MakeItTalk: Speaker-Aware Talking-Head Animation

YANG ZHOU, University of Massachusetts Amherst
XINTONG HAN, Huya Inc
ELI SHECHTMAN, Adobe Research
JOSE ECHEVARRIA, Adobe Research
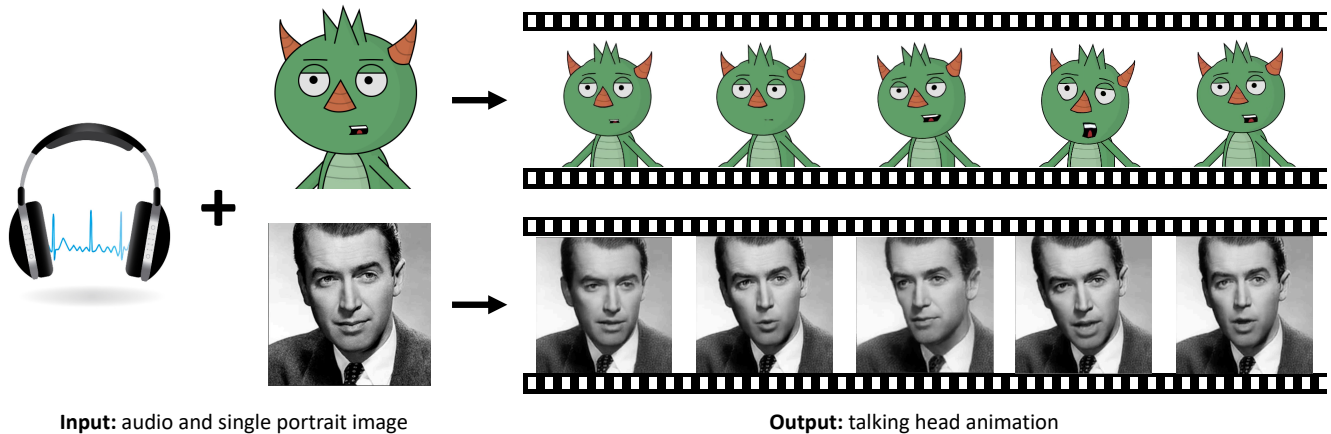EVANGELOS KALOGERAKIS, University of Massachusetts Amherst
DINGZEYU LI, Adobe Research

Input: audio and single portrait image   **Output:** talking head animation

Fig. 1. Given an audio speech signal and a single portrait image as input (left), our model generates speaker-aware talking-head animations (right). *Both the speech signal and the input face image are not observed during the model training process.* Our method creates both non-photorealistic cartoon animations (top) and natural human face videos (bottom). Please also see our supplementary video. Cartoon character *Wilk* ©Adobe Research. Natural face *James Stewart* by *studio publicity still* (public domain).

We present a method that generates expressive talking-head videos from a single facial image with audio as the only input. In contrast to previous attempts to learn direct mappings from audio to raw pixels for creating talking faces, our method first disentangles the content and speaker information in the input audio signal. The audio content robustly controls the motion of lips and nearby facial regions, while the speaker information determines the specifics of facial expressions and the rest of the talking-head dynamics. Another key component of our method is the prediction of facial landmarks reflecting the speaker-aware dynamics. Based on this intermediate representation, our method works with many portrait images in a single unified framework, including artistic paintings, sketches, 2D cartoon characters, Japanese mangas, and stylized caricatures. In addition, our method generalizes well for faces and characters that were not observed during training. We present extensive quantitative and qualitative evaluation of our method, in addition to user studies, demonstrating generated talking-heads of significantly higher quality compared to prior state-of-the-art methods.

Authors' addresses: Yang Zhou, yangzhou@cs.umass.edu, University of Massachusetts Amherst; Xintong Han, hixintonghan@gmail.com, Huya Inc; Eli Shechtman, elishe@adobe.com, Adobe Research; Jose Echevarria, echevarr@adobe.com, Adobe Research; Evangelos Kalogerakis, kalo@cs.umass.edu, University of Massachusetts Amherst; Dingzeyu Li, dinli@adobe.com, Adobe Research.

CCS Concepts: • **Computing methodologies** → **Animation**; *Machine learning approaches.*

Additional Key Words and Phrases: Facial Animation, Neural Networks

1 INTRODUCTION

Animating expressive talking-heads is essential for filmmaking, virtual avatars, video streaming, computer games, and mixed realities. Despite recent advances, generating realistic facial animation with little or no manual labor remains an open challenge in computer graphics. Several key factors contribute to this challenge. Traditionally, the *synchronization* between speech and facial movement is hard to achieve manually. Facial dynamics lie on a high-dimensional manifold, making it nontrivial to find a mapping from audio/speech [Edwards et al. 2016]. Secondly, different talking *styles* in multiple talking-heads can convey different personalities and lead to better viewing experiences [Walker et al. 1997]. Last but not least, handling lip syncing and facial animation are not sufficient for the perception of *realism* of talking-heads. The entire facial expression considering the correlation between all facial elements and head pose

Table 1. A comparison of related works across to various criteria shown on the left. "Handle unseen faces" means handling faces (or rigged models and 3d meshes) unobserved during training. "Single target image" means requiring only a single target image instead of a video for talking head generation.

	Suwajanakorn et al. [2017]	Taylor et al. [2017]	Karras et al. [2017]	Zhou et al. [2018]	Zhou et al. [2019]	Vougioukas et al. [2019]	Chen et al. [2019]	Thies et al. [2020]	Ours
animation format	image	rigged model	3d mesh	rigged model	image	image	image	image	image
beyond lip animation	✓	×	✓	×	✓	✓	✓	✓	✓
head pose	✓	×	×	×	×	×	×	×	✓
speaker-awareness	×	×	×	×	×	×	×	✓	✓
handle unseen faces	×	✓	×	×	✓	×	✓	✓	✓
single target image	×	–	–	–	✓	✓	✓	×	✓

also play an important role [Faigin 2012; Greenwood et al. 2018]. These correlations, however, are less constrained by the audio and thus hard to be estimated.

In this paper, we propose a new method based on a deep neural architecture, called *MakeItTalk*, to address the above challenges. Our method generates talking-heads from a single facial image and audio as the only input. At test time, MakeItTalk is able to produce plausible talking-head animations with both facial expressions and head motions for new faces and voices not observed during training.

Mapping audio to facial animation is challenging, since it is not a one-to-one mapping. Different speakers can have large variations in head pose and expressions given the same audio content. The key insight of our approach is to disentangle the speech content and speaker identity information in the input audio signal. The content captures the phonetic and prosodic information in the input audio and is used for robust *synchronization* of lips and nearby facial regions. The speaker information captures the rest of the facial expressions and *head motion dynamics*, which tend to be characteristic for the speaker and are important for generating expressive talking-head animation. We demonstrate that this disentanglement leads to significantly more plausible and believable head animations. To deal with the additional challenge of producing coherent head motion, we propose a combination of LSTM and self-attention mechanisms to capture both short and long-range temporal dependencies in head pose.

Another key component of our method is the prediction of facial landmarks as an intermediate representation incorporating speaker-aware dynamics. This is in contrast to previous approaches that attempt to directly generate raw pixels or 3D morphable face models from audio. Leveraging facial landmarks as the intermediate representation between audio to visual animation has several advantages. First, based on our disentangled representations, our model learns to generate landmarks that capture subtle, speaker-dependent dynamics, sidestepping the learning of low-level pixel appearance that tends to miss those. The degrees of freedom (DoFs) for landmarks is in the order of tens (68 in our implementation), as opposed to millions of pixels in raw video generation methods. As a result, our learned model is also compact, making it possible to train it from moderately sized datasets. Last but not the least, the landmarks can be easily used to drive a wide range of different types of animation content, including human face images and non-photorealistic cartoon images, such as sketches, 2D cartoon characters, Japanese mangas and stylized caricatures. In contrast, existing video synthesis methods and morphable models are limited to human faces and cannot readily generalize to non-photorealistic or non-human faces and expressions.

Our experiments demonstrate that our method achieves significantly more accurate and plausible talking heads compared to prior work qualitatively and quantitatively, especially in the challenging setting of animating static face images unseen during training. In addition, our ablation study demonstrates the advantages of disentangling speech content and speaker identity for speaker-aware facial animation.

In summary, given an audio signal and a single portrait image as input (both unseen during training), our method generates expressive talking-head animations. We highlight the following contributions:

- We introduce a new deep-learning based architecture to predict facial landmarks, capturing both facial expressions and overall head poses, from only speech signals.
- We generate speaker-aware talking-head animations based on disentangled speech content and speaker information, inspired by advances from voice conversion.
- We present two landmark-driven image synthesis methods for non-photorealistic cartoon images and human face images. These methods can handle new faces and cartoon characters not observed during training.
- We propose a set of quantitative metrics and conduct user studies for evaluation of talking-head animation methods.

2 RELATED WORK

In computer graphics, we have a long history of cross-modal synthesis. More than two decades ago, Brand [1999] pioneered *Voice Puppetry* to generating full facial animation from an audio track. Music-driven dance generation from Shiratori et al. [2006] matched the rhythm and intensity from the input melody to full-body dance motions. Gan et al. [2020a,b] separate sound sources and generate synchronized musics from videos with people playing instruments. Recently, Ginosar et al. [2019] predicted conversational gestures and skeleton movements from speech signals. We focus on audio-driven facial animation, which can supplement other body and gesture prediction methods. In the following paragraphs, we overview prior work on facial landmark synthesis, facial animation, and video generation. Table 1 summarizes differences of methods that are most related to ours based on a set of key criteria.

