MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement

Alexander Richard[1] Michael Zollhöfer[1] Yandong Wen[2] Fernando de la Torre[2] Yaser Sheikh[1]
[1]Facebook Reality Labs [2]Carnegie Mellon University
{richardalex, zollhoefer, yasers}@fb.com yandongw@andrew.cmu.edu torre@cs.cmu.edu

Abstract

This paper presents a generic method for generating full facial 3D animation from speech. Existing approaches to audio-driven facial animation exhibit uncanny or static upper face animation, fail to produce accurate and plausible co-articulation or rely on person-specific models that limit their scalability. To improve upon existing models, we propose a generic audio-driven facial animation approach that achieves highly realistic motion synthesis results for the entire face. At the core of our approach is a categorical latent space for facial animation that disentangles audio-correlated and audio-uncorrelated information based on a novel cross-modality loss. Our approach ensures highly accurate lip motion, while also synthesizing plausible animation of the parts of the face that are uncorrelated to the audio signal, such as eye blinks and eye brow motion. We demonstrate that our approach outperforms several baselines and obtains state-of-the-art quality both qualitatively and quantitatively. A perceptual user study demonstrates that our approach is deemed more realistic than the current state-of-the-art in over 75% of cases. We recommend watching the supplemental video before reading the paper.

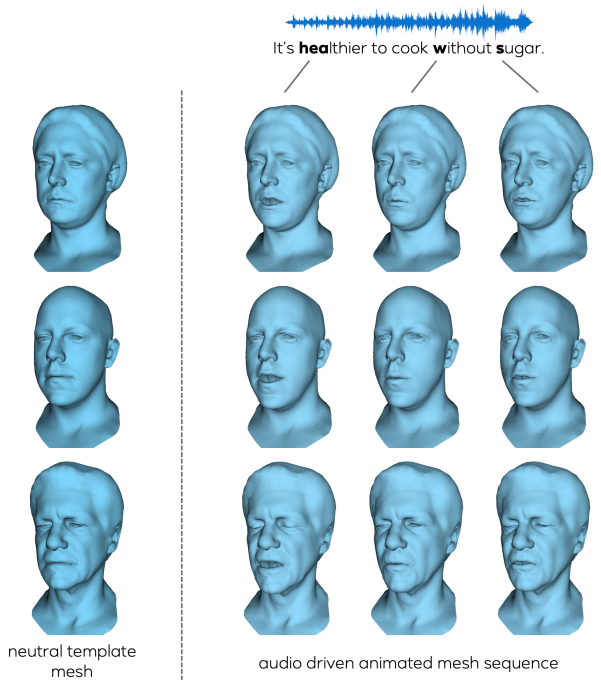

Figure 1. Given a neutral face mesh of a person and a speech signal as input, our approach generates highly realistic face animations with accurate lip shape and realistic upper face motion such as eye blinks and eyebrow raises.

1. Introduction

Speech-driven facial animation is a highly challenging research problem with several applications such as facial animation for computer games, e-commerce, or immersive VR telepresence. The demands on speech-driven facial animation differ depending on the application. Applications such as speech therapy or entertainment (e.g., Animojies or AR effects) do not require a very presice level of realism in the animation. In the production of films, movie dubbing, driven virtual avatars for e-commerce applications or immersive telepresence, on the contrary, the quality of speech animation requires a high degree of naturalness, plausibility, and has to provide intelligibility comparable to a natural speaker. The human visual system has been evolutionary adapted to understanding subtle facial motions and expressions. Thus, a poorly animated face without realistic co-articulation effects or out of lip-sync is deemed to be disturbing for the user.

Psychological literature has observed that there is an important degree of dependency between speech and facial gestures. This dependency has been exploited by audio-driven facial animation methods developed in computer vision and graphics [4, 2]. With the advances in deep learning, recent audio-driven face animation techniques make use of person-specific approaches [22, 26] that are trained in a supervised fashion based on a large corpus of paired audio and mesh data. These approaches are able to obtain high-quality lip animation and synthesize plausible upper face

motion from audio alone. To obtain the required training data, high-quality vision-based motion capture of the user is required, which renders these approaches as highly impractical for consumer-facing applications in real world settings. Recently, Cudeiro *et al*. [8] extended this work, by proposing a method that is able to generalize across different identities and is thus able to animate arbitrary users based on a given audio stream and a static neutral 3D scan of the user. While such approaches are more practical in real world settings, they normally exhibit uncanny or static upper face animation [8]. The reason for this is that audio does not encode all aspects of the facial expressions, thus the audio-driven facial animation problem tries to learn a one-to-many mapping, *i.e*., there are multiple plausible outputs for every input. This often leads to over-smoothed results, especially in the regions of the face that are only weakly or even uncorrelated to the audio signal.

