# Video Generation

Sibo Zhang

# Generative AI + Video

- **Speech2Video** (ACCV 2020)
- **Text2Video** (ICASSP 2022)

ECCV 2020

# *Speech2Video* Synthesis with 3D Skeleton Regularization and Expressive Body Poses

Miao Liao*, **Sibo Zhang***, Peng Wang, Hao Zhu, Xinxin Zuo, and Ruigang Yang

# Introduction

- Speech2Video is a task of synthesizing a video of human full body movements from a speech audio input.

# Main contributions

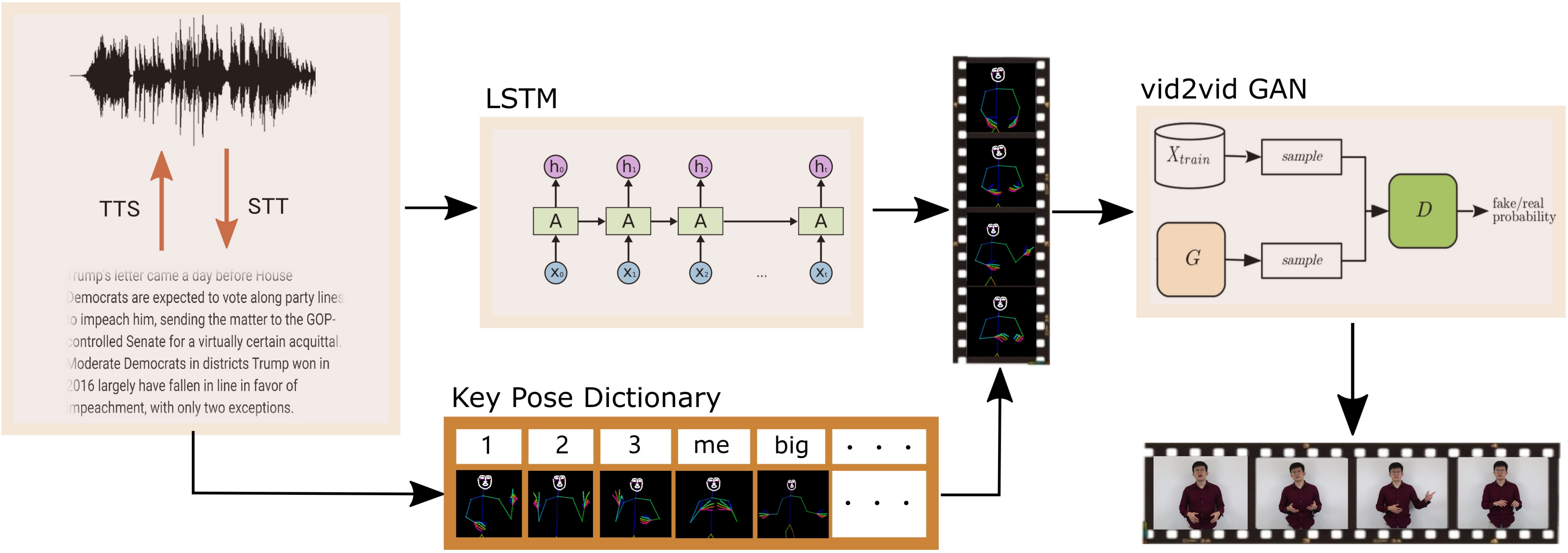
1. Novel 2-stage pipeline of generating an audio-driven virtual speaker with full-body motions.

2. A dictionary of personal key poses. An approach to insert key poses into the existing sequence.

3. 3D skeleton constraints.

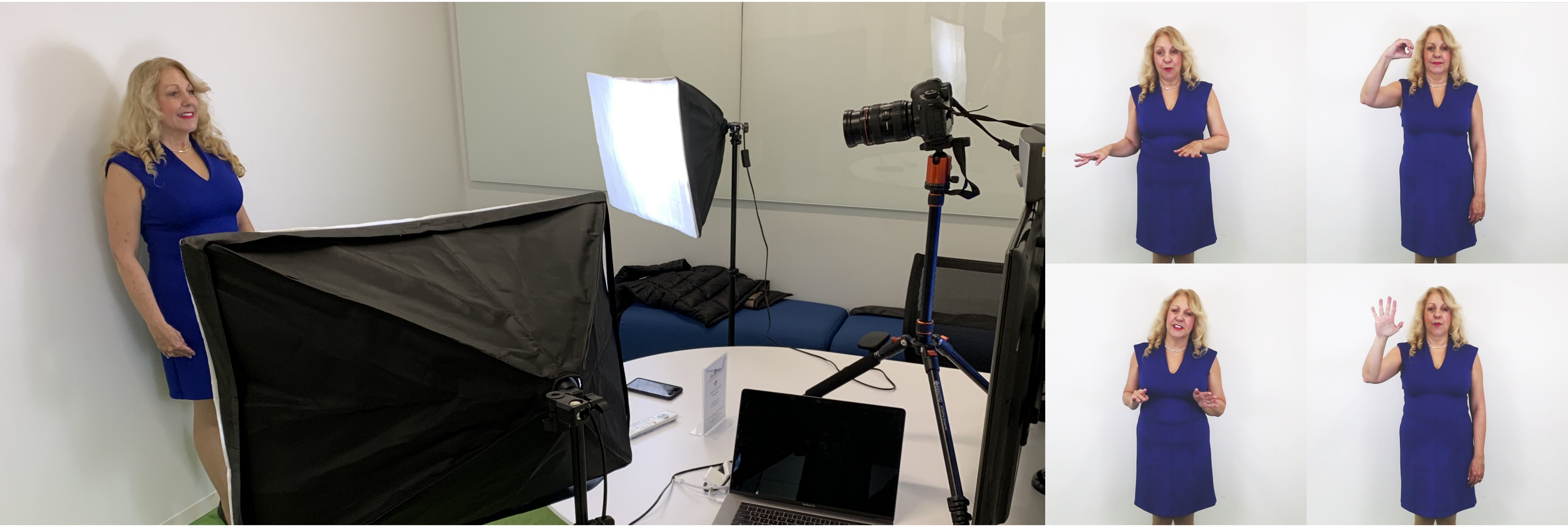
4. Modified GAN to emphasize on face and hands.