Audio-driven facial landmark synthesis. Eskimez et al. [2018, 2019] generated synchronized facial landmarks with robust noise resilience
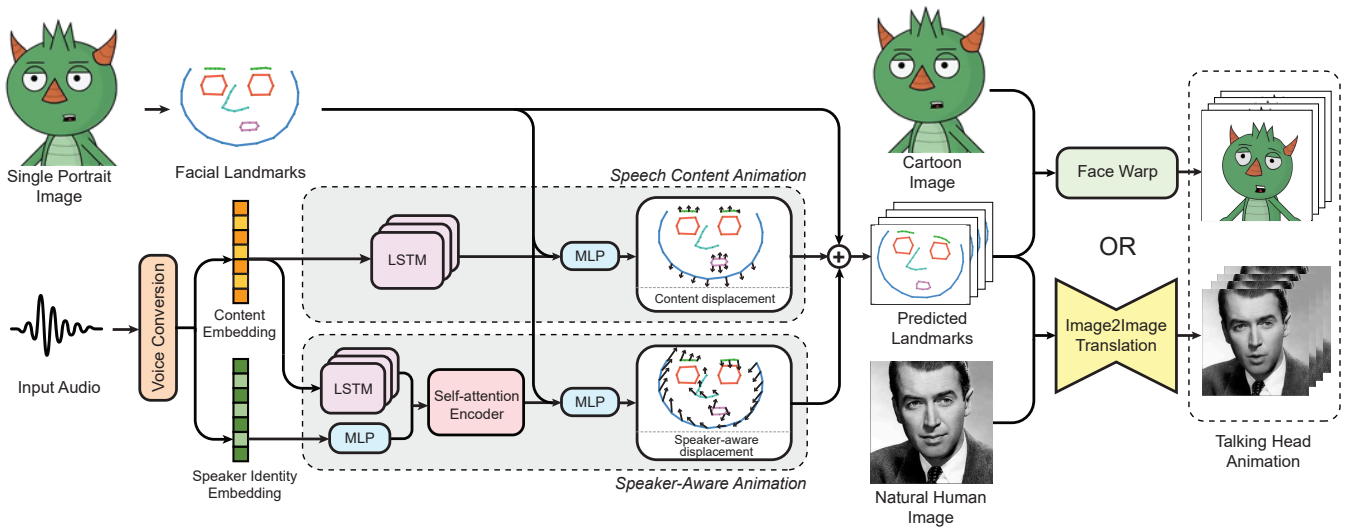
Fig. 2. Pipeline of our method ("MakeItTalk"). Given an input audio signal along with a single portrait image (cartoon or real photo), our method animates the portrait in a speaker-aware fashion driven by disentangled content and speaker embeddings. The animation is driven by intermediate predictions of 3D landmark displacements. The "speech content animation" module maps the disentangled audio content to landmark displacements synchronizing the lip, jaw, and nearby face regions with the input speech. The same set of landmarks is further modulated by the "speaker-aware animation" branch that takes into account the speaker embedding to capture the rest of the facial expressions and head motion dynamics.

using deep neural networks. Later, Chen et al. [2019] trained decoupled blocks to obtain landmarks first and then generate rasterized videos. Attention masks are used to focus on the most changing parts on the face, especially the lips. Greenwood et al. [2018] jointly learnt facial expressions and head poses in terms of landmarks from a forked Bi-directional LSTM network. Most previous audio-to-face-animation work focused on matching speech content and left out style/identity information since the identity is usually bypassed due to mode collapse or averaging during training. In contrast, our approach disentangles audio content and speaker information, and drives landmarks capturing speaker-dependent dynamics.

Lip-sync facial animation. With the increasing power of GPUs, we have seen prolific progress on end-to-end learning from audio to video frames. [Chen et al. 2018] synthesized cropped lip movements for each frames. Chung et al. [2017]; Song et al. [2019]; Vougioukas et al. [2019] generated full natural human face images with GANs or encoder-decoder CNNs. [Pham et al. 2018] estimated blendshape parameters. Taylor et al. [2017]; Zhou et al. [2018] demonstrated audio-driven talking portraits for rigged face models, however, the input cartoon models required manual rigging and retargeting, as well as artist interventions for animating the rest of the head beyond lips. Our method is instead able to automatically animate an input portrait and does not require such manual inputs. In addition, the above methods do not capture speaker identity or style. As a result, if the same sentence is spoken by two different voices, they will tend to generate the same facial animation lacking the dynamics required to make it more expressive and realistic.

"Style"-aware facial head animation. Suwajanakorn et al. [2017] used a re-timing dynamic programming method to reproduce speaker

motion dynamics. However, it was specific to a single subject (Obama), and does not generalize to faces other than Obama's. In another earlier work, Liu et al. [2015] used color, depth and audio to reproduce the facial animation of a speaker recorded with a RGBD sensor. However, it does not generalize to other unseen speakers. Cudeiro et al. [2019] attempted to model speaker style in a latent representation. Thies et al. [2020] encoded personal style in static blendshape bases. Both methods, however, focus on lower facial animation, especially lips, and do not predict head pose. More similar to ours, Zhou et al. [2019] learned a joint audio-visual representation to disentangle the identity and content from the image domain. However, their identity information primarily focus on static facial appearance and not the speaker dynamics. As we demonstrate in §5.5, speaker awareness encompasses many aspects beyond mere static appearances. The individual facial expressions and head movements are both important factors for speaker-aware animations. Our method addresses speaker identity by learning jointly the static appearance and head motion dynamics, to deliver faithfully animated talking-heads.

Warping-based character animation. Fišer et al. [2017] and Averbuch-Elor et al. [2017] demonstrated animation of portraits driven by videos and extracted landmarks of human facial performance. Weng et al. [2019] presented a system for animating the body of an input portrait by fitting a human template, then animated it using motion capture data. In our case, we aim to synthesize facial expressions and head pose from audio alone.

Evaluation metrics. Quantitatively evaluating the learned identity/style is crucial for validation; at the same time, it is nontrivial to setup an appropriate benchmark. Many prior work resorted to subjective user studies [Cudeiro et al. 2019; Karras et al. 2017].

Agarwal et al. [2019] visualized the style distribution via action units. For existing quantitative metrics, they primarily focus on pixel-level artifacts since a majority of the network capacity is used to learn pixel generation rather than high-level dynamics [Chen et al. 2019]. Action units have been proposed to be an alternative evaluation measure of expression in the context of GAN-based approaches [Pumarola et al. 2018]. We propose a new collection of metrics to evaluate the high-level dynamics that matter for facial expression and head motion.

Image-to-image translation. Image-to-image translation is a very common approach to modern talking face synthesis and editing. Face2Face and VDub are among the early explorers to demonstrate robust appearance transfer between two talking-head videos [Garrido et al. 2015; Thies et al. 2016]. Later, adversarial training was adopted to improve the quality of the transferred results. For example, Kim et al. [2019] used cycle-consistency loss to transfer styles and showed promising results on one-to-one transfers. Zakharov et al. [2019] developed a few-shot learning scheme that leveraged landmarks to generate natural human facial animation. Based on these prior works, we also employ an image-to-image translation network to generate natural human talking-head animations. Unlike Zakharov et al. [2019], our model handles generalization to faces unseen during training without the need of fine-tuning. Additionally, we are able to generate non-photorealistic images through an image deformation module.

Disentangled learning. Disentanglement of content and style in audio has been widely studied in the voice conversion community. Without diving into its long history (see [Stylianou 2009] for a detailed survey), here we only discuss recent methods that fit into our deep learning pipeline. Wan et al. [2018] developed *Resemblyzer* as a speaker identity embedding for verification purposes across different languages. Qian et al. [2019] proposed AutoVC, a few-shot voice conversion method to separate the audio into the speech content and the identity information. As a baseline, we use AutoVC for extracting voice content and Resemblyzer for extracting feature embeddings of speaker identities. We introduce the idea of voice conversion to audio-driven animation and demonstrate the advantages of *speaker-aware* talking-head generation.

3 METHOD

Overview. As summarized in Figure 2, given an audio clip and a single facial image, our architecture, called "MakeItTalk", generates a *speaker-aware* talking-head animation synchronized with the audio. In the training phase, we use an off-the-shelf face 3D landmark detector to preprocess the input videos to extract the landmarks [Bulat and Tzimiropoulos 2017]. A baseline model to animate the speech content can be trained from the input audio and the extracted landmarks directly. However, to achieve high-fidelity dynamics, we found that landmarks should instead be predicted from a disentangled content representation and speaker embedding of the input audio signal.

Specifically, we use a voice conversion neural network to disentangle the speech content and identity information [Qian et al.

2019]. The *content* is speaker-agnostic and captures the general motion of lips and nearby regions (Figure 2, *Speech Content Animation*, §3.1). The *identity* of the speaker determines the specifics of the motions and the rest of the talking-head dynamics (Figure 2, *Speaker-Aware Animation*, §3.2). For example, no matter who speaks the word 'Ha!', the lips are expected to be open, which is speaker-agnostic and only dictated by the content. As for the exact shape and size of the opening, as well as the motion of nose, eyes and head, these will depend on who speaks the word, i.e., identity. Conditioned on the content and speaker identity information, our deep model outputs a sequence of predicted landmarks for the given audio.

To generate rasterized images, we discuss two algorithms for the landmark-to-image synthesis (§3.3). For non-photorealistic images like paintings or vector arts (Fig. 7), we use a simple image warping method based on Delaunay triangulation inspired by [Averbuch-Elor et al. 2017] (Figure 2, *Face Warp*). For natural images (Fig. 8), we devised an image-to-image translation network, inspired by pix2pix [Isola et al. 2017]) to animate the given human face image with the underlying landmark predictions (Figure 2, *Image2Image Translation*). Combining all the image frames and input audio together gives us the final talking-head animations. In the following sections, we describe each module of our architecture.

3.1 Speech Content Animation

To extract the speaker-agnostic content representation of the audio, we use AutoVC encoder from Qian et al. [2019]. The AutoVC network utilizes an LSTM-based encoder that compresses the input audio into a compact representation (bottleneck) trained to abandon the original speaker identity but preserve content. In our case, we extract a content embedding $A \in \mathcal{R}^{T \times D}$ from AutoVC network, where T is the total number of input audio frames, and D is the content dimension.