This paper proposes a novel audio-driven facial animation approach that enables highly realistic motion synthesis for the entire face and also generalizes to unseen identities. To this end, we learn a novel categorical latent space of facial animation that disentangles audio-correlated and audio-uncorrelated information, *e.g*., eye closure should not be bound to a specific lip shape. The latent space is trained based on a novel cross-modality loss that encourages the model to have an accurate upper face reconstruction independent of the audio input and accurate mouth area that only depends on the provided audio input. This disentangles the motion of the lower and upper face region and prevents over-smoothed results. Motion synthesis is based on an autoregressive sampling strategy of the audio-conditioned temporal model over the learnt categorical latent space. Our approach ensures highly accurate lip motion, while also being able to sample plausible animations of parts of the face that are uncorrelated to the audio signal, such as eye blinks and eye brow motion. In summary, our contributions are:

- A novel categorical latent space for facial animation synthesis that enables highly realistic animation of the whole face by disentanglement of the upper and lower face region based on a cross-modality loss.
- An autoregressive sampling strategy for motion synthesis from an audio-conditioned temporal model over the learned categorical latent space.
- Our approach outperforms the current state-of-the-art both qualitatively and quantitatively in terms of obtained realism.

2. Related Work

Speech-based face animation has a long history in computer vision and ranges from artist-friendly stylized and viseme-based models [11, 42, 19] to neural synthesis of 2D [29, 40, 31] and 3D [30, 22, 26] faces. In the following, we review the most relevant approaches.

Viseme-based face animation. In early approaches, a viseme sequence is generated from input text [13, 14] or directly from speech using HMM-based acoustic models [35]. Visual synthesis is achieved by blending between face images from a database. For 3D face animation, Kalberer *et al*. [20] model viseme deformations via independent component analysis. In [21], 3D face masks are projected onto a low-dimensional eigenspace and smooth temporal animation is achieved by spline fitting on each component of this space, given a sequence of key viseme masks. Leaning on the concept on phonetic co-articulation, Martino *et al*. [9] seek to find context-dependent visemes rather than blending between key templates. Given the success of JALI [11], an animator-centric audio-drivable jaw and lip model, Zhou *et al*. [42] propose an LSTM-based, near real-time approach to drive an appealing lower face lip model. Due to their generic nature and artist-friendly design, viseme based approaches are popular for commercial applications particularly in virtual reality [1, 19].

Speech-driven 2D talking heads. Many speech-driven approaches are aimed at generating realistic 2D video of talking heads. Early work [4] replaced the problem of learning by searching in existing video for similar utterances as the new speech. Brand *et al*. [3] proposed a generic ML model to drive a facial control model that incorporates vocal and facial dynamic effects such as co-articulation. The approach of Suwajanakorn *et al*. [29] is able to generate video of a single person with accurate lip sync by synthesizing matching mouth textures and compositing them on top of a target video clip. However, this approach only synthesizes the mouth region and requires a large corpus of personalized training data (≈17 hours). Wav2lip [25] tackles the problem of visual dubbing, *i.e*., of lip-syncing a talking head video of an arbitrary person to match a target speech segment. Neural Voice Puppetry [31] performs audio-driven facial video synthesis via neural rendering to generate photo-realistic output frames. X2Face [40] is an encoder/decoder approach for 2D face animation, *e.g*., from audio, that can be trained fully self-supervised using a large collection of videos. Other talking face video techniques [28, 7, 36] are based on generative adversarial networks (GANs). The approach of Vougioukas *et al*. [36] synthesizes new upper face motion, but the results are of low resolution and look uncanny. The lower face animation approach of Chung *et al*. [7] requires only a few still images of the target actor and a speech snippet as input. To achieve this, an encoder-decoder model is employed that discovers a joint embedding of the face and audio. Zhou *et al*. [41] improve on Chung *et al*. [7] by learning an audio-visual latent space in combination with an adversarial loss that allows to synthesize 2D talking heads from either video or audio. All the described 2D approaches operate in pixel space and can

not be easily generalized to 3D.

Speech-driven 3D models. Approaches to drive 3D face models mostly use visual input. While earlier works map from motion captures or 2D video to 3D blendshape models [10, 37, 15], more recent works provide solutions to animate photo-realistic 3D avatars using sensors on a VR headset [23, 26, 39, 6]. These approaches achieve highly realistic results, but they are typically personalized and are not audio-driven. Most fully speech-driven 3D face animation techniques require either personalized models [5, 22, 26] or map to lower fidelity blendshape models [24] or facial landmarks [12, 16]. Cao *et al*. [5] propose speech-driven animation of a realistic textured personalized 3D face model that requires mocap data from the person to be animated, offline processing and blending of motion snippets. The fully speech-driven approach of Richard *et al*. [26] enables real-time photo-realistic avatars, but is personalized and relies on hours of training data from a single subject. Karras *et al*. [22] learn a speech-driven 3D face mesh from as little as 3-5 minutes of data per subject and condition their model on emotion states that lead to facial expressions. In contrast to our approach, however, this model has lower fidelity lip sync and upper face expressions, and does not generalize to new subjects. In [30], a single-speaker model is generalized via re-targeting techniques to arbitrary stylized avatars. Most closely related to our approach is VOCA [8], which allows to animate arbitrary neutral face meshes from audio and achieves convincing lip synchronization. While generating appealing lip motion, their model can not synthesize upper face motion and tends to generate muted expressions. Moreover, the approach expects a training identity as conditional input to the model. As shown by the authors in their supplemental video, this identity code has a high impact on the quality of the generated lip synchronization. Consequentially, we found VOCA to struggle on large scale datasets with hundreds of training subjects. In contrast to the discussed works, our approach is non-personalized, generates realistic upper face motion, and leads to highly accurate lip synchronization.

