Text2Video: Text-driven Talking-head Video Synthesis with Phonetic Dictionary

Sibo Zhang, Jiahong Yuan, Miao Liao, Liangjun Zhang

Baidu Research, USA

sibozhang1@gmail.com, jiahong.yuan@gmail.com, miao.liao@gmail.com,
liangjunzhang@baidu.com

Abstract

With the advance of deep learning technology, automatic video generation from audio or text has become an emerging and promising research topic. In this paper, we present a novel approach to synthesize video from the text. The method builds a phoneme-pose dictionary and trains a generative adversarial network (GAN) to generate video from interpolated phoneme poses. Compared to audio-driven video generation algorithms, our approach has a number of advantages: 1) It only needs a fraction of the training data used by an audio-driven approach; 2) It is more flexible and not subject to vulnerability due to speaker variation; 3) It significantly reduces the preprocessing, training and inference time. We perform extensive experiments to compare the proposed method with state-of-the-art talking face generation methods on a benchmark dataset and datasets of our own. The results demonstrate the effectiveness and superiority of our approach.

Index Terms: talking head video generation, text driven, multimodal synthesis, phoneme-pose dictionary

1. Introduction

With the advance of deep learning technology, automatic video generation from audio (*speech2video*) or text (*text2video*) has become an emerging and promising research topic [1, 2, 3]. It introduces exciting opportunities for applications such as AI news broadcasts, video synthesis, and digital humans.

Speech2Video models are trained to map from speech to video. Because of speaker variability in speech, *Speech2Video* models need to be trained on a large amount of data, and they are not robust to different speakers. It is also less flexible to use speech as input compared to text. Furthermore, most previous methods that generate video from speech are based on LSTM to learn audio information. However, LSTM-based methods have some limitations: 1) The network needs a lot of training data. 2) The voice of a different person degrades output motion quality. 3) We can not manipulate motion output such as changing speaker attitude since the network is a black box on what is learned. Compared to audio-based methods, text-based methods have advantages. We here define *Text2Video* as a task of synthesizing talking-head video from any text input. The video generated from a text-based method should be agnostic to the voice identity of a different person.

In this paper, we propose a novel method to generate video from text. The technique builds a phoneme-pose dictionary and trains a generative adversarial network (GAN) to generate video from interpolated phoneme poses. Forced alignment is employed to extract phonemes and their timestamps from training data to build a phoneme-pose dictionary. We applied the method to both English and Mandarin Chinese. To demonstrate our the effectiveness of approach, we conducted experiments on a number of public and private datasets. Results showed that our method achieved higher overall visual quality scores compared to state-of-the-art systems.

The main contributions of this paper are summarized as follows: 1) We propose a novel pipeline of generating talking-head speech videos from any text input, including English, Chinese, numbers, and punctuation. The inference time is as fast as ten frames per minute on our pipeline. 2) We develop an automatic pose extraction method to build a phoneme - pose dictionary from any video, online or purposely recorded. With only 44 words or 20 sentences, we can build a phoneme - pose dictionary that contains all phonemes in English. 3) To generate natural pose sequences and videos, we introduce an interpolation and smoothness method and further utilize a GAN-based video generation network to convert sequences of poses to photo-realistic videos.

2. Related Works

Text-Driven Video Generation. There are some earlier works on visual speech synthesis (from text). Ezzat [4] introduced MikeTalk, a text-to-audiovisual speech synthesizer that converts input text into an audiovisual speech stream. Taylor [5] proposed a method for automatic redubbing of video that exploited the many-to-many mapping of phoneme sequences to lip movements modeled as dynamic visemes. Text-based Mouth Editing [6] is a method to overwrite an existing video with new text input. The method conducts a viseme search to find video segments with mouth movements matching the edited text. However, their synthesis approach requires a re-timed background video as input and their phoneme retrieval is agnostic to the mood in which the phoneme was spoken.

Audio-driven video generation. Audio-driven Video Synthesis (Speech2Video) is to drive movements of human bodies with input audio. Much exciting work has been done in this area. For example, SythesisObama [1] focused on synthesizing a talking-head video by driving mouth motion with speech using an RNN. A mouth sequence was first generated via texture mapping and then pasted onto an existing human speech video. However, SythesisObama needs approximately 17 hours of training data for one person, so it is not scalable. [7] utilized facial landmarks to generate video from identity image and audio signal. [8] generated high-quality talking face videos using disentangled audio-visual representation. Wang [9] proposed a GAN-based network based on the attentional multiple representations to synthesize talking head video from a given speech. Taylor [10] introduce a deep learning approach using sliding window regression for generating realistic speech animation. However, their animation predictions are made in terms of the reference face AAM parameterization re-targeting to a character, which introduces a potential source of errors.

