach. In: The 1st DALCOM workshop, CVPR (2017)
17. Pham, H.X., Wang, Y., Pavlovic, V.: End-to-end learning for 3d facial animation from raw waveforms of speech. arXiv preprint arXiv:1710.00920 (2017)
18. Richie, S., Warburton, C., Carter, M.: Audiovisual database of spoken american english. Linguistic Data Consortium (2009)
19. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) 36(4), 95 (2017)
20. Terissi, L.D., Gómez, J.C.: Audio-to-visual conversion via hmm inversion for speech-driven facial animation. In: Brazilian Symposium on Artificial Intelligence. pp. 33–42. Springer (2008)
21. Tillman, T.W., Carhart, R.: An expanded test for speech discrimination utilizing cnc monosyllabic words: Northwestern university auditory test no. 6. Tech. rep., Northwestern University Evanston Auditory Research Lab (1966)
22. Wan, V., Anderson, R., Blokland, A., Braunschweiler, N., Chen, L., Kolluru, B., Latorre, J., Maia, R., Stenger, B., Yanagisawa, K., et al.: Photo-realistic expressive text to talking head synthesis. In: INTERSPEECH. pp. 2667–2669 (2013)
23. Wang, L., Han, W., Soong, F.K., Huo, Q.: Text driven 3d photo-realistic talking head. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
24. Xie, L., Liu, Z.Q.: A coupled hmm approach to video-realistic speech animation. Pattern Recognition 40, 2325–2340 (2007)
25. Zhang, X., Wang, L., Li, G., Seide, F., Soong, F.K.: A new language independent, photo-realistic talking head driven by voice only. In: Interspeech. pp. 2743–2747 (2013)