3. Method

Our goal is to animate an arbitrary neutral face mesh using only speech. Since speech does not encode all aspects of the facial expressions – eye-blinks are a simple example of uncorrelated expressive information – most existing audio-driven approaches exhibit uncanny or static upper face animation [8]. To overcome this issue, we learn a categorical latent space for facial expressions. At inference time, an autoregressive sampling from a speech-conditioned temporal model over this latent space ensures accurate lip motion while synthesizing plausible animation of face parts that are uncorrelated to speech. With this in mind, the latent space

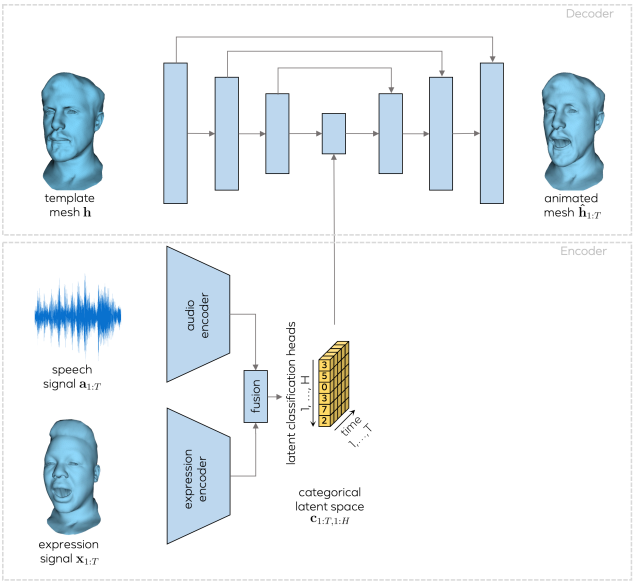

Figure 2. System overview. A sequence of animated face meshes (the expression signal) and a speech signal are mapped to a categorical latent expression space. A UNet-style decoder is then used to animate a given neutral-face template mesh according to the encoded expressions.

should have the following properties:
Categorical. Most successful temporal models operate on categorical spaces [33, 32, 34]. In order to use such models, the latent expression space should be categorical as well.
Expressive. The latent space must be capable of encoding diverse facial expressions, including sparse events like eye blinks.
Semantically disentangled. Speech-correlated and speech-uncorrelated information should be at least partially disentangled, *e.g*., eye closure should not be bound to a specific lip shape.

3.1. Modeling and Learning the Expression Space

Let $\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T), \boldsymbol{x}_t \in \mathbb{R}^{V\times 3}$ be a sequence of T face meshes, each represented by V vertices. Let further $\boldsymbol{a}_{1:T} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T), \boldsymbol{a}_t \in \mathbb{R}^D$ be a sequence of T speech snippets, each with D samples, aligned to the corresponding (visual) frame t. Moreover, we denote the template mesh that is required as input by $\boldsymbol{h} \in \mathbb{R}^{V\times 3}$.

In order to achieve high expressiveness of the categorical latent space, the space must be sufficiently large. Since this leads to an infeasibly large number of categories C for a single latent categorical layer, we model H latent classification heads of C-way categoricals, allowing for a large expression space with a comparably small number of categories as the number of configurations of the latent space is C^H and therefore grows exponentially in H. Throughout this paper, we use $C = 128$ and $H = 64$.

The mapping from expression and audio input signals to

the multi-head categorical latent space is realized by an encoder $\tilde{\mathcal{E}}$ which maps from the space of all audio sequences and all expression sequences (*i.e.*, sequences of animated face meshes) to a $T \times H \times C$-dimensional encoding

$$\mathbf{enc}_{1:T,1:H,1:C} = \tilde{\mathcal{E}}(\boldsymbol{x}_{1:T}, \boldsymbol{a}_{1:T}) \in \mathbb{R}^{T \times H \times C}. \quad (1)$$

This continuous-valued encoding is then transformed into a categorical representation using a Gumbel-softmax [18] over each latent classification head,

$$\boldsymbol{c}_{1:T,1:H} = \Big[\text{Gumbel}\big(\mathbf{enc}_{t,h,1:C}\big)\Big]_{1:T,1:H} \quad (2)$$

such that each categorical component at time step t and in the latent classification head h gets assigned one of C categorical labels, $c_{t,h} \in \{1, \ldots, C\}$. We denote the complete encoding function, *i.e.*, $\tilde{\mathcal{E}}$ followed by categorization, as $\mathcal{E}$.