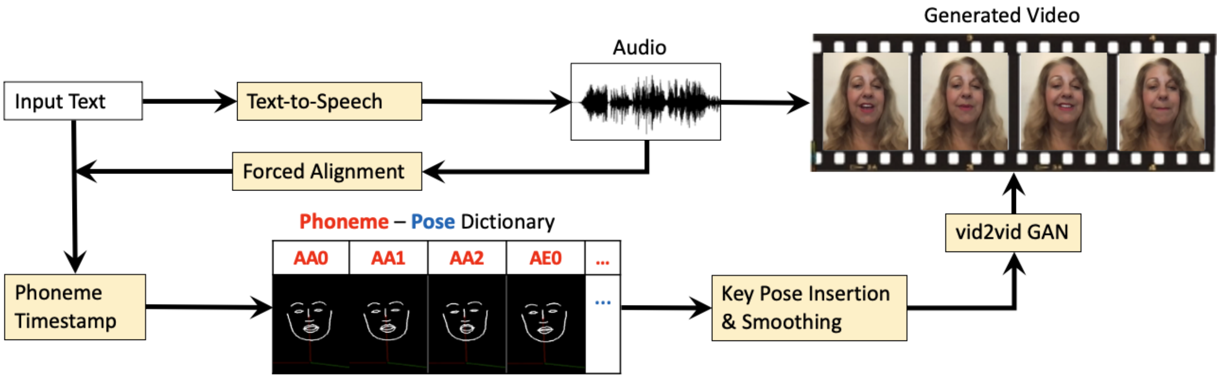

Figure 1: *Pipeline of Text2Video system. Given input text, we generate audio from the text. Then we apply forced alignment to get phoneme timestamps and lookup poses in a phoneme-pose dictionary. Next, we apply the key pose interpolation and smooth module to get a sequence of poses. In the end, we utilize a modified GAN to generate the final output video.*

Ginosar [11] proposed a method to learn individual styles of speech gesture in two stages. However, final generated videos from their rendering stage have a few artifacts. Thies [2] developed a 3D face model by audio and rendered the output video using a technique called neural rendering [12]. They proposed Audio2ExpressionNet, a temporal network architecture to map an audio stream to a 3D blend shape basis representing person-specific talking styles. This method needs a long time to train. Previously, mouth movement synthesis is mostly deterministic: given a pronunciation, the mouth's movement or shape is similar across different persons and contexts. Alternately, Liao [3] proposed a novel 2-stage pipeline of generating an audio-driven virtual speaker with full-body movements. Their method was able to add personalized gestures in the speech by interpolating key poses. They also utilized 3D skeleton constraints to guarantee that the final video is physically plausible. However, this method is audio-based and has the limitations as mentioned earlier.

3. Method

3.1. Text2Video Framework

As shown in figure 1, the input to our system is text, which can be English, Chinese, numbers, and punctuation. The output is generated video of a talking human. Given an input text, we use TTS to generate speech from the text. Then we apply forced alignment to obtain phoneme timestamps, and lookup phoneme poses in our phoneme-pose dictionary. Next, we apply the key pose interpolation and smooth module to generate a sequence of poses. Finally, we use GAN to generate videos. Our method contains two key components: building a phoneme-pose dictionary from training data (audio and video of speech) and training a model to generate video from phoneme poses.

3.2. Build Phoneme-Pose Dictionary

Phonemes are the basic units of the sound structure of a language. They are produced with different positions of the tongue and lips, for example, with lips rounded (e.g. /u/) or spread (e.g. /i/), or wide open (e.g., /a/) or closed (e.g., /m/). English has 40 phonemes if we don't count lexical stress. The phonemes are listed in supplemental materials Appendix 1. There are three levels of lexical stress in English: primary stress, secondary stress, and unstress. Stress may influence the position of the lips in speech production. For example, the vowel 'er' in the word *permit* is stressed when the word is a noun and is unstressed when it is a verb. The mouth is slightly more open when pronouncing the stressed 'er'. Therefore, we distinguish stress in the English phoneme-pose dictionary. For Mandarin Chinese, we use initials and finals as the basic units in the phoneme-pose dictionary. This is because phonemes in the finals in Chinese are more blended and don't have a clear boundary between each other [13]. Appendix 2 is a list of Mandarin initials and finals. We build a phoneme-pose dictionary for English and Mandarin Chinese, respectively, mapping from phonemes to lip postures extracted from a speech production video.