# System Pipeline



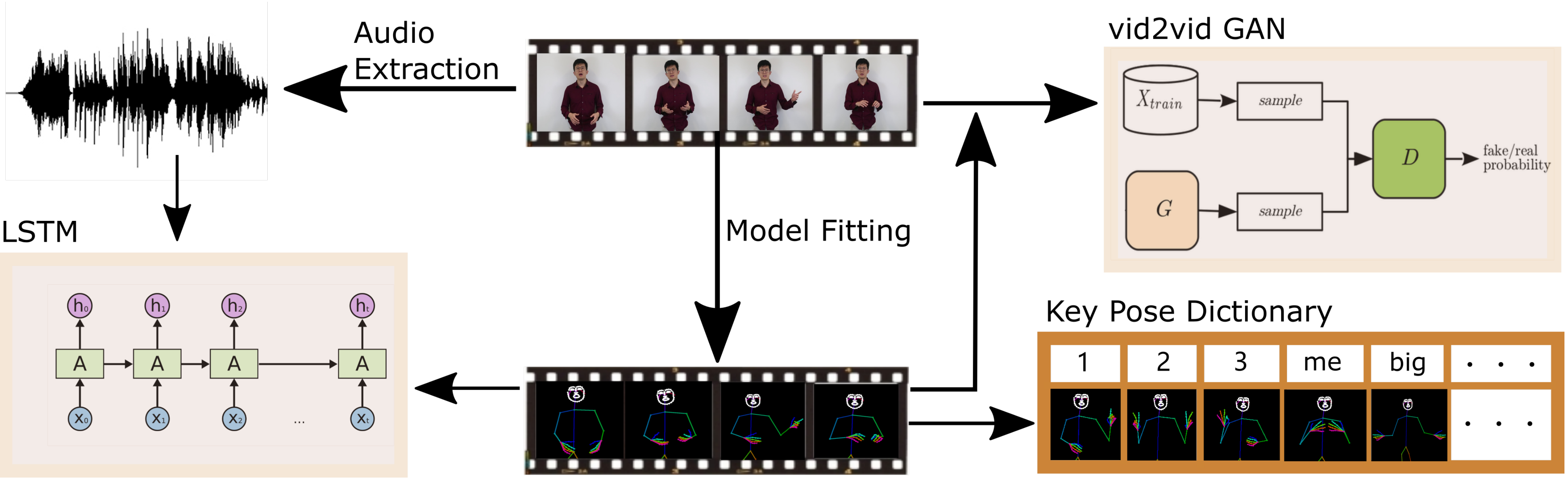
rump's letter came a day before House
Democrats are expected to vote along party lines
o impeach him, sending the matter to the GOP-
controlled Senate for a virtually certain acquittal.
Moderate Democrats in districts Trump won in
2016 largely have fallen in line in favor of
mpeachment, with only two exceptions.

# Speech2Video Dataset



Left figure shows our data capture room; Right figure are 4 frames from captured video.