The animation of the input template mesh $\boldsymbol{h}$ is realized by a decoder $\mathcal{D}$,

$$\hat{\boldsymbol{h}}_{1:T} = \mathcal{D}(\boldsymbol{h}, \boldsymbol{c}_{1:T,1:H}), \quad (3)$$

which maps the encoded expression onto the provided template $\boldsymbol{h}$. It thereby generates an animated sequence $\hat{\boldsymbol{h}}_{1:T}$ of face meshes that looks like the person represented by template $\boldsymbol{h}$, but moves according to the expression code $\boldsymbol{c}_{1:T,1:H}$. See Figure 2 for an overview.

Learning the Latent Space. At training time, ground-truth correspondences are only available for the case where *(a)* template mesh, speech signal, and expression signal are from the same identity, and *(b)* the desired decoder output $\hat{\boldsymbol{h}}_{1:T}$ is equal to the expression input $\boldsymbol{x}_{1:T}$. Consequentially, training with a simple ℓ_2 reconstruction loss between $\hat{\boldsymbol{h}}_{1:T}$ and $\boldsymbol{x}_{1:T}$ would lead to the audio input being completely ignored as the expression signal already contains all information necessary for perfect reconstruction – a problem that leads to poor speech-to-lip synchronization, as we show in Section 4.1.

We therefore propose a cross-modality loss that ensures information from both input modalities is utilized in the latent space. Let $\boldsymbol{x}_{1:T}$ and $\boldsymbol{a}_{1:T}$ be a given expression and speech sequence. Let further $\boldsymbol{h}_x$ denote the template mesh for the person represented in the signal $\boldsymbol{x}_{1:T}$. Instead of a single reconstruction $\hat{\boldsymbol{h}}_{1:T}$, we generate two different reconstructions

$$\hat{\boldsymbol{h}}^{(\text{audio})}_{1:T} = \mathcal{D}\big(\boldsymbol{h}_x, \mathcal{E}(\tilde{\boldsymbol{x}}_{1:T}, \boldsymbol{a}_{1:T})\big) \quad \text{and} \quad (4)$$

$$\hat{\boldsymbol{h}}^{(\text{expr})}_{1:T} = \mathcal{D}\big(\boldsymbol{h}_x, \mathcal{E}(\boldsymbol{x}_{1:T}, \tilde{\boldsymbol{a}}_{1:T})\big), \quad (5)$$

where $\tilde{\boldsymbol{x}}_{1:T}$ and $\tilde{\boldsymbol{a}}_{1:T}$ are a randomly sampled expression and audio sequence from the training set. In other words, $\hat{\boldsymbol{h}}^{(\text{audio})}_{1:T}$ is a reconstruction given the correct audio but a random expression sequence and $\hat{\boldsymbol{h}}^{(\text{expr})}_{1:T}$ is a reconstruction given the correct expression sequence but random audio. Our novel cross-modality loss is then defined as

$$\mathcal{L}_{\text{xMod}} = \sum_{t=1}^{T}\sum_{v=1}^{V} \mathcal{M}^{(\text{upper})}_v \big(\|\hat{h}^{(\text{expr})}_{t,v} - x_{t,v}\|^2\big) + \sum_{t=1}^{T}\sum_{v=1}^{V} \mathcal{M}^{(\text{mouth})}_v \big(\|\hat{h}^{(\text{audio})}_{t,v} - x_{t,v}\|^2\big), \quad (6)$$

where $\mathcal{M}^{(\text{upper})}$ is a mask that assigns a high weight to vertices on the upper face and a low weight to vertices around the mouth. Similarly, $\mathcal{M}^{(\text{mouth})}$ assigns a high weight to vertices around the mouth and a low weight to other vertices.

The cross-modality loss encourages the model to have an accurate upper face reconstruction independent of the audio input and, accordingly, to have an accurate reconstruction of the mouth area based on audio independent of the expression sequence that is provided. Since eye blinks are quick and sparse events that affect only a few vertices, we also found it crucial to emphasize the loss on the eye lid vertices during training. We therefore add a specific eye lid loss

$$\mathcal{L}_{\text{eyelid}} = \sum_{t=1}^{T}\sum_{v=1}^{V} \mathcal{M}^{(\text{eyelid})}_v \big(\|\hat{h}_{t,v} - x_{t,v}\|^2\big), \quad (7)$$

where $\mathcal{M}^{(\text{eyelid})}$ is a binary mask with ones for eye lid vertices and zeros for all other vertices. The final loss we optimize is $\mathcal{L} = \mathcal{L}_{xMod} + \mathcal{L}_{\text{eyelid}}$. We found that an equal weighting of the two terms works well in practice.