Key Pose Extraction. First, we use Openpose [14] to extract key poses from training videos. Then we build up the phoneme-pose dictionary from our phoneme extraction pipeline described below.

Phoneme Extraction. We employed the P2FA aligner [15] to determine phonemes and their time positions in an utterance. The task requires two inputs: audio and word transcriptions. The transcribed words are mapped into a phone sequence in advance using a pronouncing dictionary or grapheme to phoneme rules. Phone boundaries are determined by comparing the observed speech signal and pre-trained, Hidden Markov Model (HMM) based acoustic models. In forced alignment, the speech signal is analyzed as a successive set of frames (e.g., every 10 ms). The alignment of frames with phonemes is determined by finding the most likely sequence of hidden states (which are constrained by the known sequence of phonemes derived from transcription) given the observed data and the acoustic models represented by the HMMs. Then, we store a sequence of poses for each phoneme in the dictionary based on the alignment. The width of the phoneme-poses is determined based on the dataset video frame rate and average speaking rate.

3.3. Text to Speech

We use Baidu TTS to generate audio from text input. The system's default female and male voices are used. For personalized video generation, one can use any technique to generate a voice of his/her own choice. The voice of a different person will not affect the generated video quality of our method.

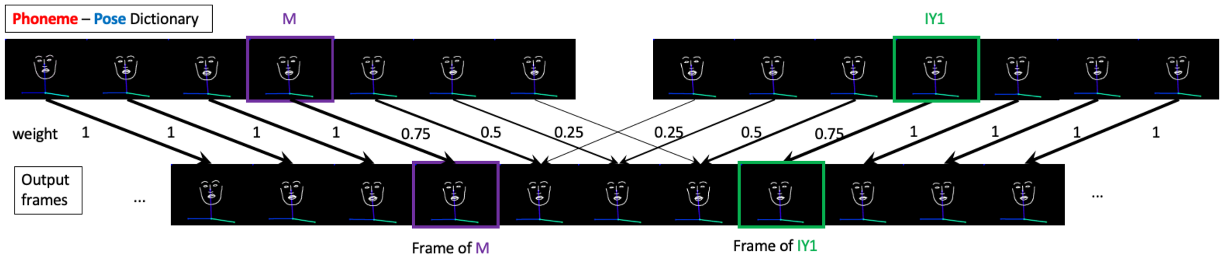

Figure 2: *Interpolation method. To generate the output sequence including "me" or "M IY1" in phonemes, we first find out these two key pose sequences in the phoneme-pose dictionary and the timestamps of these two phonemes in the output frames. The figure shows the case of the interval length between two phonemes is larger than the minimum key pose distance, we copy these two phoneme sequences to the output frames and apply the interpolation to the middle poses between these two adjacent key poses.*

3.4. Key Pose Insertion

To generate a sequence of poses, we need to do key pose insertion for the missing poses between key poses. We go through all phonemes one by one in speech and find their corresponding poses in the phoneme-pose dictionary. When we insert a pose into a video, we do a interpolation in their pose parameter space. We determine the interpolation strategies by taking consideration of the following factors: phoneme poses width (which represents the number of frames for a key pose sequence extracted from the phoneme-pose dictionary), and minimum key pose distance (which determine if we need to do interpolation).

Our interpolation strategies are as follows: If the interval length between two phoneme key pose frames is larger than or equal to the minimum key pose distance, we will do the interpolation using the key pose$_i$ and key pose$_{i+1}$. If the interval length between two phoneme key pose frames is smaller than the minimum key pose distance, we will skip the key pose$_{i+1}$ and using the key pose$_i$ and key pose$_{i+2}$ to do the interpolation. Then, we will blend key poses between two key pose sequences with a weighted sum of phoneme poses using interpolation which is illustrated in figure 2. The new frames in the output sequence are interpolated between two key pose frames, weighted by their distance to those two frames. Weight is inverse proportional to the distance with a key frame which means the larger the distance, the smaller the weight.

3.5. Smoothing

Smoothing is implemented after the interpolation step. The phoneme pose is directly copied to its time point within the video. The smoothing of the motion of poses is controlled by a smooth width parameter. To make human motion more stable, we smooth all face keypoints except the mouth part. Because smoothing the mouth directly will sacrifice the accuracy of the mouth shape corresponding to phonemes, we calculate the mouth center and shift for all mouth key points corresponding to the center of the mouth. The new frames are linearly interpolated, weighted by their distance to other frames in the sliding window. Finally, we copy mouth key points to the mouth center of each frame. We smooth the frames one by one in the sliding window till the end of the pose sequences.