# Training Process



Audio Extraction

vid2vid GAN

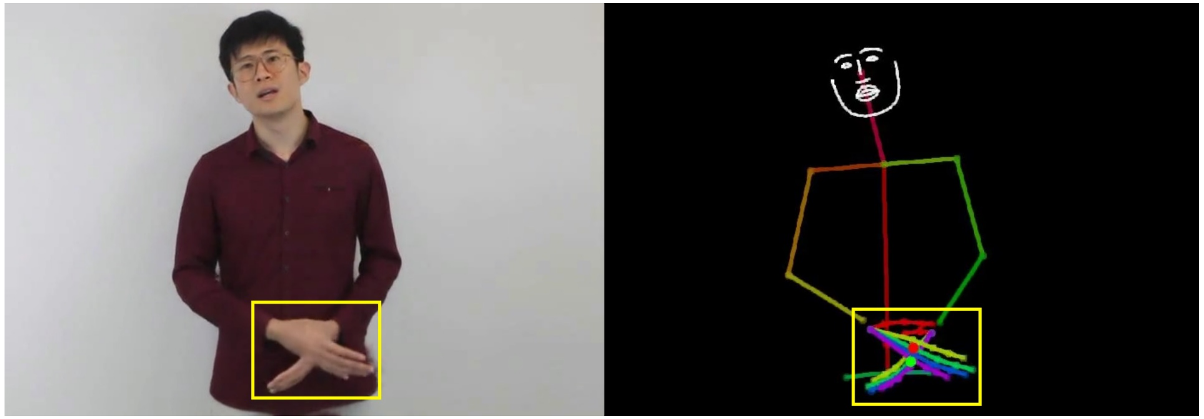
Model Fitting

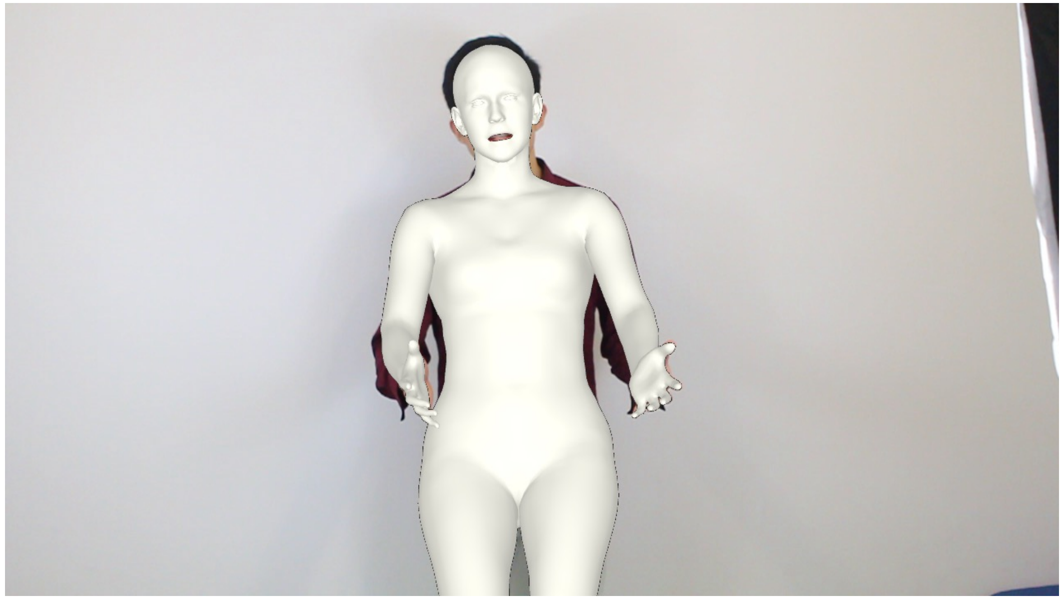
LSTM

Key Pose Dictionary

| | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | me | big | . . . |

# Body Model Fitting

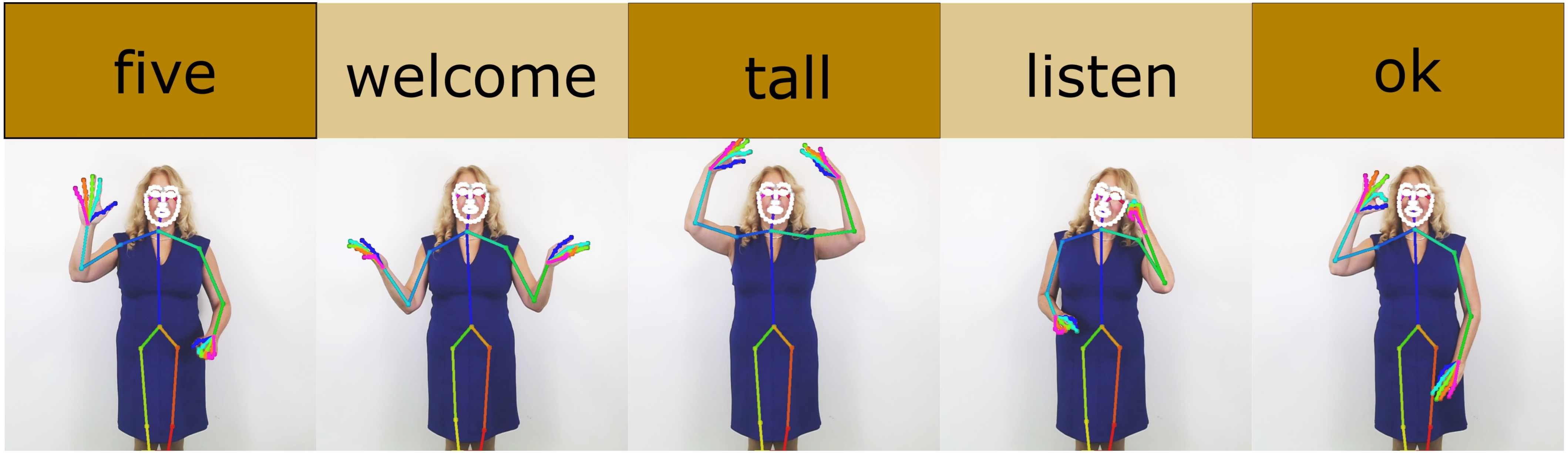




- Failure case with 2d model: elongated fingers.