Network Architectures. The audio encoder is a four-layer 1D temporal convolutional network similar to the one used in [26]. The expression encoder has three fully connected layers followed by a single LSTM layer to capture temporal dependencies. The fusion module is a three-layer MLP. The decoder $\mathcal{D}$ has a UNet-style architecture with additive skip connections, see Figure 2. This architectural inductive bias prevents the network from diverging from the given template mesh too much. In the bottleneck layer, the expression code $\boldsymbol{c}_{1:T,1:H}$ is concatenated with the encoded template mesh. The bottleneck layer is followed by two LSTM layers to model temporal dependencies between frames followed by three fully connected layers remapping the representation to vertex space.

See the supplementary material for more details.

3.2. Audio-Conditioned Autoregressive Modeling

When driving a template mesh using audio input alone, the expression input $\boldsymbol{x}_{1:T}$ is not available. With only one modality given, missing information that can not be inferred from audio has to be synthesized. Therefore, we learn an autoregressive temporal model over the categorical latent space. This model allows to sample a latent sequence that generates plausible expressions and is consistent with the

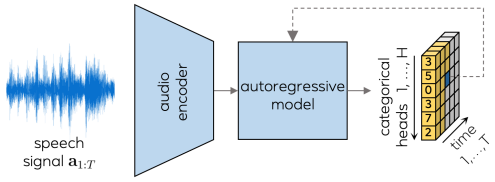

Figure 3. Autoregressive model. Audio-conditioned latent codes are sampled for each position $c_{t,h}$ in the latent expression space, where the model only has access to previously generated labels as defined in Equation (8).

audio input. Following Bayes' Rule, the probability of a latent embedding $\boldsymbol{c}_{1:T,1:H}$ given the audio input $\boldsymbol{a}_{1:T}$ can be decomposed as

$$p(\boldsymbol{c}_{1:T,1:H}|\boldsymbol{a}_{1:T}) = \prod_{t=1}^{T}\prod_{h=1}^{H} p(c_{t,h}|\boldsymbol{c}_{<t,1:H}, \boldsymbol{c}_{t,<h}, \boldsymbol{a}_{\leq t}). \tag{8}$$

Note that we assumed temporal causality in the decomposition, *i.e.*, a category $c_{t,h}$ at time t only depends on current and past audio information $\boldsymbol{a}_{\leq t}$ rather than on past and future context $\boldsymbol{a}_{1:T}$. We model this quantity with an autoregressive convolutional network similar to PixelCNN [33]. Our autoregressive temporal CNN has four convolutional layers with increasing dilation along the temporal axis. The convolutions are masked such that for the prediction of $c_{t,h}$ the model only has access to information from all categorical heads in the past, $\boldsymbol{c}_{<t,1:H}$, and the preceding categorical heads at the current time step, $\boldsymbol{c}_{t,<h}$, see yellow blocks in Figure 3. To train the autoregressive model, we generate training data using the pretrained encoder $\mathcal{E}$, which maps the expression and audio sequences $(\boldsymbol{x}_{1:T}, \boldsymbol{a}_{1:T})$ in the training set to their categorical embeddings, see Equation (1). The autoregressive model is then optimized on these correspondences using teacher forcing and a cross-entropy loss over the latent categorical labels. At inference time, a categorical expression code is sequentially sampled for each position $c_{t,h}$ using the trained audio-conditioned autoregressive network.

4. Evaluation

Dataset. Existing works are typically trained on less than a dozen subjects [22, 30, 26] and available datasets are tracked with low fidelity [8], *e.g.*, not including eye lids, facial hair, or eyebrows, and therefore render unfit to demonstrate high fidelity full-face motion from speech generalizing over arbitrary identities.

In this work, we use an in-house dataset of 250 subjects, each of which is reading a total of 50 phonetically balanced sentences. The sequences are captured at 30 frames per second and face meshes are tracked from 80 synchronized cameras surrounding the subject's head. We use a face model with 6,172 vertices that has a high level of detail including eye lids, upper face structure, and different hair styles. In total, our data amounts to 13h of paired audio-visual data, or 1.4 million frames of tracked 3D face meshes. We train our model on the first 40 sentences of 200 subjects and use the remaining 10 sentences of the remaining 50 subjects as validation (10 subjects) and test set (40 subjects). All shown qualitative and quantitative results are obtained using the held-out subjects and sentences in the test set.

encoder inputs	decoder loss	reconstruction error (in mm)	autoregr. model perplexity
expression	ℓ_2	1.156	1.853
expr. + audio	ℓ_2	1.124	1.879
expr. + audio	$\mathcal{L}_{\text{xMod}}$	1.244	**1.669**

Table 1. Different strategies to learn the latent space and their impact on the autoregressive model.