3.6. Train Video Generation Network

We utilize the generative network vid2vid [16] to convert our pose sequences into real human speech videos. We modified the GAN network to put more weight on the face to emphasize this part.

4. Experiments

4.1. Dataset and Settings

Dataset. To validate our approach, we used the VidTIMIT dataset [17]. The VidTIMIT dataset consists of video and corresponding audio recordings of 43 people (19 female and 24 male), reading sentences chosen from the TIMIT corpus [18]. There are ten sentences for each person. The sentences' mean duration is 4.25 seconds, or about 106 video frames (25 fps). To test our algorithm, we also recorded a dataset of our own. We invited a female native English speaker to do recording via zoom meeting. We prepared prompts, including 44 words and 20 sentences. The word examples are transcribed in ARPABET (see appendix). We also tested our algorithm in other languages like Mandarin Chinese. We used a native Mandarin Chinese speaker (female) as a model and captured a video of her reading a list of 386 syllables in Pinyin. The total recorded video is approximately 8 mins. Besides, we used online Youtube videos of a Chinese news broadcaster to test our algorithm in the wild. Details of the four datasets are compared in table 1.

Implementation Details. In our experiments, the video frame rate is 25 fps. We set phoneme poses width equals to 7, minimum key pose distance to 4, and smooth sliding window size to 9. Fig 3 shows VidTIMIT output. The result videos are in the supplementary multimedia file and the demo video is at `https://youtu.be/d5MFzHxeOTs`.

4.2. Evaluation

To evaluate the generated videos' quality, we conducted a human subjective test on Amazon Mechanical Turk (AMT) with 401 participants. We showed a total of 5 videos to the participants. The participants were required to rate those videos' quality on a Likert scale from 1 (very bad) to 5 (very good). The ratings include 1) The face in the video is clear; 2) The face motion in the video looks natural and smooth; 3) The audio-visual alignment (lip-sync) quality; 4) The overall visual quality of the video. We choose to compare our results with SoTA approaches using user study, including LearningGesture [11], neural-voice-puppetry [2], and Speech2Video [3]. Since these three methods are audio-based and use the real human voice in their demo videos. We also used a real human voice for the comparison. Table 2 shows the scores from the user study for all methods. Our method has the best overall quality score compared to the other 3 SOTA methods. Besides, our text-based method is more flexible than the aforementioned audio-based method and not subject to vulnerability due to speaker variation.

Table 1: *Dataset detail of VidTIMIT, data from two models we hired and Youtube. Dataset details include training video duration, data source, recording resolution, and how a phoneme-pose dictionary was built from the data.*

	Training video time	Captured by/ Data From	Video Resolution	Phoneme-pose dictionary built from
VidTIMIT dataset	1 min per person, 43 people	Digital camera	512*384	10 English sentences
American Female	Total 6 min.	Zoom	640*480	44 English words
Chinese Female	Total 8 min.	Digital camera	1920*1080	386 Chinese pinyin
Chinese Male	Total 10 min.	Youtube	512*448	Chinese pinyin extracted from video

	Q1	Q2	Q3	Q4
LearningGesture	3.424	3.267	3.544	3.204
Neural-voice-puppetry	3.585	3.521	3.214	3.465
Speech2Video	3.513	3.308	3.094	3.262
Text2Video	**3.761**	**3.924**	**3.567**	**3.848**

Table 2: *User Study. Average scores of 401 participants on 4 questions. Q1: face is clear. Q2: The face motion in the video looks natural and smooth. Q3: The audio-visual alignment (lip sync) quality. Q4: Overall visual quality.*

	Q1	Q2	Q3	Q4
Text2Video(w/TTS)	3.33	3.51	3.23	3.09
Text2Video(w/Human voice)	3.28	3.50	3.21	3.18
Real video	3.52	3.87	3.86	3.46

Table 3: *Ablation study on different voice quality. Average scores of 401 participants on same questions as Table 2.*

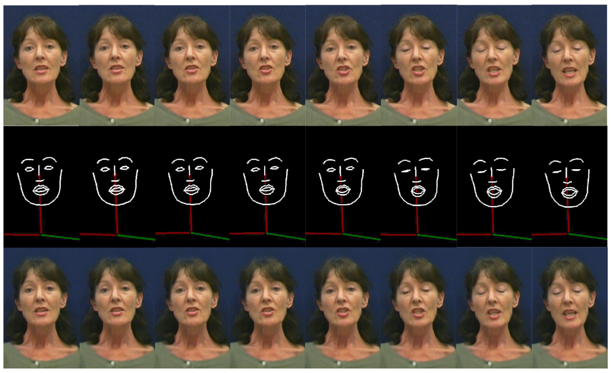

Figure 3: *The output of our method from the VidTIMIT dataset. The first line shows the ground truth video clips of "She" or "SH IY1" in phonemes, the second line shows the output pose sequences, and the third line shows the synthesized image sequences generate from pose sequences.*