- SMPL-X 3d model

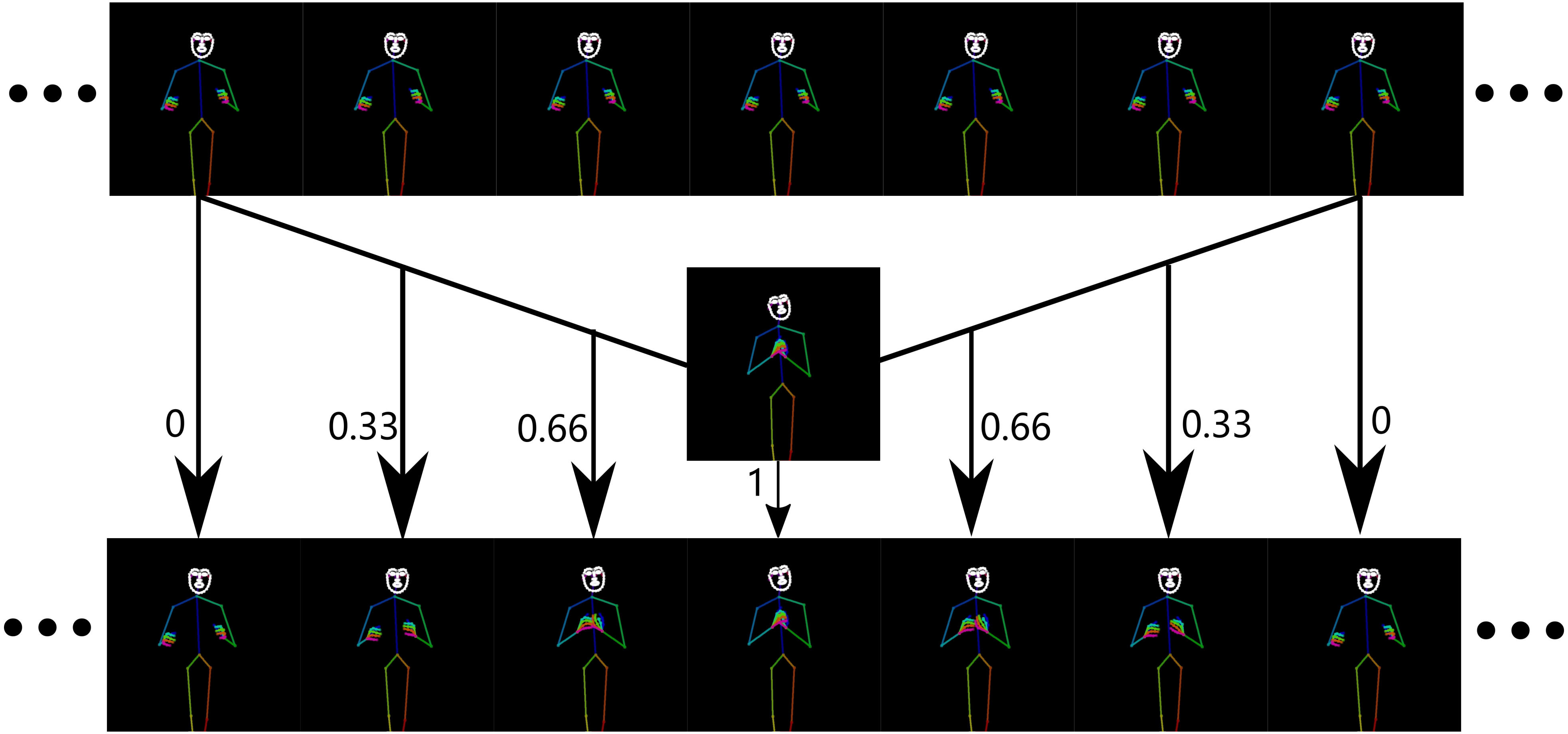
# Dictionary Building



five | welcome | tall | listen | ok

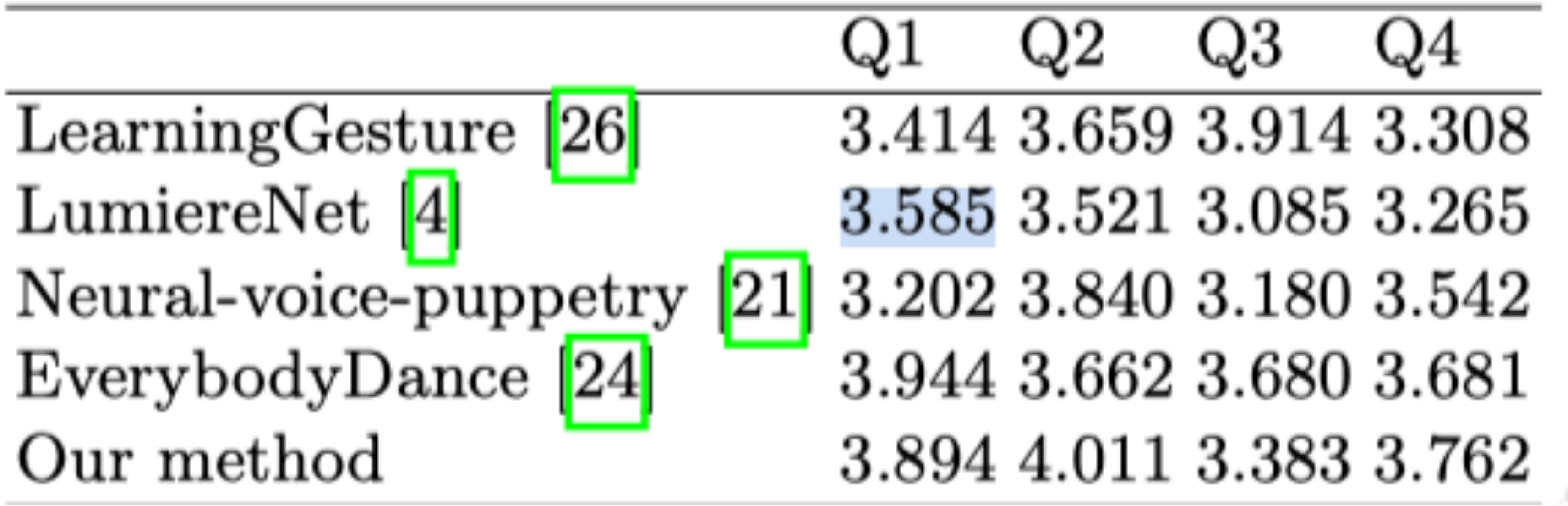
- Example poses in our dictionary. We also have motion poses (a sequence of frames).

# Key Pose Insertion



Inserting a key pose smoothly into an existing video sequence

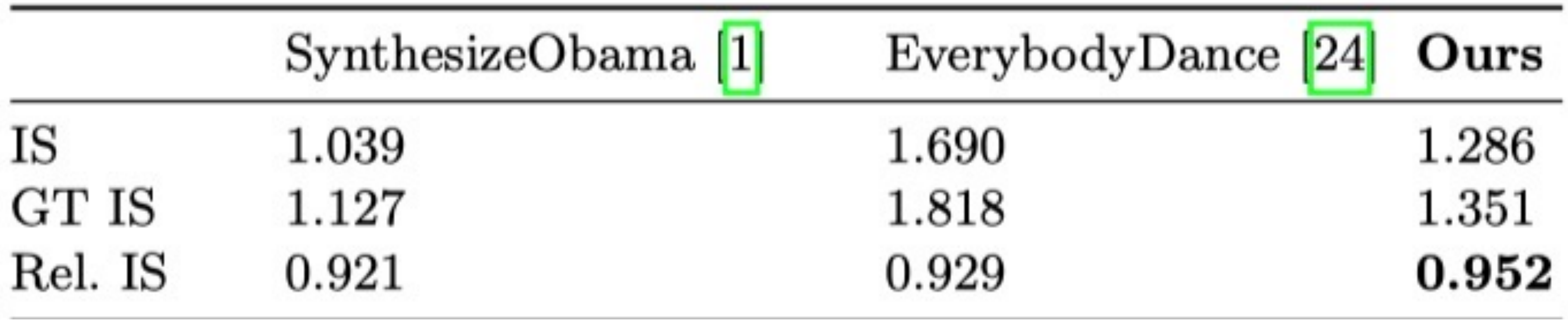
# User Study



| | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| LearningGesture [26] | 3.414 | 3.659 | 3.914 | 3.308 |
| LumiereNet [4] | 3.585 | 3.521 | 3.085 | 3.265 |
| Neural-voice-puppetry [21] | 3.202 | 3.840 | 3.180 | 3.542 |
| EverybodyDance [24] | 3.944 | 3.662 | 3.680 | 3.681 |
| Our method | 3.894 | 4.011 | 3.383 | 3.762 |

Average scores of 248 participants on 4 questions. Q1: Completeness of body. Q2: The face is clear. Q3: The body movement is correlated with audio. Q4: Overall quality.

# Inception Score Comparison



| | SynthesizeObama [1] | EverybodyDance [24] | **Ours** |
|---|---|---|---|
| IS | 1.039 | 1.690 | 1.286 |
| GT IS | 1.127 | 1.818 | 1.351 |
| Rel. IS | 0.921 | 0.929 | **0.952** |

Inception scores measure Quality and Diversity for generated videos (IS) and ground truth videos (GT IS) of different methods. The relative incpetion score (Rel. IS) is the ratio of the first to the second. (higher is better)

- Result Video: https://youtu.be/MUlRtgbGeUs
- Paper: https://arxiv.org/abs/2007.09198

Speech2Video Synthesis with 3D Skeleton Regularization and Expressive Body Poses

Sibo Zhang, Miao Liao, Feng Wang, Hao Zhu, Xinxin Zuo, Ruigang Yang

Baidu Research

ACCV 2020

Given input audio,
our method produce following video

icassp 2022 Singapore

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

***Sibo Zhang**, Jiahong Yuan, Miao Liao, Liangjun Zhang*

IEEE Signal Processing Society

IEEE

# Background and Motivation

- 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 since the network is a black box on what is learned.

- **Text2Video** is 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.

# 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. The inference time is as fast as 10 frames per second.

2) 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.