The goal of the content animation component is to map the content embedding A to facial landmark positions with a neutral style. In our experiments, we found that recurrent networks are much better suited for the task than feedforward networks, since they are designed to capture such sequential dependencies between the audio content and landmarks. We experimented with vanilla RNNs and LSTMs [Graves and Jaitly 2014], and found that LSTMs offered better performance. Specifically, at each frame t, the LSTM module takes as input the audio content A within a window $[t \rightarrow t+\tau]$. We set $\tau = 18$ frames (a window size of 0.3s in our experiments). To animate any input 3D static landmarks q, where $q \in \mathcal{R}^{68 \times 3}$ that are extracted using a landmark detector, the output from LSTM layers is fed into a Multi-Layer Perceptron (MLP) and finally predicts displacements Δq_t, which put the input landmarks in motion at each frame.

To summarize, the speech content animation module models sequential dependencies to output landmarks based on the following transformations:

$$c_t = LSTM_c(A_{t \rightarrow t+\tau}; w_{lstm,c}), \tag{1}$$

$$\Delta q_t = MLP_c(c_t, q; w_{mlp,c}), \tag{2}$$
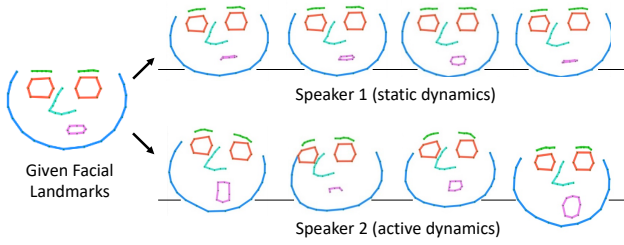
$$p_t = q + \Delta q_t, \tag{3}$$

Fig. 3. Landmark prediction for different speaker identities. Left: static facial landmarks from a given portrait image. Right-top: predicted landmark sequence from a speaker who tends to be conservative in terms of head motion. Right-bottom: predicted landmark sequence from another speaker who tends to be more active.

where $\{w_{lstm,c}, w_{mlp,c}\}$ are learnable parameters for the LSTM and MLP networks respectively. The LSTM has three layers of units, each having an internal hidden state vector of size 256. The decoder MLP network has three layers with internal hidden state vector size of 512, 256 and 204 (68×3), respectively.

3.2 Speaker-Aware Animation

Matching just the lip motion to the audio content is not sufficient. The motion of the head or the subtle correlation between mouth and eyebrows are also crucial clues to generate plausible talking-heads. For example, Figure 3 shows our speaker-aware predictions for two different speaker embeddings: one originates from a speaker whose head motion tends to be more static, and another that is more active (see also our supplementary video). Our method successfully differentiates the head motion dynamics between these two speakers.

To achieve this, we extract the speaker identity embedding with a speaker verification model [Wan et al. 2018] which maximizes the embedding similarity among different utterances of the same speaker, and minimizes the similarity among different speakers. The original identity embedding vector s has a size of 256. We found that reducing its dimensionality from 256 to 128 via a single-layer MLP improved the generalization of facial animations especially for speakers not observed during training. Given the identity embedding s extracted, we further modulate the per-frame landmarks p_t such that they reflect the speaker's identity. More specifically, the landmarks are perturbed to match the head motion distribution and facial expression dynamics observed in speakers during training. In this manner, our method reproduces a speaker-specific distribution of plausible head movements reflected by the modulated landmarks.

As shown in the bottom stream of Figure 2, we first use an LSTM to encode the content representation within time windows, which has the same network architecture and time window length as the LSTM used in the speech content animation module. We found, however, that it is better to have different learned parameters for this LSTM, such that the resulting representation $\tilde{c}_t$ is more tailored for capturing head motion and facial expression dynamics:

$$\tilde{c}_t = LSTM_s(A_{t \to t+\tau}; w_{lstm,s}), \tag{4}$$

where $w_{lstm,s}$ are trainable parameters. Then, the following model takes as input the speaker embedding s, the audio content representation $\tilde{c}_t$, and the initial static landmarks q to generate speaker-aware landmark displacement. Notably, we found that producing coherent head motions and facial expressions requires capturing longer time-dependencies compared to the speech content animation module. While phonemes typically last for a few tens of milliseconds, head motions, e.g., a head swinging left-to-right, may last for one or few seconds, several magnitudes longer. To capture such long and structured dependencies, we adopted a self-attention network [Devlin et al. 2018; Vaswani et al. 2017]. The self-attention layers compute an output expressed as a weighted combination of learned per-frame representations i.e., the audio content representation $\tilde{c}_t$ extracted by the above LSTM concatenated with the speaker embedding s. The weight assigned to each frame is computed by a compatibility function comparing all-pairs frame representations within a window. We set the window size to $\tau' = 256$ frames (4 sec) in all experiments. The output from the last self-attention layer and the initial static landmarks are processed by an MLP to predict the final per-frame landmarks.

Mathematically, our speaker-aware animation models structural dependencies to perturb landmarks that capture head motion and personalized expressions, which can be formulated as follows:

$$h_t = Attn_s(\tilde{c}_{t \to t+\tau'}, s; w_{attn,s}), \tag{5}$$

$$\Delta p_t = MLP_s(h_t, q; w_{mlp,s}), \tag{6}$$

$$y_t = p_t + \Delta p_t, \tag{7}$$

where $\{w_{attn,s}, w_{mlp,c}\}$ are trainable parameters of the self-attention encoder and MLP decoder, p_t is computed by Eq. (3), and y_t are the final per-frame landmarks capturing both speech content and speaker identity. In our implementation, the attention network follows the encoder block in Vaswani et al. [2017]. More details about its architecture are provided in the appendix.

3.3 Single-Image Animation

The last step of our model creates the final animation of the input portrait. Given an input image Q and the predicted landmarks $\{y_t\}$ for each frame t, we produce a sequence of images $\{F_t\}$ representing the facial animation. The input portrait might either depict a cartoon face, or a natural human face image. We use different implementations for each of these two types of portraits. In the next paragraphs, we explain the variant used for each type.

Cartoon Images (non-photorealistic). These images usually have sharp feature edges, e.g., from vector arts or flat shaded drawings. To preserve these sharp features, we propose a morphing-based method to animate them, avoiding pixel-level artifacts. From the input image, we extract the facial landmarks using [Yaniv et al. 2019]. We then run Delaunay triangulation on these landmarks to generate semantic triangles. By mapping the initial pixels as texture maps to the triangles, the subsequent animation process becomes straightforward. As long as the landmark topology remains the same, the textures on each triangle naturally transfer across frames. An illustration is shown in Figure 4. An analogy with our approach is the vertex and fragment shader pipeline in rendering. The textures are bind to each fragment at the very beginning and
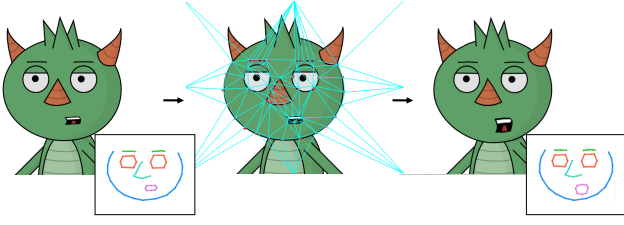
Fig. 4. Cartoon image face warping through facial landmarks and Delaunay Triangulation. Left: Given cartoon image and its facial landmarks. Middle: Delaunay triangulation. Right: Warped image guided by the displaced landmarks.

from then on, only the vertex shader is changing the location of these vertices (landmark positions). In practice, we implement a GLSL-based C++ code that uses vertex/fragment shaders and runs in real-time.

Natural Images. The goal here is to synthesize a sequence of frames given the input photo and the predicted animated landmarks from our model. Our image synthesis approach is inspired by the landmark-based facial animation from Zakharov et al. [2019], which translates landmarks to natural images based on a trained network. Instead of using a separate embedder and adaptive instance normalization layers to encode the target face appearance, our module is designed based on the UNet architecture to process displaced landmarks and portraits. Specifically, we first create an image representation Y_t of the predicted landmarks y_t by connecting consecutive facial landmarks and rendering them as line segments of predefined color (Figure 2). The image Y_t is concatenated channel-wise with the input portrait image Q to form a 6-channel image of resolution 256×256. The image is passed to an encoder-decoder network that performs image translation to produce the image F_t per frame. Its architecture follows the generators proposed in Esser et al. [2018] and Han et al. [2019]. Specifically, the encoder employs 6 convolutional layers, where each layer contains one 2-strided convolution followed by two residual blocks, and produces a bottleneck, which is then decoded through symmetric upsampling decoders. Skip connections are utilized between symmetric layers of the encoder and decoder, as in U-net architectures [Ronneberger et al. 2015]. The generation proceeds for each frame. Since the landmarks change smoothly over time, the output images formed as an interpolation of these landmarks exhibit temporal coherence. Examples of generated image sequences are shown in Figure 8.

4 TRAINING

We now describe our training procedure to learn the parameters of each module in our architecture.

Voice Conversion Training. We follow the training setup described in [Qian et al. 2019] with the speaker embedding initialized by the pretrained model provided by Wan et al. [2018]. A training source speech from each speaker is processed through the content encoder. Then another utterance of the same source speaker is used to extract the speaker embedding, which is passed to the decoder along with the audio content embedding to reconstruct the original source
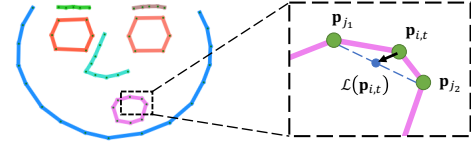
Fig. 5. Graph Laplacian coordinates illustration. Left: 8 facial parts that contain subsets of landmarks. Right: Zoom-in graph Laplacian vector and related neighboring landmark points.

speech. The content encoder, decoder, and MLP are trained to minimize the self-reconstruction error of the source speech spectrograms [Qian et al. 2019]. Training is performed on the *VCTK* corpus [Veaux et al. 2016], which is a speech dataset including utterances by 109 native English speakers with various accents.