Speech Features. Our audio data is recorded at 16kHz. For each tracked mesh, we compute the Mel spectrogram of a 600ms audio snippet starting 500ms before and ending 100ms after the respective visual frame. We extract 80-dimensional Mel spectral features every 10ms, using 1,024 frequency bins and a window size of 800 for the underlying Fourier transform.

4.1. Disentangling Audio and Expression

In this section, we show that our model successfully learns a latent representation that allows to control upper face motion and lip synchronization from different input modalities. Specifically, we aim to answer these questions:
(1) In Figure 2, the encoder maps a sequence of audio features and face meshes to the latent space. At training time, the expression input $\boldsymbol{x}_{1:T}$ is exactly the target signal that would minimize the loss at the output of the decoder. *Why is it crucial to have audio as input to learn the latent space?*
(2) Many multi-modal approaches suffer from the problem that the weaker modality is ignored [27, 38]. *How can this effect be avoided, considering that the expression input is already signal-complete?*
(3) Given the issues discussed above, it is interesting to investigate the structure of the latent space. *Are latent space and animated mesh regions semantically disentangled across modalities?*

(1) Multi-modal encoder inputs. We start with a discussion on learning a latent space from audio and expression inputs. One straightforward way to learn a latent space is to *omit the audio input* to the encoder in Figure 2. Limited capacity of the latent space and the inductive bias of the decoder, specifically its skip connections, make sure that even in this case, sufficient information is used from the tem-

plate geometry (see supplemental video for an example). Not surprisingly, this setup also leads to a low reconstruction error[1] as shown in Table 1. The major caveat of this strategy, however, is the structure of the latent space. Since there is no incentive to disentangle information about independent parts of the face, latent representations can have a strong entanglement between eye motion and mouth shape. As a consequence, the autoregressive model (Equation (8)) fails to generate latent representations of accurate lip shape if it is required to produce temporally consistent and plausible upper face motion at the same time. Consequentially, lip motion is less accurate and facial motion more muted when compared to our model that uses audio and expression to learn the latent space.

We quantify this effect by evaluating the perplexity of the autoregressive model. Given a categorical latent representation $\boldsymbol{c}_{1:T,1:H}$ of test set data, the perplexity is

$$PP = p(\boldsymbol{c}_{1:T,1:H}|\boldsymbol{a}_{1:T})^{-\frac{1}{T \cdot H}}, \tag{9}$$

i.e., the inverse geometric average of the likelihood of the latent representations under the autoregressive model. Intuitively, a low perplexity means that at each prediction step the autoregressive model only has a small number of potential categories to choose from, whereas high perplexity means the model is less certain which categorical representation to choose next. A perplexity of 1 would mean the autoregressive model is fully deterministic, *i.e.*, the latent embedding is fully defined by the conditioning audio input. As there are face motions uncorrelated with audio, this does not happen in practice. A comparison of row one and three in Table 1 shows that learning the latent space from audio and expression input leads to a stronger and more confident audio-conditioned autoregressive model than learning the latent space from expression inputs alone. This observation is consistent with the qualitative evaluation in the supplemental video.

(2) Enforcing the usage of audio input. The training loss of the decoder has major impact on how the model makes use of the different input modalities. Since the expression input is sufficient for exact reconstruction, a simple ℓ_2 loss on the desired output meshes will cause the model to ignore the audio input completely and the results are similar to the above case where no audio was given as encoder input, see Table 1 row one and two. The cross-modality loss $\mathcal{L}_{\text{xMod}}$ offers an effective solution to this issue. The model is encouraged to learn accurate lip shape even if the expression input is exchanged by different, random expressions. Similarly, upper face motion is encouraged to remain accurate independent of the audio input. The last row in Table 1 shows that the cross-modality loss does not negatively affect *expressiveness* of the learnt latent space (*i.e.*, the reconstruction error is small for all latent space variants) but positively affects the autoregressive model's perplexity.

[1]We refer to the reconstruction error of a model as its ℓ_2 error on the test set when all inputs, *i.e.*, template mesh and encoder input, are from the same person. The autoregressive model is not used for reconstruction.

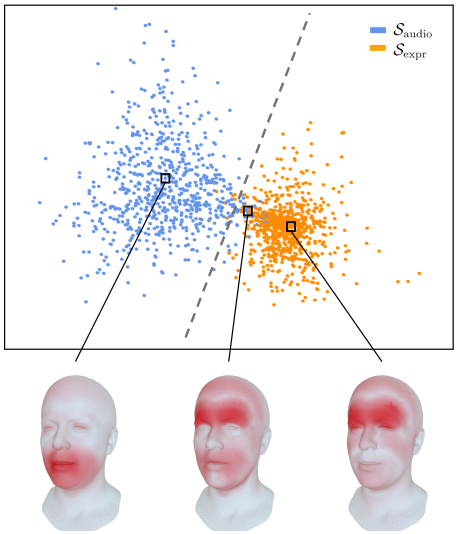

Figure 4. Visualization of the latent space. Latent configurations caused by changes in the audio input are clustered together. Latent configurations caused by changes in the expression input form another cluster. Both clusters can be well separated with minimal leakage into each other.