# Text2Video Framework

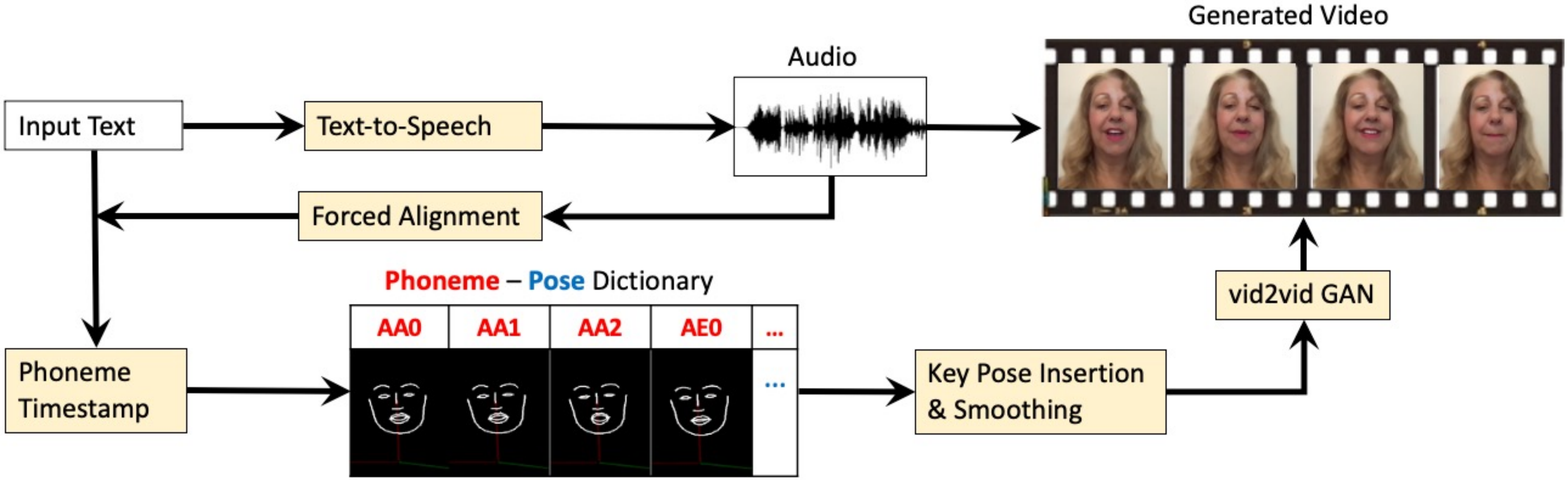


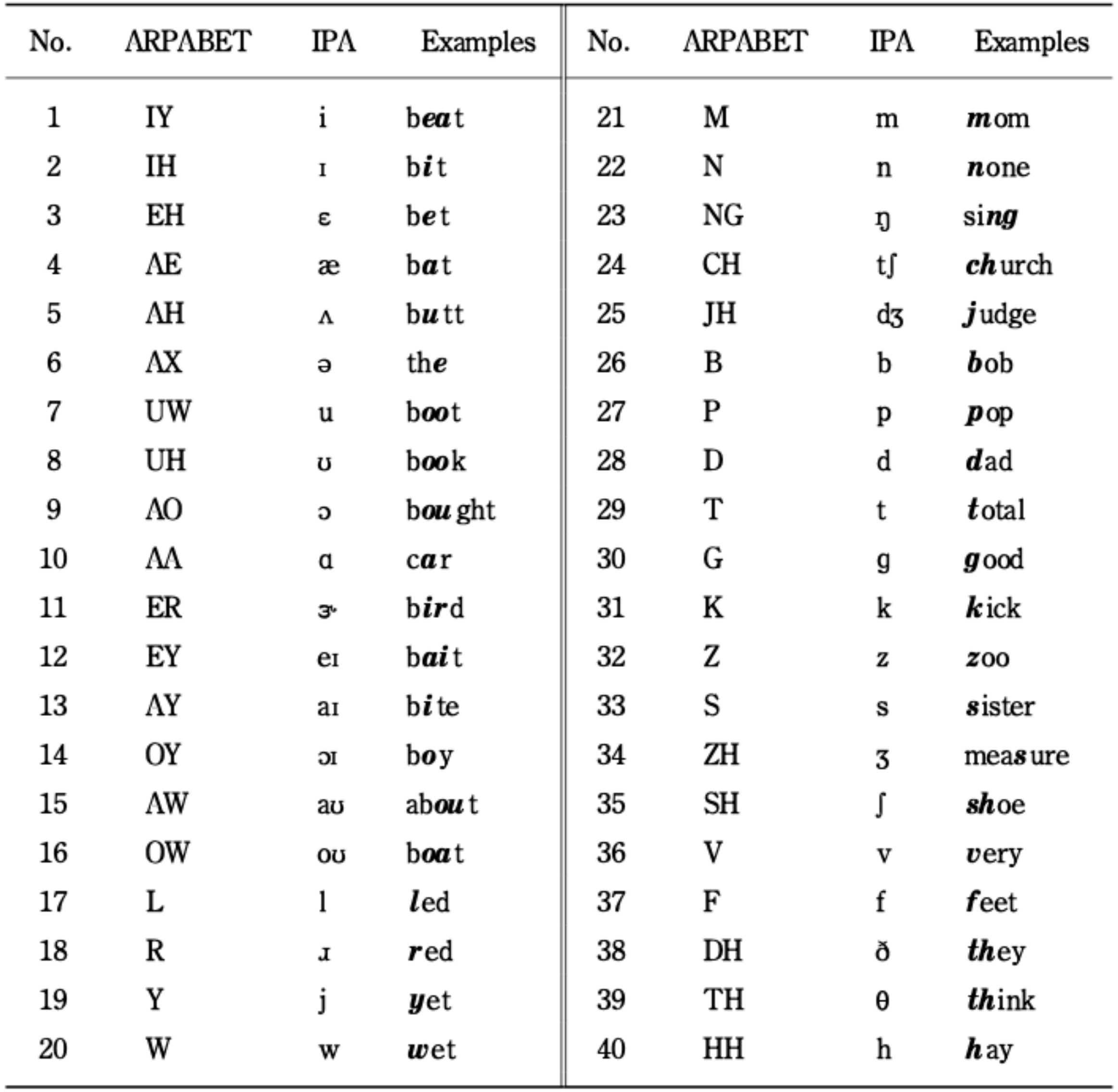
**Fig. 1.** Framework of Text2Video including two parts: building Phoneme – Pose dictionary and generating video from text. For generating video from text, we generating audio from text firstly, applying forced alignment to get phoneme timestamps, searching in a phoneme-pose dictionary, applying the key pose interpolation/ smoothing module to get a sequence of poses, and generating video using modified GAN.