4.1 Speech Content Animation Training

Dataset. To train a content-based animation model, we use an audio-visual dataset that provides high-quality facial landmarks and corresponding audio. To this end, we use the *Obama Weekly Address* dataset [Suwajanakorn et al. 2017] containing 6 hours of video featuring various Obama's speeches. Due to its high resolution and relatively consistent front facing camera angle, we can obtain accurate facial landmark detection results using [Bulat and Tzimiropoulos 2017]. We also register the facial landmarks to a front-facing standard facial template [Beeler and Bradley 2014] using a best-estimated affine transformation [Segal et al. 2009]. This also results in factoring out the speaker-dependent head pose motion. We emphasize that one registered speaker is enough to train this module, since our goal here is to learn a generic lip motion based on audio. The lip sync will be specialized to particular speaker IDs by including the speaker-aware branch and training on multiple human subjects, as we will explain in §4.2.

Loss function. To learn the parameters $\{w_{lstm,c}, w_{mlp,c}\}$ used in the LSTM and MLP, we minimize a loss function that evaluates (a) the distance between the registered reference landmark positions $\hat{p}_t$ and predicted ones p_t, and (b) the distance between their respective graph Laplacian coordinates, which promotes correct placement of landmarks with respect to each other and preserves facial shape details [Sorkine 2006]. Specifically, our loss is:

$$L_c = \sum_{t=1}^{T}\sum_{i=1}^{N}\left\|p_{i,t} - \hat{p}_{i,t}\right\|_2^2 + \lambda_c \sum_{t=1}^{T}\sum_{i=1}^{N}\left\|\mathcal{L}(p_{i,t}) - \mathcal{L}(\hat{p}_{i,t})\right\|_2^2, \quad (8)$$

where i is the index for each individual landmark, and λ_c weighs the second term ($\lambda_c = 1$ in our implementation, set through hold-out validation). We use the following graph Laplacian $\mathcal{L}(p_t)$:

$$\mathcal{L}(p_{i,t}) = p_{i,t} - \frac{1}{|\mathcal{N}(p_i)|}\sum_{p_j \in \mathcal{N}(p_i)} p_{j,t}, \quad (9)$$

where $\mathcal{N}(p_i)$ includes the landmark neighbors connected to p_i within a distinct facial part (Figure 5). We use 8 facial parts that contain subsets of landmarks predefined for the facial template.

4.2 Speaker-Aware Animation Training

Dataset. To learn the speaker-aware dynamics of head motion and facial expressions, we need an audio-visual dataset featuring a diverse set of speakers. We found the *VoxCeleb2* dataset is well-suited for our purpose since it contains video segments from a variety of speakers [Chung et al. 2018]. VoxCeleb2 was originally designed for speaker verification. Since our goal is different, i.e., capturing speaker dynamics for talking head synthesis, we chose a subset of 67 speakers with a total of 1,232 videos clips from Vox-Celeb2. On average, we have around 5-10 minutes of videos for each speaker. The criterion for selection was accurate landmark detection in the videos based on manual verification. Speakers were selected based on Poisson disk sampling on the speaker representation space. We split this dataset as 60% / 20% / 20% for training, hold-out validation and testing respectively. In contrast to the content animation step, we do not register the landmarks to a front-facing template since here we are interested in learning the overall head motion.

Adversarial network. Apart from capturing landmark position, we also aim to match the speaker's head motion and facial expression dynamics during training. To this end, we incorporate a GAN approach. Specifically, we create a discriminator network $Attn_d$ which follows a similar structure with the self-attention generator network in §4.2. More details about its architecture are provided in the appendix. The goal of the discriminator is to find out if the temporal dynamics of the speaker's facial landmarks look "realistic" or fake. It takes as input the sequence of facial landmarks within the same window used in the generator, along with audio content and speaker's embedding. It returns an output characterizing the "realism" r_t per frame t:

$$r_t = Attn_d(y_{t \to t+\tau'}, \tilde{c}_{t \to t+\tau'}, s; w_{\text{attn},d}), \qquad (10)$$

where $w_{\text{attn},d}$ are the parameters of the discriminator. We use the LSGAN loss function [Mao et al. 2017] to train the discriminator parameters treating the training landmarks as "real" and the generated ones as "fake" for each frame:

$$L_{gan} = \sum_{t=1}^{T} (\hat{r}_t - 1)^2 + r_t^2, \qquad (11)$$

where $\hat{r}_t$ denotes the discriminator output when the training landmarks $\hat{y}_t$ are used as its input.

Loss function. To train the parameters $w_{\text{attn},s}$ of the self-attention generator network, we attempt to maximize the "realism" of the output landmarks, and also consider the distance to the training ones in terms of absolute position and Laplacian coordinates:

$$L_s = \sum_{t=1}^{T} \sum_{i=1}^{N} \left\| y_{i,t} - \hat{y}_{i,t} \right\|_2^2 + \lambda_s \sum_{t=1}^{T} \sum_{i=1}^{N} \left\| \mathcal{L}(y_{i,t}) - \mathcal{L}(\hat{y}_{i,t}) \right\|_2^2$$
$$+ \mu_s \sum_{t=1}^{T} (r_t - 1)^2, \qquad (12)$$

where $\lambda_s = 1$ and $\mu_s = 0.001$ are set through hold-out validation. We alternate training between the generator (minimizing L_s) and

the discriminator (minimizing L_{gan}) to improve each other as done in GAN approaches [Mao et al. 2017].

4.3 Image-to-Image Translation Training

Finally, we train our image-to-image translation module to handle natural human face animation outputs. The encoder/decoder pair used for image translation is first trained on paired video frames from VoxCeleb2. Then, we fine-tune the network on a subset which contains high-resolution video crops provided by Siarohin et al. [2019]. In particular, based on a video of a talking person in the dataset, we randomly sample a frame pair: a source training frame $\hat{Q}_{src}$ and a target frame $\hat{Q}_{trg}$ of this person. The facial landmarks of the target face are extracted and rasterized into an RGB image $\hat{Y}_{trg}$ based on the approach described in §3.3. The encoder/decoder network takes as input the concatenation of $\hat{Q}_{src}$ and $\hat{Y}_{trg}$ and outputs a reconstructed face Q_{trg}. The loss function aims to minimize the L^1 per-pixel distance and perceptual feature distance between the reconstructed face Q_{trg} and training target face $\hat{Q}_{trg}$ as in [Johnson et al. 2016]:

$$L_a = \sum_{\{\text{src,trg}\}} ||Q_{\text{trg}} - \hat{Q}_{\text{trg}}||_1 + \lambda_a \sum_{\{\text{src,trg}\}} ||\phi(Q_{\text{trg}}) - \phi(\hat{Q}_{\text{trg}})||_1,$$

where $\lambda_a = 1$, and ϕ concatenates feature map activations from the pretrained VGG19 network [Simonyan and Zisserman 2014].

4.4 Implementation Details

All landmarks in our dataset are converted to 62.5 frames per second and audio waveforms are sampled under $16K$ Hz frequency. Both of these rates followed Qian et al. [2019], i.e. 62.5 Hz for the mel-spetrogram and 16 kHz for the speech waveform. We experimented with other common frame rates, and we found the above worked well for the entire pipeline. We note that the facial landmarks are extracted from the input video at its original frame rate and the interpolation is performed on landmarks rather than the original pixels. We trained both the speech content animation and speaker-aware animation modules with the Adam optimizer using PyTorch. The learning rate was set to 10^{-4}, and weight decay to 10^{-6}. The speech content animation module contains $1.9M$ parameters and took 12 hours to train on a Nvidia 1080Ti GPU. The speaker-aware animation module has $3.8M$ parameters and took 30 hours to train on the same GPU. The single-face animation module for generating natural human faces was also trained with the Adam optimizer, a learning rate of 10^{-4}, and a batch size of 16. The network has $30.7M$ parameters and was trained for 20 hours on 8 Nvidia 1080Ti GPUs. At test time, our network produces natural human face videos at around 22 FPS, and animations for non-photorealistic videos at around 28 FPS.

Our source code and models are available at the following project page: https://people.umass.edu/ yangzhou/MakeItTalk/

5 RESULTS

With all the pieces in place, we now present results of talking-head videos from a single input image and a given audio file. Figure 6

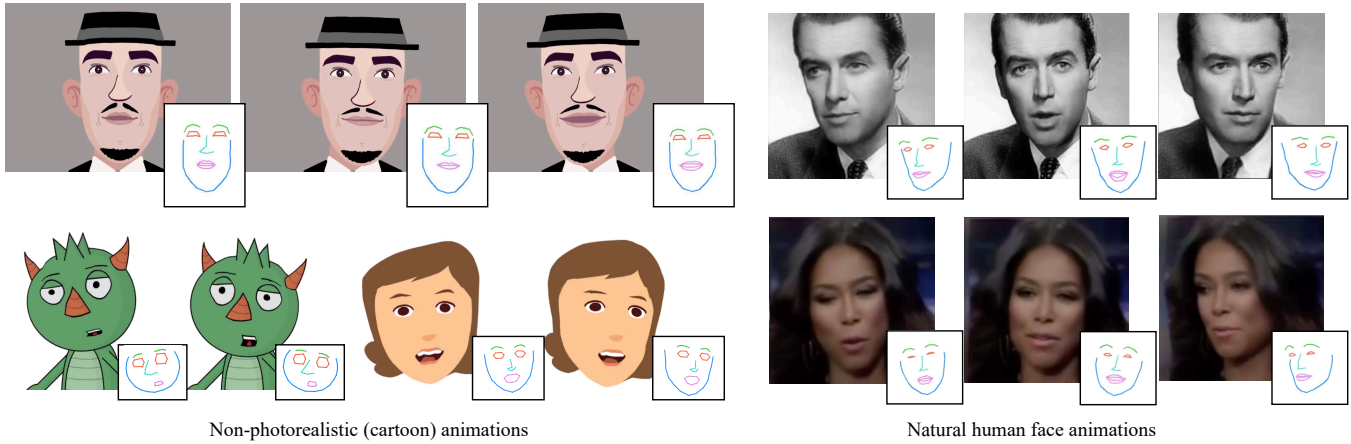Non-photorealistic (cartoon) animations   Natural human face animations

