Paper Number: 1272

Text2Video: Text-driven Talking-head Video Synthesis with Personalized Phoneme-Pose Dictionary

Sibo Zhang, Jiahong Yuan, Miao Liao, Liangjun Zhang
Baidu Research USA
sibozhang1@gmail.com

Problem with Speech2Video

- Automatic video generation from audio (**Speech2Video**) or text (**Text2Video**) has become an emerging and promising research topic
- Previous Speech2Video 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.

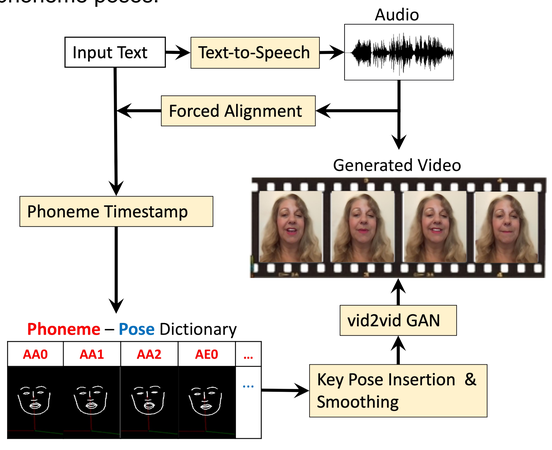
  3) Users can not manipulate motion output such as changing speaker attitude since the network is a black box on what is learned.

Main Contributions

1. A novel pipeline of generating talking-head speech videos from any text input, including numbers and punctuation, in both English and Mandarin Chinese.
2. An automatic pose extraction method to build a phoneme - pose dictionary from any video, online or purposely recorded.
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.

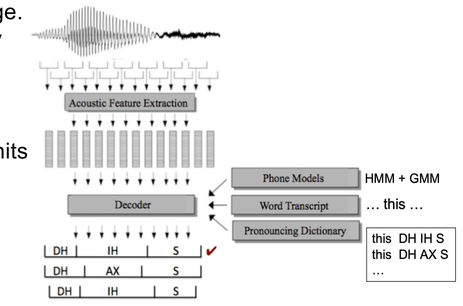
Text2Video Framework

As shown in Figure, the input to our system is text, and 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.



Phoneme - Pose Dictionary

- Phonemes are the basic units of the sound structure of a language. English has 40 phonemes. They are produced with different positions of the tongue and lips.
- For Mandarin Chinese, we use initials and finals as the basic units in the phoneme-pose dictionary.
- 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 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, an interpolation is performed in their pose parameter space.
- We determine the interpolation strategies by taking consideration of the following factors:

  1) Phoneme poses width (which represents the number of frames for a key pose sequence extracted from the phoneme-pose dictionary), and minimum key poses distance (which determine if we need to do interpolation).

  2) Minimum key poses distance between two phonemes equals to the sum of (half of the first phoneme pose width + half of the second phoneme pose width). The equation is defined as:

$$distance = \frac{1}{2} \times width_i + \frac{1}{2} \times width_{i+1}, \qquad (1)$$

Smoothing

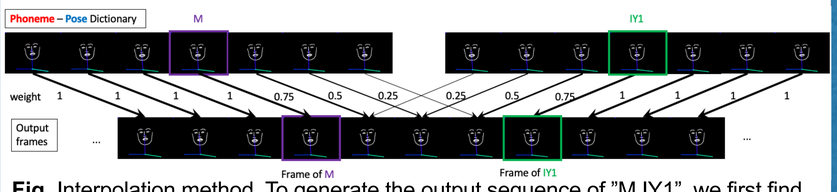


Fig. Interpolation method. To generate the output sequence of ”M IY1”, we first find the key-pose sequences of "M" and "IY" in the phoneme-pose dictionary, as well as the timestamps of the two phonemes in the output. Then we copy the two key-pose sequences to the output frames and apply interpolation to the middle frames between the two adjacent key poses.

Datasets

1. VidTIMIT dataset. 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.
2. Female English speaker.
3. Female Mandarin Chinese speaker.
4. Male Chinese news broadcaster from Youtube video.

Fig. 1. left up 2. left right 3.left down 4. right down

User Studies

As shown in Table, our output video with human voice got 3.68, and the real video got 4.06 (out of 5) on overall visual quality. The generated video is 90.6\% of the overall quality of the real 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 1. 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)	3.73	3.91	3.63	3.55
Text2Video(w/Human voice)	3.78	4.01	3.71	3.68
Real video	4.02	4.47	4.46	4.06

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

Results

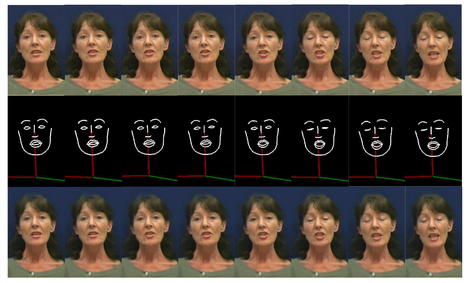


Figure. The output of our method from the VidTIMIT dataset. The first line shows the ground truth video clips of ``She" or ``SH IY1" in phonemes, the second line shows the output pose sequences, and the third line shows the synthesized image sequences generate from pose sequences.

More Information

Full video is available on YouTube

Text2Video: Text-driven Talking-head Video Synthesis with Personalized Phoneme - Pose Dictionary

Sibo Zhang, Jiahong Yuan, Miao Liao, Liangjun Zhang

Check out code on GitHub
https://github.com/sibozhang/Text2Video

ICASSP 2022 Paper:
https://arxiv.org/abs/2104.14631

Demo video:
https://youtu.be/TQJCyQ4ISEg

icassp 2022 Singapore, China, Online

The 47th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 7-13 May, 2022 · Virtual, Singapore, China

IEEE Signal Processing Society

IEEE