# Building Phoneme - Pose Dictionary

1. What is phonemes?
2. Key Pose Extraction
3. Phoneme Extraction
4. Mapping Phoneme to key poses
    1) Key Pose Insertion
    2) Smoothing

# What is phoneme?

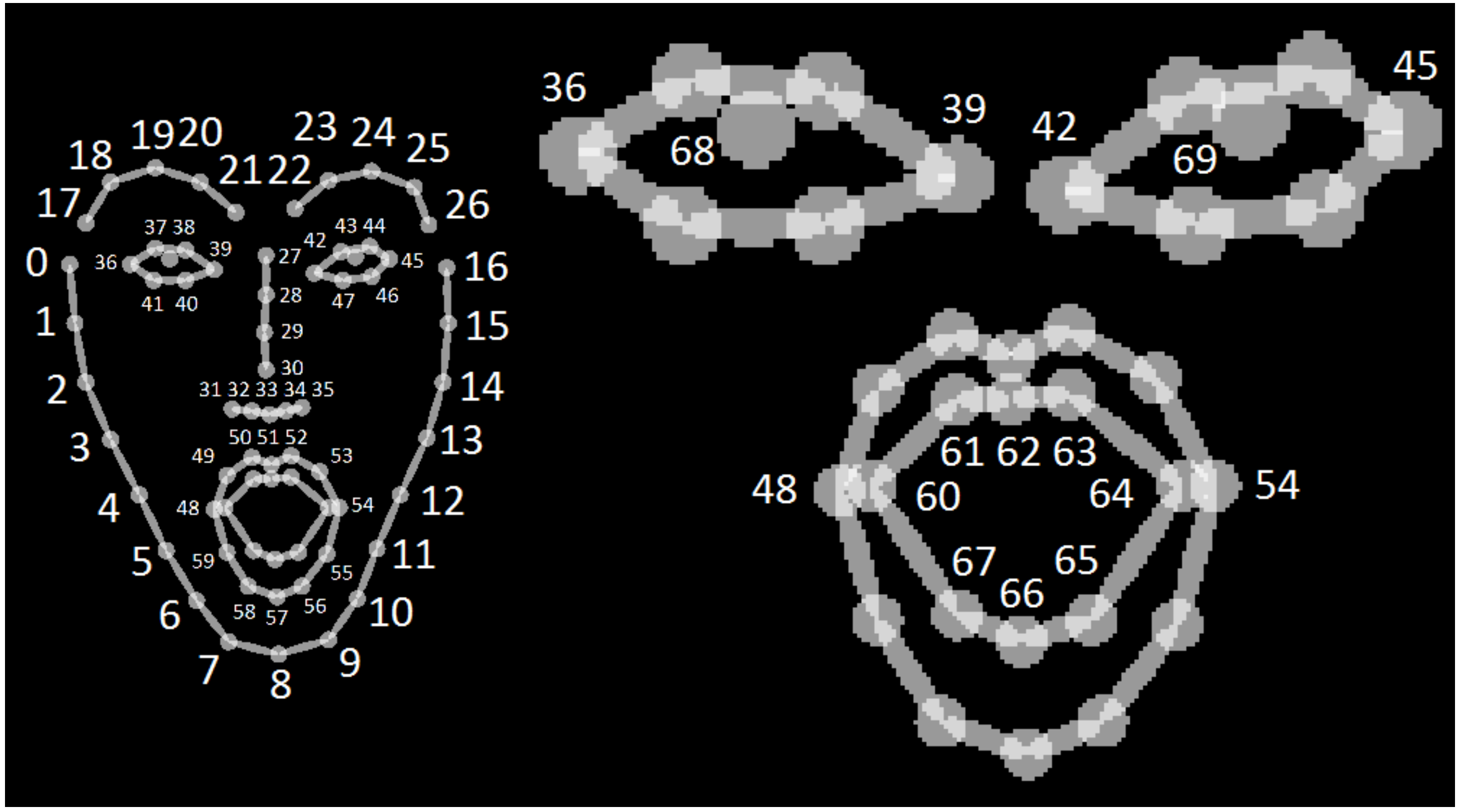
- Phonemes are the basic units of the sound structure of a language. English has 40 phonemes. They are produced with different positions of the tongue and lips.
- For Mandarin Chinese, we use initials and finals as the basic units in the phoneme-pose dictionary.
- We build a phoneme-pose dictionary for English and Mandarin Chinese, respectively, mapping from phonemes to lip postures extracted from a speech production video.



| No. | ΛRPΛBET | IPΛ | Examples | No. | ΛRPΛBET | IPΛ | Examples |
|---|---|---|---|---|---|---|---|
| 1 | IY | i | b***ea***t | 21 | M | m | ***m***om |
| 2 | IH | ɪ | b***i***t | 22 | N | n | ***n***one |
| 3 | EH | ɛ | b***e***t | 23 | NG | ŋ | si***ng*** |
| 4 | ΛE | æ | b***a***t | 24 | CH | tʃ | ***ch***urch |
| 5 | ΛH | ʌ | b***u***tt | 25 | JH | dʒ | ***j***udge |
| 6 | ΛX | ə | th***e*** | 26 | B | b | ***b***ob |
| 7 | UW | u | b***oo***t | 27 | P | p | ***p***op |
| 8 | UH | ʊ | b***oo***k | 28 | D | d | ***d***ad |
| 9 | ΛO | ɔ | b***ou***ght | 29 | T | t | ***t***otal |
| 10 | ΛΛ | ɑ | c***a***r | 30 | G | g | ***g***ood |
| 11 | ER | ɝ | b***ir***d | 31 | K | k | ***k***ick |
| 12 | EY | eɪ | b***ai***t | 32 | Z | z | ***z***oo |
| 13 | ΛY | aɪ | b***i***te | 33 | S | s | ***s***ister |
| 14 | OY | ɔɪ | b***oy*** | 34 | ZH | ʒ | mea***s***ure |
| 15 | ΛW | aʊ | ab***ou***t | 35 | SH | ʃ | ***sh***oe |
| 16 | OW | oʊ | b***oa***t | 36 | V | v | ***v***ery |
| 17 | L | l | ***l***ed | 37 | F | f | ***f***eet |
| 18 | R | ɹ | ***r***ed | 38 | DH | ð | ***th***ey |
| 19 | Y | j | ***y***et | 39 | TH | θ | ***th***ink |
| 20 | W | w | ***w***et | 40 | HH | h | ***h***ay |

# Key Pose Extraction

- We use OpenPose to extract key poses from training videos by averaging all the phoneme-poses present in the training video. Then we build up the phoneme-pose dictionary from our phoneme extraction pipeline.