Fig. 6. Generated talking-head animation gallery for non-photorealistic cartoon faces (left) and natural human faces (right). The corresponding intermediate facial landmark predictions are also shown on the right-bottom corner of each animation frame. Our method synthesizes not only facial expressions, but also different head poses. Cartoon *Man with hat* and *Girl with brown hair* ©*Yang Zhou*. Natural face (at right bottom corner) from *VoxCeleb2* dataset [Chung et al. 2018] ©*Visual Geometry Group* (CC BY).
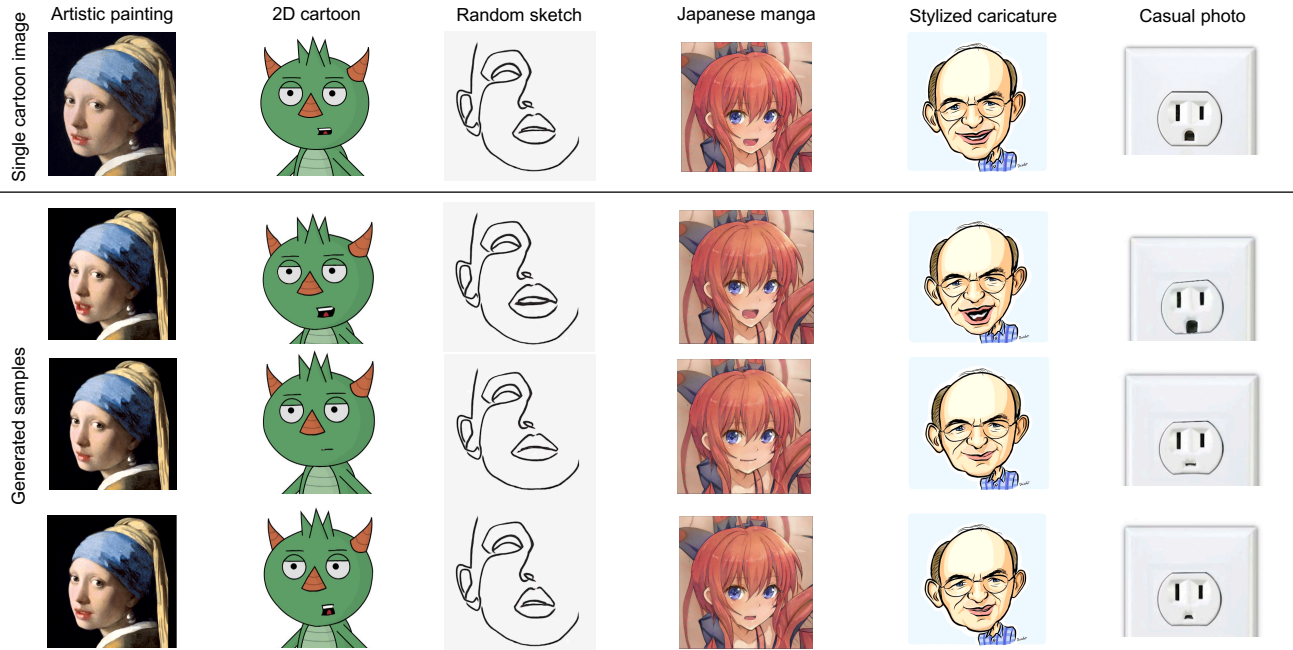

Fig. 7. Our model works for a variety types of non-photorealistic (cartoon) portrait images, including artistic paintings, 2D cartoon characters, random sketches, Japanese mangas, stylized caricatures and casual photos. Top row: input cartoon images. Next rows: generated talking face examples by face warping. Please also see our supplementary video. Artistic painting *Girl with a pearl earring* ©*Johannes Vermeer* (public domain). *Random sketch* ©*Yang Zhou*. *Japanese manga* ©*Gwern Branwen* (CC-0). Stylized caricature *Roy* ©*Adobe Research*.

shows a gallery of our generated animations for cartoon and natural human face images. We note that the resulting animations include both full facial expressions and dynamic head poses. We refer readers to our supplementary video.

In the following sections, we discuss more results for generating non-photorealistic animations, and natural human facial video,

along with qualitative comparisons. Then we present detailed numerical evaluation, an ablation study, and applications of our method.

5.1 Animating Non-Photorealistic Images

Figure 7 shows a gallery of our generated non-photorealistic animations. Each animation is generated from input audio and a single
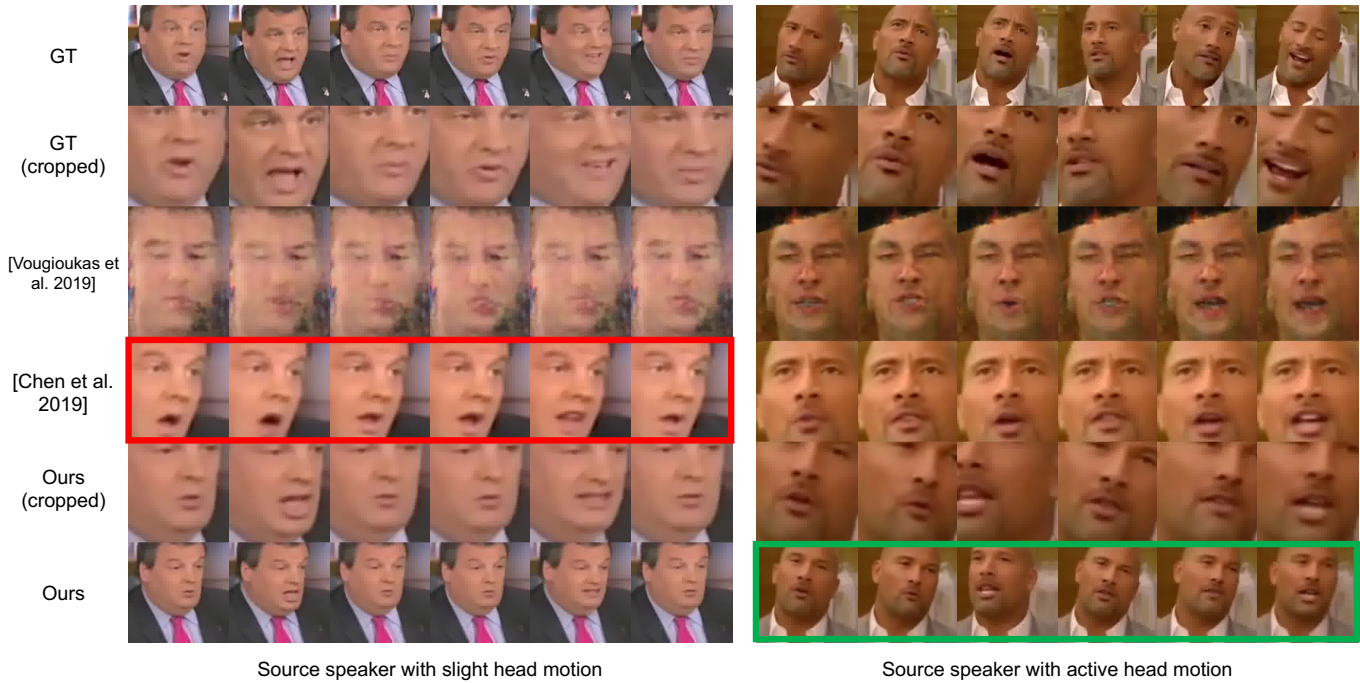
Fig. 8. Comparison with state-of-the-art methods for video generation of human talking-heads. The compared methods crop the face and predict primarily the lip region while ours generates both facial expression and head motion. GT and our results are full faces and are cropped for a better visualization of the lip region. Left example: Chen et al. [2019] has worse lip synchronization for side-faces (see the red box). Right example: our method predicts speaker-aware head pose dynamics (see the green box). Note that the predicted head pose is different than the one in the ground-truth video, but it exhibits similar *dynamics* that are characteristic for the speaker. Natural faces from *VoxCeleb2* dataset [Chung et al. 2018] ©*Visual Geometry Group* (CC BY).

portrait image. The portrait images can be artistic paintings, random sketches, 2D cartoon characters, Japanese mangas, stylized caricatures and casual photos. Despite being only trained on human facial landmarks, our method can successfully generalize to a large variety of stylized cartoon faces. This is because our method uses landmarks as intermediate representations, and also learns relative landmark displacements instead of absolute positions.

5.2 Animating Human Facial Images

Figure 8 shows synthesized talking head videos featuring talking people as well as comparisons with state-of-the-art video generation methods [Chen et al. 2019; Vougioukas et al. 2019]. The ground-truth (GT) and our results are cropped to highlight the differences in the lip region (see row 2 and 5). We notice that the videos generated by Chen et al. [2019]; Vougioukas et al. [2019] predict primarily the lip region on cropped faces and therefore miss the head poses. Vougioukas et al. [2019] does not generalize well to faces unseen during training. Chen et al. [2019] lacks synchronization with the input audio, especially for side-facing portraits (see the red box). Compared to these methods, our method predicts facial expressions more accurately and also captures head motion to some degree (see the green box). On the other hand, to be fair, we also note that our method is not artifact-free: the head motion often distorts the background, which is due to the fact that our network attempts

to produce the whole image, without separating foreground from background.

Even only trained on natural face images, our image translation module can also generate plausible facial animations not only limited to real faces, but also to 2D paintings, picture of statue heads, or rendered images of 3D models. Please check our supplementary video for results.

Our supplementary video also includes a comparison with the concurrent work by [Thies et al. 2020]. Given the same audio and target speaker image at test time, we found that our lip synchronization appears to be more accurate than their method. We also emphasize that our method learns to generate head poses, while in their case the head pose is not explicitly handled or is added back heuristically in a post-processing step (not detailed in their paper). Their synthesized video frames appear sharper than ours perhaps due to their neural renderer of their 3D face model. However, their method requires additional training on target-specific reference videos of length around 30 hours, while ours animates a single target photo immediately without any retraining. Thus, our network has the distinctive advantage of driving a diverse set of single images for which long training videos are not available. These include static cartoon characters, casual photos, paintings, and sketches.