4.3. Ablation Study

We also implemented the following user study to validate the effectiveness of our method. We showed three videos to the participants. We used the same text input as the real speech video to generate two synthesized videos, one with the real person speech and the other with a TTS voice. The remaining one is the real speech video. We put all videos randomly without telling the participants which one is real. As shown in table 3, our output video with human voice got 3.18, and the real video got 3.46 (out of 5) on overall visual quality. The generated video is 91.9% of the overall quality of the real video. In particular, our proposed method has similar performance on face clarity and motion smoothness compared to the real video. Our TTS one got 89.0% of the overall quality of the real video. The little difference should come from the quality of the TTS audio. Here we simply picked an average female voice in our experiment. Using a better TTS or using a learning method to train a personalized human voice could improve the overall audio quality. Based on the user study, our text-based video generation method showed an overall visual quality that has barely correlated with the voice quality.

4.4. Running Times and Hardware

Here we compare our method with SythesisObama [1], neural-voice-puppetry [2], and Speech2Video [3] on training data duration, data preprocessing time, training time, and inference time. Our method needs the least amount of data to train a model. For instance, using our fine-grained 40 words or 20 sentence list to capture all phonemes in English, the training video input is less than 1 minute. Using existing videos to extract a phoneme-pose dictionary will also make the training data a similar size. The total number of images we need to train is around 1250 images for 25 fps 60s video.

Besides, our method needs the least preprocessing and training time among all four approaches. Preprocessing time of our approach includes running Openpose and building up a phoneme-pose dictionary. The training time of our method is relatively short. It took about 4 hours to finish 15 epochs of training on a cluster of 8 NVIDIA Tesla M40 24G GPUs while other methods need at least 30 hours. For the VidTIMIT dataset which has a resolution of 512*384, a model trained on 15 epochs is good for inference. The inference time of our method is around 0.1 second per frame, which is similar to Neural-voice-puppetry but much faster than SythesisObama (1.5 s/frame) and Speech2Video (0.5 s/frame) on Nvidia 1080Ti. Details of comparison can be found in appendix table 1.

5. Conclusion and Future Work

In this paper, we proposed a novel method to synthesize talking-head speech video from any text input. Our method includes an automatic pose extraction to build a phoneme - pose dictionary from any video. Compared to SOTA audio-driven methods, our text-based video synthesis method only needs a fraction of the training data and significantly reduces inference time. We demonstrated the effectiveness of our approach for both English and Mandarin Chinese text inputs. In future, we will extend our framework and build phoneme-pose dictionaries for other languages. We will also integrate voice learning methods into our training pipeline to generate personalized voices.

6. Acknowledgements

The paper is submitted to Interspeech 2021.

7. References

[1] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 95, 2017.

[2] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," *arXiv preprint arXiv:1912.05566*, 2019.

[3] M. Liao, S. Zhang, P. Wang, H. Zhu, X. Zuo, and R. Yang, "Speech2video synthesis with 3d skeleton regularization and expressive body poses," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[4] T. Ezzat and T. Poggio, "Visual speech synthesis by morphing visemes," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 45–57, 2000.

[5] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," in *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, 2012, pp. 275–284.

[6] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, "Text-based editing of talking-head video," *arXiv preprint arXiv:1906.01524*, 2019.

[7] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7832–7841.

[8] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9299–9306.

[9] W. Wang, Y. Wang, J. Sun, Q. Liu, J. Liang, and T. Li, "Speech driven talking head generation via attentional landmarks based representation," *Proc. Interspeech 2020*, pp. 1326–1330, 2020.

[10] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.

[11] 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*, 2019, pp. 3497–3506.

[12] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[13] H. Ren, "On the acoustic structure of diphthongal syllables," Ph.D. dissertation, UCLA, 1986.

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," in *arXiv preprint arXiv:1812.08008*, 2018.

[15] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *The Journal of the Acoustical Society of America*, vol. 123, p. 3878, 2008.

[16] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[17] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *International conference on biometrics*. Springer, 2009, pp. 199–208.

[18] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.