(3) Cross-Modality Disentanglement. We demonstrate that the cross-modal disentanglement with its properties analyzed above in fact leads to a structured latent space and that each input modality has different effects on the decoded face meshes. To this end, we generate two different sets of latent representations, $\mathcal{S}_{\text{audio}}$ and $\mathcal{S}_{\text{expr}}$. $\mathcal{S}_{\text{audio}}$ contains latent codes obtained by fixing the expression input to the encoder and varying the audio input. Similarly, $\mathcal{S}_{\text{expr}}$ contains latent codes obtained by fixing the audio input and varying the expression input. In the extreme case of perfect cross-modal disentanglement, $\mathcal{S}_{\text{audio}}$ and $\mathcal{S}_{\text{expr}}$ form two non-overlapping clusters. We fit a separating hyper-plane on the points in $\mathcal{S}_{\text{audio}} \cup \mathcal{S}_{\text{expr}}$ and visualize a 2D projection of the result in Figure 4. Note that there is only minimal leakage between the clusters formed by $\mathcal{S}_{\text{audio}}$ and $\mathcal{S}_{\text{expr}}$. Below the plot, we visualize which face vertices are most moved by latent representations within the cluster of $\mathcal{S}_{\text{audio}}$ (left), within the cluster of $\mathcal{S}_{\text{expr}}$ (right), and close to the decision boundary (middle). While audio mostly controls the mouth area and expression controls the upper face, latent representations close to the decision boundaries influence face vertices in all areas, which suggests that certain upper face expressions are correlated to speech, e.g., raising the eyebrows.

In Figure 5, we show the vertices that are most affected by decoding of all latent representations in $\mathcal{S}_{\text{audio}}$ and $\mathcal{S}_{\text{expr}}$,

(a) Vertices most influenced by audio input.

(b) Vertices most influenced by expression input.

Figure 5. Impact of the audio and expression modalities on the generated face meshes. Audio steers primarily the mouth area but has also a visible impact on eyebrow motion. Expression meshes influence primarily the upper face parts including the eye lids.

	lip vertex error (in mm)
VOCA [8]	3.720
VOCA [8] + our audio encoder	3.472
Ours	**3.184**

Table 2. Lip errors of our approach compared to VOCA.

respectively. It becomes evident that the loss $\mathcal{L}_{\text{xMod}}$ leads to a clear cross-modality disentanglement into upper and lower face motion. Yet, it is notable that audio, besides its impact on lips and jaw, has a considerable impact on the eyebrow area (Figure 5a). We show in the supplemental video that our model, in fact, learns correlations between speech and eyebrow motion, *e.g.*, raising the eyebrows when words are emphasized.

4.2. Audio-Driven Evaluation

Lip-sync Evaluation. We compare our approach to VOCA [8], the state of the art in audio-driven animation of arbitrary template face meshes. To evaluate the quality of the generated lip synchronization, we define the lip error of a single frame to be the maximal ℓ_2 error of all lip vertices and report the average over all frames in the test set. Since upper lip and mouth corners move much less than the lower lip, we found that the average over all lip vertex errors tends to mask inaccurate lip shapes, while the maximal lip vertex error per frame correlates better with the perceptual quality.

For the comparison with VOCA, which expects a conditioning on a training identity, we sample said identity at random. In addition to the original implementation of VOCA, we also compare to a variant where we replace the DeepSpeech features [17] used in [8] with Mel spectrograms and our audio encoder. Table 2 shows that our approach achieves a lower lip error per frame on average. As shown in the supplemental material of [8], the quality of VOCA is strongly dependent on the chosen conditioning identity. On a challenging dataset with highly detailed meshes, we find this effect to be amplified even more and observed VOCA to produce much less pronounced lip motion, particularly missing most lip closures. We provide a side-by-side comparison in the supplemental video.

	favorability			ours better or equal
	competitor	equal	ours	
ours vs. VOCA [8]				
full-face	24.7%	20.9%	54.4%	**75.3%**
lip sync	23.0%	19.8%	57.2%	**77.0%**
upper face	33.6%	21.6%	44.8%	**66.4%**
ours vs. ground truth				
full-face	42.1%	35.7%	22.2%	**57.9%**
lip sync	45.1%	34.1%	20.8%	**54.9%**
upper face	68.5%	6.9%	24.6%	**31.5%**

Table 3. Perceptual study. Human participants where asked which of two presented clips are more realistic as full-face clips, upper face only, or in terms of lip sync. For each row, 400 pairs of side-by-side clips have been ranked by favorability.