5.3 Evaluation Protocol

We evaluated MakeItTalk and compared with related methods quantitatively and with user studies. We created a test split from the VoxCeleb2 subset, containing 268 video segments from 67 speakers. The speaker identities were observed during training, however, their test speech and video are different from the training ones. Each video clip lasts 5 to 30 seconds. Landmarks were extracted using [Bulat and Tzimiropoulos 2017] from test clips and their quality was manually verified. We call these as "reference landmarks" and use them in the evaluation metrics explained below.

Evaluation Metrics. To evaluate how well the synthesized landmarks represent accurate lip movements, we use the following metrics:

- **Landmark distance for jaw-lips (D-LL)** represents the average Euclidean distance between predicted facial landmark locations of the jaw and lips and reference ones. The landmark positions are normalized according to the maximum width of the reference lips for each test video clip.
- **Landmark velocity difference for jaw-lips (D-VL)** represents the average Euclidean distance between reference landmark velocities of the jaw and lips and predicted ones. Velocity is computed as the difference of landmark locations between consecutive frames. The metric captures differences in first-order jaw-lips dynamics.
- **Difference in open mouth area (D-A:)**: the average difference between the area of the predicted mouth shape and reference one. It is expressed as percentage of the maximum area of the reference mouth for each test video clip.

To evaluate how well the landmarks produced by our method and others reproduce overall head motion, facial expressions, and their dynamics, we use the following metrics:

- **Landmark distance (D-L)**: the average Euclidean distance between all predicted facial landmark locations and reference ones (normalized by the width of the face).
- **Landmark velocity difference (D-V)**: the average Euclidean distance between reference landmark velocities and predicted ones (again normalized by the width of the face). Velocity is computed as the difference of landmark locations between consecutive frames. This metric serves as an indicator of landmark motion dynamics.
- **Head rotation and position difference (D-Rot/Pos)**: the average difference between the reference and predicted head rotation angles (measured in degrees) and head position (again normalized by the width of the face). The measure indicates head pose differences, like nods and tilts.

5.4 Content Animation Evaluation

We begin our evaluation by comparing MakeItTalk with state-of-the-art methods for synthesis of facial expressions driven by landmarks. Specifically, we compare with Eskimez et al. [2018], Zhou et al. [2018], and Chen et al. [2019]. All these methods attempt to synthesize facial landmarks, but cannot produce head motion. Head movements are either generated procedurally or copied from a source video. Thus, to perform a fair evaluation, we factor out

Table 2. Quantitative comparison of facial landmark predictions of MakeItTalk versus state-of-the-art methods.

Methods	D-LL ↓	D-VL ↓	D-A ↓
[Zhou et al. 2018]	6.2%	0.63%	15.2%
[Eskimez et al. 2018]	4.0%	0.42%	7.5%
[Chen et al. 2019]	5.0%	0.41%	5.0%
Ours (no separation)	2.9%	0.64%	17.1%
Ours (no speaker branch)	2.2%	0.29%	5.9%
Ours (no content branch)	3.1%	0.38%	10.2%
Ours (full)	**2.0%**	**0.27%**	**4.2%**

head motion from our method, and focus only on comparing predicted landmarks under an identical "neutral" head pose for all methods. For the purpose of this evaluation, we focus on the lip synchronization metrics (**D-LL**, **D-VL**, **D-A**), and ignore head pose metrics. Quantitatively, Table 2 reports these metrics for the above-mentioned methods and ours. We include our full method, including three reduced variants: (a) *"Ours (no separation)"*, where we eliminate the voice conversion module and feed the raw audio features as input directly to the speaker-aware animation branch trained and tested alone; in this manner, there is no separation (disentanglement) between audio content and speaker identity, (b) *"Ours (no speaker branch)"*, where we keep the voice conversion module for disentanglement, but we train and test the speech content branch alone without the speaker-aware branch, (c) *"Ours (no content branch)"*, where we again perform disentanglement, but we train and test the speaker-aware branch alone without the speech content branch. We discuss these three variants in more detail in our ablation study (Section 5.6). The result shows that our method achieves the lowest errors for all measures. In particular, our method has 2x times less **D-LL** error in lip landmark positions compared to [Eskimez et al. 2018], and 2.5x times less **D-LL** error compared to [Chen et al. 2019].

Figure 9 shows characteristic examples of facial landmark outputs for the above methods and ours from our test set. Each row shows one output frame. Zhou et al. [2018] is only able to predict the lower part of the face and cannot reproduce closed mouths accurately (see second row). Eskimez et al. [2018] and Chen et al. [2019], on the other hand, tend to favor conservative mouth opening. In particular, Chen et al. [2019] predicts bottom and upper lips that sometimes overlap with each other (see second row, red box). In contrast, our method captures facial expressions that match better the reference ones. Ours can also predict subtle facial expressions, such as the lip-corner lifting (see first row, red box).

5.5 Speaker-Aware Animation Evaluation

Head Pose Prediction and Speaker Awareness. Existing speech-driven facial animation methods do not synthesize head motion. Instead, a common strategy is to copy head poses from another existing video. Based on this observation, we evaluate our method against two baselines: *"retrieve-same ID"* and *"retrieve-random ID"*. These baselines retrieve the head pose and position sequence from another video clip randomly picked from our training set. Then the facial landmarks are translated and rotated to reproduce the copied head
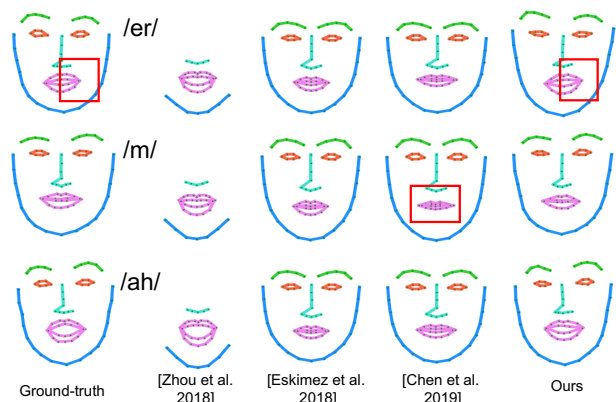
Fig. 9. Facial expression landmark comparison. Each row shows an example frame prediction for different methods. The GT landmark and uttered phonemes are shown on left.

poses and positions. The first baseline *"retrieve-same ID"* uses a training video with the same speaker as in the test video. This strategy makes this baseline stronger since it re-uses dynamics from the same speaker. The second baseline *"retrieve-random ID"* uses a video from a different random speaker. This baseline is useful to examine whether our method and alternatives produce head pose and facial expressions better than random or not.

Table 3 reports the **D-L**, **D-V**, and **D-Rot/Pos** metrics. Our full method achieves much smaller errors compared to both baselines, indicating our speaker-aware prediction is more faithful compared to merely copying head motion from another video. In particular, we observe that our method produces $2.7x$ less error in head pose (D-Rot), and $1.7x$ less error in head position (D-Pos) compared to using a random speaker identity (see *"retrieve-random ID"*). This result also confirms that the head motion dynamics of random speakers largely differ from ground-truth ones. Compared to the stronger baseline of re-using video from the same speaker (see *"retrieve-same ID"*), we observe that our method still produces $1.3x$ less error in head pose (D-Rot), and $1.5x$ less error in head position (D-Pos). This result confirms that re-using head motion from a video clip even from the right speaker still results in significant discrepancies, since the copied head pose and position does not necessarily synchronize well with the audio. Our full method instead captures the head motion dynamics and facial expressions more consistently w.r.t. the input audio and speaker identity.

Figure 6 shows a gallery of our generated cartoon images and natural human faces under different predicted head poses. The corresponding generated facial landmarks are also shown on the right-bottom corner of each image. The demonstrated examples show that our method is able to synthesize head pose well, including nods and swings. Figure 10 shows another qualitative validation of our method's ability to capture personalized head motion dynamics. The figure embeds 8 representative speakers from our dataset based on their variance in Action Units (AUs), head pose and position variance. The AUs are computed from the predicted landmarks based on the definitions from [Ekman and Friesen 1978]. The embedding is performed through t-SNE [Maaten and Hinton 2008]. These 8 representatives were selected using furthest sampling i.e.,
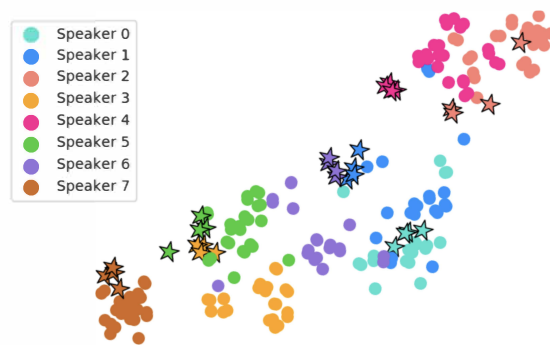

Fig. 10. t-SNE visualization for AUs, head pose and position variance based on 8 reference speakers videos (solid dots) and our predictions (stars). Different speakers are marked with different colors as shown in the legend.

Table 3. Head pose prediction comparison with the baseline methods in §5.5 and our variants based on our head pose metrics.

Methods	D-L ↓	D-V ↓	D-Rot/Pos ↓
retrieve-same ID	17.1%	1.2%	10.3/8.1%
retrieve-random ID	20.8%	1.1%	21.4/9.2%
Ours (no separation)	12.4%	1.1%	8.8/5.4%
Ours (random ID)	33.0%	2.4%	28.7/12.3%
Ours (no speaker branch)	13.8%	1.2%	12.6/6.9%
Ours (no content branch)	12.5%	0.9%	8.6/5.7%
Ours (full)	**12.3%**	**0.8%**	**8.0/5.4%**

their AUs, head pose and position differ most from the rest of the speakers in our dataset. We use different colors for different speakers and use *solid dots* for embeddings produced based on the reference videos in AUs, head pose and position variance, and *stars* for embeddings resulting from our method. The visualization demonstrates that our method produces head motion dynamics that tend to be located more closely to reference ones.