# Phoneme Extraction

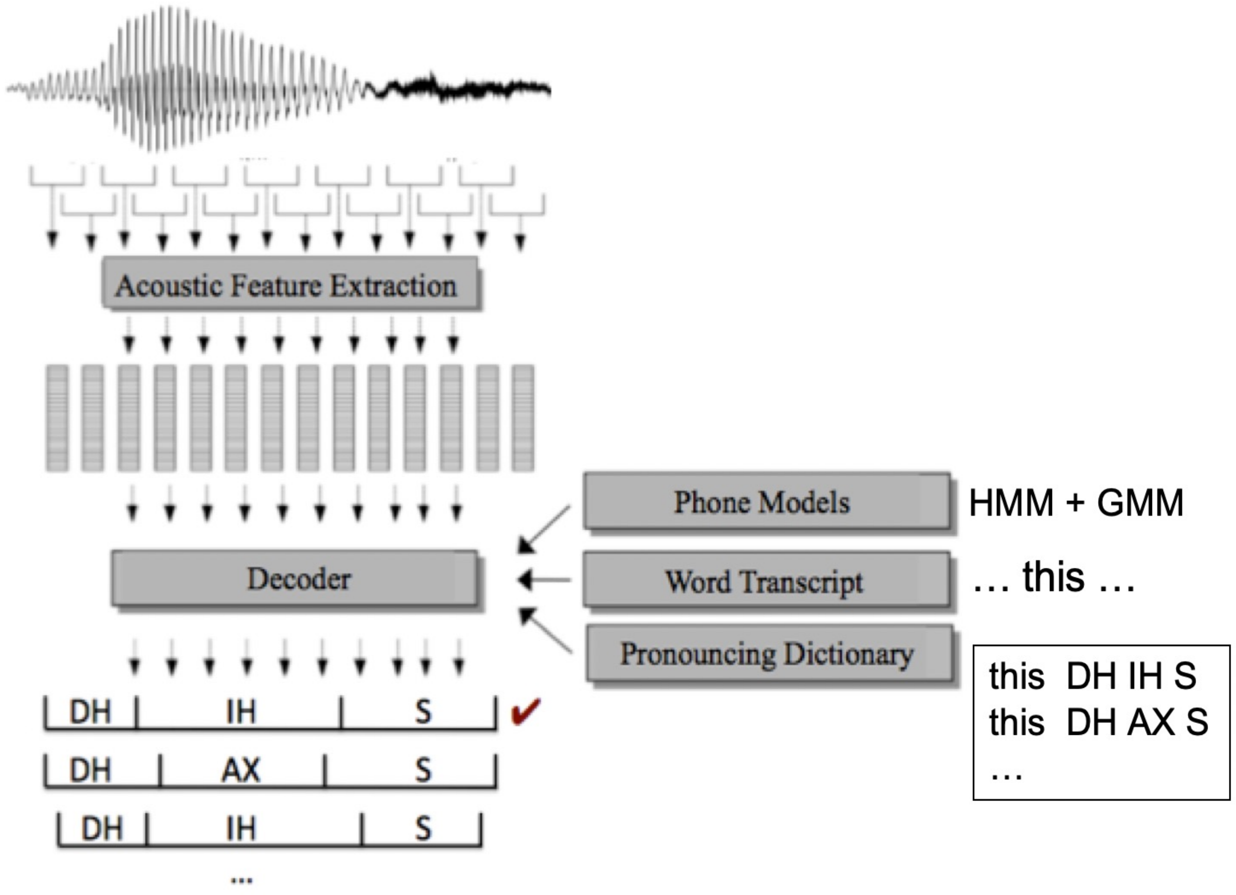


**Figure**. Forced aligner for phoneme extraction.

- We employed the P2FA aligner 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.
- Phone boundaries are determined by comparing the observed speech signal and pre-trained, Hidden Markov Model (HMM) based acoustic models.

# Key Pose Insertion

- 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).
- 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

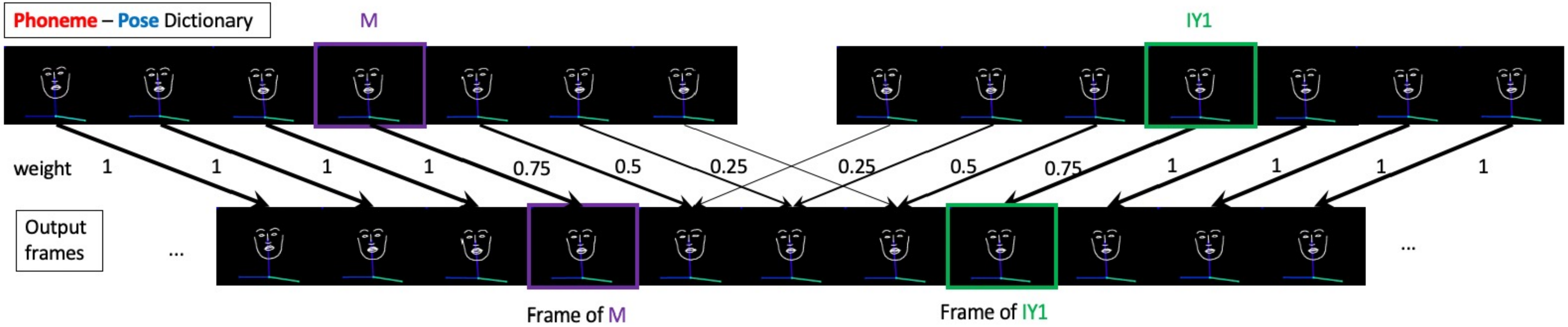
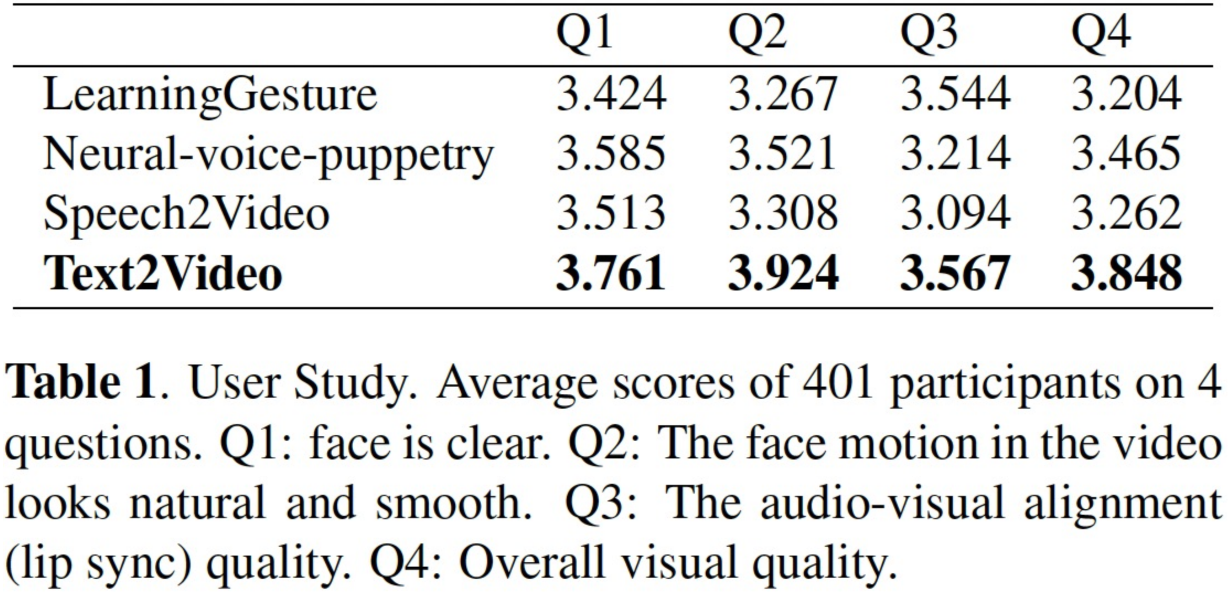