5.6 Ablation study

Individual branch performance. We performed an ablation study by training and testing the three reduced variants of our network described in §5.4: *"Ours (no separation)"* (no disentanglement between content and speaker identity), *"Ours (no speaker branch)"*, and *"Ours (no content branch)"*. The aim of the last two variants is to check whether a single network can jointly learn both lip synchronization and speaker-aware head motion. The results of these three variants and our full method are shown in in Table 2 and Table 3. We also refer readers to the supplementary video for more visual comparisons.

The variant *"Ours (no speaker branch)"* only predicts facial landmarks from the audio content without considering the speaker identity. It performs well in terms of capturing the lip landmarks, since these are synchronized with the audio content. The variant is slightly worse than our method based on the lip evaluation metrics (see Table 2). However, it results in $1.6x$ larger errors in head pose and $1.3x$ larger error in head position (see Table 3) since head motion is a function of both speaker identity and content.
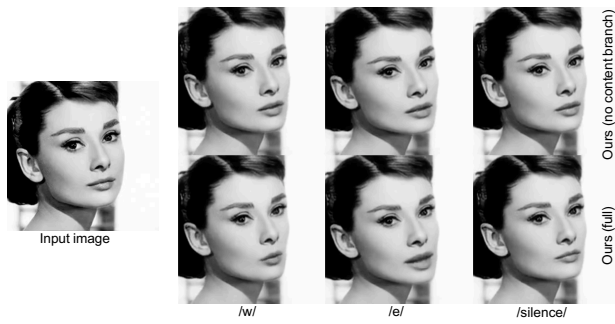
Fig. 11. Comparison to *"Ours (no content branch)"* variant (right-top) which uses only the *speaker-aware animation* branch. The full model (right-bottom) result has much better articulation in the lower-part of the face. It demonstrates that a single network architecture cannot jointly learn both lip synchronization and speaker-aware head motion. *Audrey Hepburn ©Me Pixels* (CC-0).

The results of the variant *"Ours (no content branch)"* has the opposite behaviour: it performs well in terms of capturing head pose and position (it is slightly worse than our method, see Table 3). However, it has $1.6x$ higher error in jaw-lip landmark difference and $2.4x$ higher error in open mouth area difference (see Table 2), which indicates that the lower part of face dynamics are not synchronized well with the audio content. Figure 11 demonstrates that using the speaker-aware animation branch alone i.e., the "Ours (no content)" variant results in noticeable artifacts in the jaw-lip landmark displacements. Using both branches in our full method offers the best performance according to all evaluation metrics.

The results of the variant *"Ours (no separation)"* are similar to the variant *"Ours (no content branch)"*: it achieves slightly worse head pose performance than our full method (Table 3), and much worse results in terms of lip movement accuracy (Table 2). Specifically, it has $1.5x$, $2.4x$, and $4.1x$ higher error in jaw-lip landmark position, velocity, and open mouth area difference respectively. We hypothesize this is because the content and the speaker identity information are still entangled and therefore it is hard for the network to disambiguate a one-to-one mapping between audio and landmarks.

Random speaker ID injection. We tested one more variant of our method called *"Ours (random ID)"*. For this variant, we use our full network, however, instead of using the correct speaker embedding, we inject another random speaker identity embedding. The result of this variant is shown in Table 3. Again we observe that the performance is significantly worse ($3.6x$ more error for head pose). This indicates that our method successfully splits the content and speaker-aware motion dynamics, and captures the correct speaker head motion dynamics (i.e., it does not reproduce random ones).

5.7 User Studies

We also evaluated our method through perceptual user studies via Amazon Mechanical Turk service. We obtained 6480 query responses from 324 different MTurk participants in our two different user studies described below.

User study for speaker awareness. Our first study evaluated the *speaker awareness* of different variants of our method while synthesizing cartoon animations. Specifically, we assembled a pool of 300 queries displayed on different webpages. On top of the webpage, we showed a reference video of a real person talking, and on the bottom we showed two cartoon animations: one cartoon animation generated using our full method and another cartoon animation based on one of the two variants discussed above: (*"Ours (random ID)"* and *"Ours (no speaker ID)"*. The two cartoon videos were placed in randomly picked left/right positions for each webpage. Each query included the following question: "Suppose that you want to see the real person of the video on the top as the cartoon character shown on the bottom. Which of the two cartoon animations best represents the person's talking style in terms of facial expressions and head motion?" The MTurk participants were asked to pick one of the following choices: "left animation", "right animation", "can't tell - both represent the person quite well", "can't tell - none represent the person well". Each MTurk participant was asked to complete a questionnaire with 20 queries randomly picked from our pool. Queries were shown at a random order. Each query was repeated twice (i.e., we had 10 unique queries per questionnaire), with the two cartoon videos randomly flipped each time to detect unreliable participants giving inconsistent answers. We filtered out unreliable MTurk participants who gave two different answers to more than 5 out of the 10 unique queries in the questionnaire, or took less than a minute to complete it. Each participant was allowed to answer one questionnaire maximum to ensure participant diversity. We had 90 different, reliable MTurk participants for this user study. For each of our 300 queries, we got votes from 3 different MTurk participants. Since each MTurk participant voted for 10 unique queries twice, we gathered 1800 responses (300 queries × 3 votes × 2 repetitions) from our 90 MTurk participants. Figure 12 (top) shows the study result. We see that the majority of MTurkers picked our full method more frequently, when compared with either of the two variants.

User study for natural human facial video. To validate our landmark-driven *human facial animation* method, we conducted one more user study. Each MTurk participant was shown a questionnaire with 20 queries involving random pairwise comparisons out of a pool with 780 queries we generated. For each query, we showed a single frame showing the head of a real person on top, and two generated videos below (randomly placed at left/right positions): one video synthesized from our method, and another from either Vougioukas et al. [2019] or Chen et al. [2019]. The participants were asked which person's facial expression and head motion look more realistic and plausible. We also explicitly instruct them to ignore the particular camera position or zoom factor and focus on the face. Participants were asked to pick one of four choices ("left", "right", "both" or "none") as in the previous study. We also employed the same random and repeated query design and MTurker consistency and reliability checks to filter out unreliable answers. We had 234 different MTurk participants for this user study. Like in the previous study, each query received votes from 3 different, reliable MTurk participants. As a result, we gathered 780 queries × 3 votes
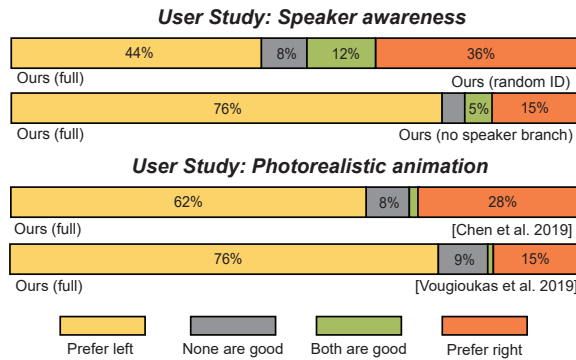
Fig. 12. User study results for speaker awareness (top) and natural human facial animation (bottom).

× 2 repetitions = 4680 responses from our 234 participants. Figure 12 (bottom) shows the study result. Our method was voted as the most "realistic" and "plausible" by a large majority, when compared to Chen et al. [2019] or Vougioukas et al. [2019].

5.8 Applications

Synthesizing plausible talking-heads has numerous applications [Kim et al. 2019; Zhou et al. 2019]. One common application is dubbing using voice from a person different from the one in the original video, or even using voice in a different language. In Figure 13(top), given a single frame of the target actor and a dubbing audio spoken by another person, we can generate a video of the target actor talking according to that other person's speech.

Another application is bandwidth-limited video conference. In scenarios where the visual frames cannot be delivered with high fidelity and frame-rate, we can make use only of the audio signal to drive the talking-head video. Audio signal can be preserved under much lower bandwidth compared to its visual counterpart. Yet, it is still important to preserve visual facial expressions, especially lip motions, since they heavily contribute to understanding in communication [McGurk and MacDonald 1976]. Figure 13(middle) shows that we can synthesize talking heads with facial expressions and lip motions with only the audio and an initial high-quality user profile image as input. Figure 13(bottom) shows examples of both natural human and cartoon talking-head animation that can be used in teleconferencing for entertainment reasons, or due to privacy concerns related to video recording. We also refer readers to the supplementary video.

Our supplementary video also demonstrates a text-to-video application, where we synthesize natural human face video from text input, after converting it to audio through a speech synthesizer [Notevibes 2020]. Finally, our video demonstrates the possibility of interactively editing the pose of our synthesized talking heads by applying a rotation to the intermediate landmarks predicted by our network.

6 CONCLUSION

We have introduced a deep learning based approach to generate speaker-aware talking-head animations from an audio clip and a
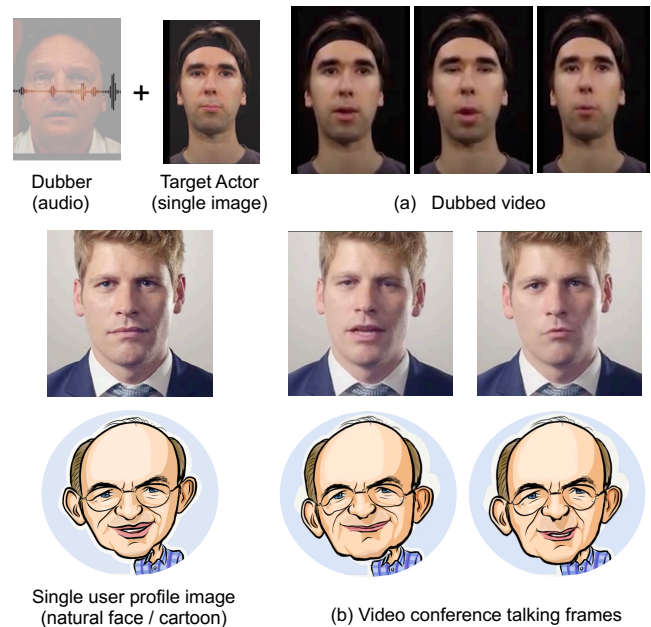

Fig. 13. Applications. Top row: video dubbing for target actor given only audio as input. Middle and bottom row: video conference for natural human and cartoon user profile images. Please also see our supplementary video. Video conference application *natural face ©PxHere* (CC-0).

single image. Our method can handle new audio clips and new portrait images not seen during training. Our key insight was to predict landmarks from disentangled audio content and speaker, such that they capture better lip synchronization, personalized facial expressions and head motion dynamics. This led to much more expressive animations with higher overall quality compared to the state-of-the-art.

Limitations and Future Work. There are still several avenues for future research. Although our method captures aspects of the speaker's style e.g., predicting head swings reflecting active speech, there are several other factors that can influence head motion dynamics. For example, the speaker's mood can also play a significant role in determining head motion and facial expressions. Further incorporating sentiment analysis into the animation pipeline is a promising research direction.

Our speech content animation currently does not always capture well the bilabial and fricative sounds, i.e. /b/m/p/f/v/. We believe this is caused by the voice conversion module that we used, which tends to miss those short phoneme representations when performing voice spectrum reconstruction. While we observe promising results via directly adopting a state-of-the-art voice conversion architecture, a more domain-specific adaptation for audio-driven animation may address such discrepancies between voice conversion and animation.

Improving the image translation from landmarks to videos can also be further investigated. The current image-to-image translation network takes only the 2D facial landmarks as input. Incorporating more phoneme- or viseme-related features as input may

improve the quality of the generated video in terms of articulation. Moreover, background distortion and artifacts are noticeable in our current solution. Our image translation module warps both the background and the foreground head to produce the animation, which gives an impression of a camera motion mixed with head motion. Adapting fore/background separation or portrait matting in the network architecture during training and testing may help generate better results [Sengupta et al. 2020]. Capturing long-range temporal and spatial dependencies between pixels could further reduce low-level artifacts. Another limitation is that in the case of large head motion, more artifacts tend to appear: since we attempt to create animations from a single input image, large rotations/translations require sufficient extrapolation to unseen parts of the head (e.g., neck, shoulders, hair), which are more challenging for our current image translation net to hallucinate especially for natural head images.

Our method heavily relies on the intermediate sparse landmark representation to guide the final video output. The representation has the advantage of being low-dimensional and handling a variety of faces beyond human-looking ones. On the other hand, landmarks serve mostly as coarse proxies for modeling heads, thus, for large motion, they sometimes cause face distortions especially for natural head images (see also our supplementary video, last clip). In the particular case of human face animation, an alternative representation could be denser landmarks or parameters of a morphable model that may result in more accurate face reconstructions. A particular challenge here would be to train such models in the zero-shot learning regime, where the input portrait has not been observed during training; current methods seem to require additional fine-tuning on target faces [Thies et al. 2020].

Finally, our current algorithm focuses on a fully automatic pipeline. It remains an open challenge to incorporate user interaction within a human-in-the-loop approach. An important question is how an animator could edit landmarks in certain frames and propagate those edits to the rest of the video. We look forward to future endeavors on high-quality expressive talking-head animations with intuitive controls.

7 ETHICAL CONSIDERATIONS

"Deepfake videos" are becoming more prevalent in our everyday life. The general public might still think that talking head videos are hard or impossible to generate synthetically. As a result, algorithms for talking head generation can be misused to spread misinformation or for other malicious acts. We hope our code will help people understand that generating such videos is entirely feasible. Our main intention is to spread awareness and demystify this technology. Our code includes a watermark to the generated videos making it clear that they are synthetic.

ACKNOWLEDGMENTS

REFERENCES

Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. Protecting World Leaders Against Deep Fakes. In *Proc. CVPRW*.

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. 2017. Bringing portraits to life. *ACM Trans. Graphics* (2017).

Thabo Beeler and Derek Bradley. 2014. Rigid stabilization of facial expressions. *ACM Trans. Graphics* (2014).

Matthew Brand. 1999. Voice puppetry. In *Proc. SIGGRAPH*.

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proc. ICCV*.

Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip movements generation at a glance. In *Proc. ECCV*.

Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proc. CVPR*.

Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that?. In *Proc. BMVC*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proc. INTERSPEECH*.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *Proc. CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805* (2018).

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graphics* (2016).

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system: a technique for the measurement of facial movement. (1978).

Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. 2018. Generating talking face landmarks from speech. In *Proc. LVA/ICA*.

Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. 2019. Noise-Resilient Training Method for Face Landmark Generation From Speech. *IEEE/ACM Trans. Audio, Speech, and Language Processing* (2019).

Patrick Esser, Ekaterina Sutter, and Björn Ommer. 2018. A variational u-net for conditional appearance and shape generation. In *Proc. CVPR*.

Gary Faigin. 2012. *The artist's complete guide to facial expression*. Watson-Guptill.

Jakub Fišer, Ondřej Jamriška, David Simons, Eli Shechtman, Jingwan Lu, Paul Asente, Michal Lukáč, and Daniel Sýkora. 2017. Example-Based Synthesis of Stylized Facial Animations. *ACM Trans. Graphics* (2017).

Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. 2020a. Foley Music: Learning to Generate Music from Videos. *ECCV* (2020).

Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. 2020b. Music Gesture for Visual Sound Separation. In *Proc. CVPR*. 10478–10487.

Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. 2015. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer graphics forum*.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning Individual Styles of Conversational Gesture. In *Proc. CVPR*.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proc. ICML*.

David Greenwood, Iain Matthews, and Stephen Laycock. 2018. Joint learning of facial expression and head pose from speech. Interspeech.

Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. 2019. FiNet: Compatible and Diverse Fashion Image Inpainting. In *Proc. ICCV*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graphics* (2017).

Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. 2019. Neural style-preserving visual dubbing. *ACM Trans. Graphics* (2019).

Yilong Liu, Feng Xu, Jinxiang Chai, Xin Tong, Lijuan Wang, and Qiang Huo. 2015. Video-audio driven real-time facial animation. *ACM Trans. Graphics* (2015).

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* (2008).

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proc. ICCV*.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* (1976).

Notevibes. 2020. Text to Speech converter. https://notevibes.com/.

Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. 2018. End-to-end learning for 3d facial animation from speech. In *Proc. ICMI*.

Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proc. ECCV*.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. In *Proc. ICML*. 5210–5219.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*.

Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. 2009. Generalized-icp.. In *Proc. Robotics: science and systems*.

Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. Background Matting: The World is Your Green Screen. In *Proc. CVPR*.

Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. 2006. Dancing-to-music character animation. In *Computer Graphics Forum*.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*.

Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. 2019. Talking Face Generation by Conditional Recurrent Adversarial Network. In *Proc. IJCAI*.

Olga Sorkine. 2006. Differential representations for mesh processing. In *Computer Graphics Forum*.

Yannis Stylianou. 2009. Voice transformation: a survey. In *Proc. ICASSP*.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Trans. Graphics* (2017).

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Trans. Graphics* (2017).

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural Voice Puppetry: Audio-driven Facial Reenactment. In *Proc. CVPR, to appear*.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. CVPR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. NeurIPS*.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2016. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. (2016).

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2019. Realistic Speech-Driven Facial Animation with GANs. *IJCV* (2019).

Marilyn A Walker, Janet E Cahn, and Stephen J Whittaker. 1997. Improvising linguistic style: Social and affective bases for agent personality. In *Proc. IAA*.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *Proc. ICASSP*.

Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. 2019. Photo wake-up: 3d character animation from a single photo. In *Proc. CVPR*.

Jordan Yaniv, Yael Newman, and Ariel Shamir. 2019. The face of art: landmark detection and geometric style in portraits. *ACM Trans. Graphics* (2019).

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *Proc. ICCV*.

Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proc. AAAI*.

Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. 2018. Visemenet: Audio-driven animator-centric speech animation. *ACM Trans. Graphics* (2018).

A SPEAKER-AWARE ANIMATION NETWORK

The attention network follows the encoder block structure in Vaswani et al. [2017]. It is composed of a stack of $N = 2$ identical layers. Each layer has (a) a multi-head self-attention mechanism with $N_{head} = 2$ heads and dimensionality $d_{model} = 32$ and (b) a fully connected

Table 4. Generator architecture for synthesizing natural face images.

	Landmark Representation Y_t
256×256	Input Image Q
128×128	ResBlock down $(3 + 3) \rightarrow 64$
64×64	ResBlock down $64 \rightarrow 128$
32×32	ResBlock down $128 \rightarrow 256$
16×16	ResBlock down $256 \rightarrow 512$
8×8	ResBlock down $512 \rightarrow 512$
4×4	ResBlock down $512 \rightarrow 512$
8×8	ResBlock up $512 \rightarrow 512$
16×16	Skip + ResBlock up $(512 + 512) \rightarrow 512$
32×32	Skip + ResBlock up $(512 + 512) \rightarrow 256$
64×64	Skip + ResBlock up $(256 + 256) \rightarrow 128$
128×128	Skip + ResBlock up $(128 + 128) \rightarrow 64$
256×256	Skip + ResBlock up $(64 + 64) \rightarrow 3$
256×256	Tanh

feed-forward layer whose size is the same to the input size. We also use the embedding layer (a one-layer MLP with hidden size 32) and the position encoder layer as mentioned in Vaswani et al. [2017]. The discriminator network has a similar network architecture to the attention network. The difference is that the discriminator also includes a three-layer MLP network which has 512, 256, 1 hidden size respectively.

B IMAGE-TO-IMAGE TRANSLATION NETWORK

The layers of the network architecture used to generate natural human face images are listed in Table 4. In this table, the left column indicates the spatial resolution of the feature map output. **ResBlock down** means a 2-strided convolutional layer with 3×3 kernel followed by two residual blocks, **ResBlock up** means a nearest-neighbor upsampling with a scale of 2, followed by a 3×3 convolutional layer and then two residual blocks, and **Skip** means a skip connection concatenating the feature maps of an encoding layer and decoding layer with the same spatial resolution.