**Fig. 2**. Interpolation method. To generate the output sequence of ”M IY1”, we first find the key-pose sequences of ”M” and ”IY” in the phoneme-pose dictionary, as well as the timestamps of the two phonemes in the output. Then we copy the two key-pose sequences to the output frames and apply interpolation to the middle frames between the two adjacent key poses.

# Datasets

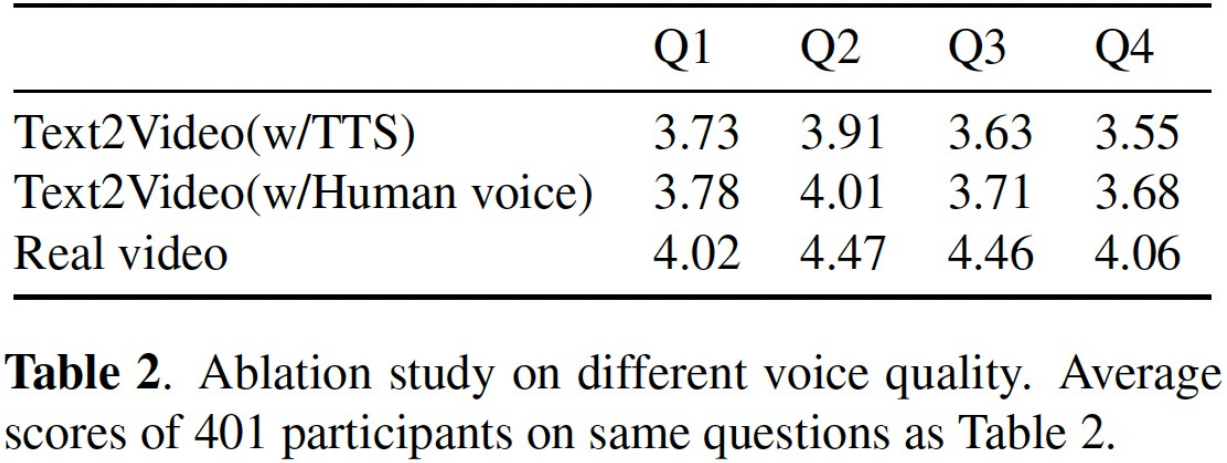
- Left to right:
- 1) VidTIMIT dataset. The VidTIMIT dataset consists of video and corresponding audio recordings of 43 people (19 female and 24 male), reading sentences chosen from the TIMIT corpus.
- 2) Female English speaker.
- 3) Female Mandarin Chinese speaker.
- 4) Male Chinese news broadcaster from Youtube video.

# Evaluation



| | 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.

# Training Data, Preprocessing and Training Time

| | Synthesizing Obama | Neural-voice-puppetry | Speech2Video | **Our method** |
|---|---|---|---|---|
| Training data | 17 hr | 3 min + | ~ 20 min | **1 min** |
| Preprocessing time | 2 weeks | a few hours | A few hours | **10 mins** |
| Training time | 2 hours | Audio2ExpressionNet: ~28 hour;<br>Rendering networks: ~30 hours training time | 3 days | **4 hours** |
| Inference time<br>(per frame) | 1.5s | 0.01 - 0.1 s | 0.5 s | **0.01 - 0.1 s** |

# VidTIMIT Dataset Result (Human Voice)

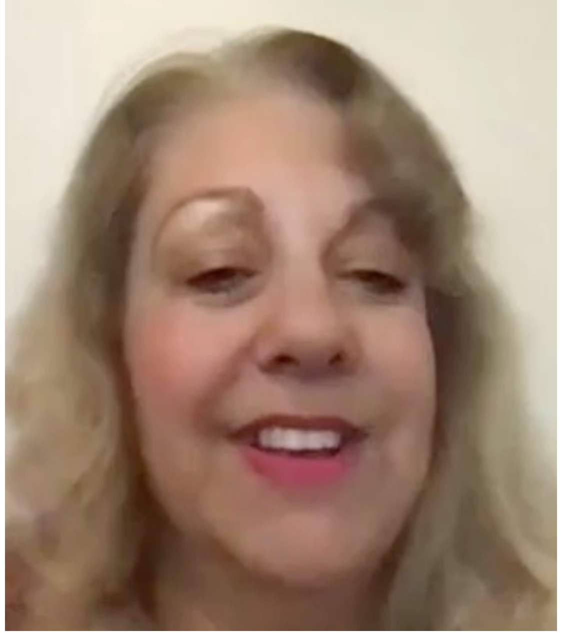
**She had your dark suit in greasy wash water all year.**

Input text

Output video

# English Female Result (TTS)

**It suffers from a lack of unity of purpose and respect for heroic leadership.**



Input text

Output video

# Chinese Female Result (TTS)

**Checking the weather in Hefei, China for you. Today is February 24, 2020. Hefei is cloudy today, the lowest temperature is 9 degrees Celsius, the highest temperature is 15 degrees Celsius, and there is light breeze.**

正在为您查询合肥的天气情况。今天是2020年2月24日，合肥市今天多云，最低温度9摄氏度，最高温度15摄氏度，微风。

Input text

Output video

# Youtube Video Result (Chinese TTS)

**Kobe knows basketball, knows the game, knows what he needs to do to win the championship (won the state championship without a good point guard and a good shooter); He is a true leader, often facing multiplayer double teams and still dominates the game , good three-point shooting; a very determined competitor.**

科比了解篮球，了解比赛，知道自己需要做什么才能赢得冠军（在没有优秀控卫和优秀射手的情况下赢得了州冠军）；是个真正的领袖，经常会面对多人包夹依然统治了比赛，三分命中率不错；是个非常有决心的竞争者。

Input text

Output video

# Conclusion

In this paper, we proposed a novel method to synthesize talking-head 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 needs significantly less training data and has 10 times faster preprocessing and training time. We demonstrated the effectiveness of our approach for both English and Mandarin Chinese text inputs.

# Thank you!