Perceptual Evaluation. We presented side-by-side clips of our approach versus either VOCA or tracked ground truth to a total of 100 participants and let them judge three subtasks: a full face comparison, a lip sync comparison, where we showed only the region between the chin and the nose, and an upper face comparison, where we showed only the face from the nose upwards, see Table 3. For each row, 400 pairs of short clips each containing one sentence spoken by a subject from the test set have been evaluated. Participants could choose to either favor one clip over the other or rank them both as equally good.

In a comparison against VOCA, our approach has been ranked better or equal in more than 75% of the cases for full face animation and lip sync and is ranked better or equal than VOCA in 77% of the cases. For upper face motion alone, the results are slightly lower but participants still heavily favored our approach (66.4% of cases better than or equal to VOCA). Note that most frequently our approach has been strictly preferred over VOCA; equal favorability only makes up a minority of cases. When comparing our approach to tracked ground truth, most participants found ground truth favorable. Yet, with 35.7% of full-face examples being ranked equally good and only 42.1% ranked less favorable than ground truth, our approach proves to be a strong competitor.

Qualitative Examples. Figure 1 shows audio-driven examples generated with our approach. The lip shapes are consistent with the respective speech parts among all subjects, while unique and diverse upper face motion such as eye brow raises and eye blinks are generated separately for each sequence. Further examples generated with our approach and comparisons to other models can be found in the supplemental video.

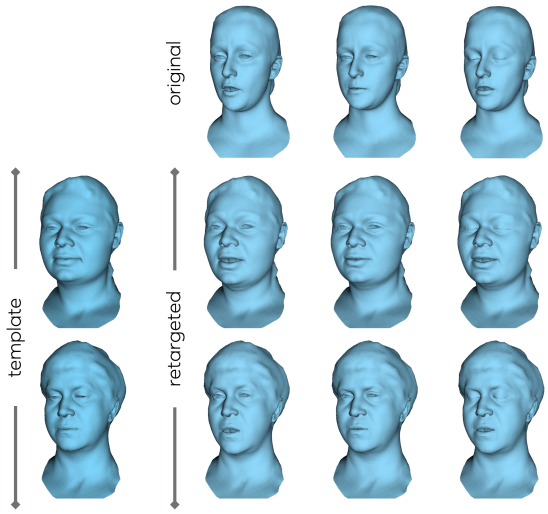

Figure 6. Re-targeting. Given an animated mesh and neutral templates of other identities, our approach accurately re-targets facial expressions such as lip shape, eye closure, and eyebrow raises.

4.3. Re-Targeting

Re-targeting is the process of mapping facial motion from one identity's face onto another identity's face. Typical applications are movies or computer games, where an actor animates a face that is not his own. Our system can successfully address this problem, as we demonstrate in the supplemental video. Using the architecture displayed in Figure 2, the latent expression embedding is computed from the original (tracked) face mesh and the audio of the source identity. The template mesh is of the desired target identity. Our system maps the audio and original animated face mesh to a latent code and decodes it to an animated version of the template mesh. Note that no autoregressive modeling is required for this task. Besides the video, Figure 6 shows examples of re-targeted facial expressions from one identity to another.

4.4. Mesh Dubbing

Usually, when dubbing videos, a speech translation that is fully consistent with the lip motion in the original language is not possible. To overcome this, our system can be used to re-synthesize 3D face meshes with matching lip motion in another language while keeping upper face motion intact. In this task, the original animated mesh and a new audio snippet are provided as input. As before, we use the architecture in Figure 2 directly to re-synthesize lip motion in the new language. Since the latent space is disentangled across modalities (see Figure 4 and Figure 5), lip motion will be adapted to the new audio snippet but the general upper face motion such as eye blinks are maintained from the original clip. Figure 7 shows an example of this application. See the supplemental video for further examples.

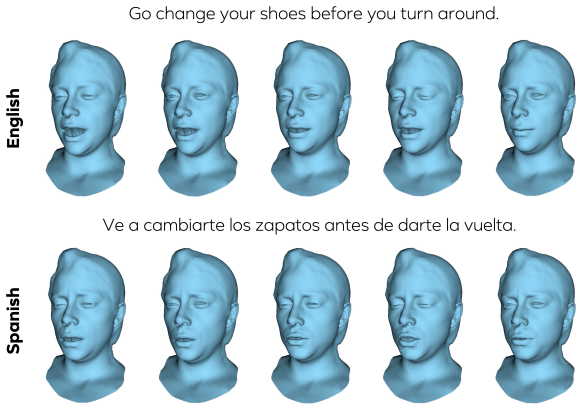

Figure 7. Dubbing. We re-synthesize an English sentence with a new Spanish audio snippet. Note how the lip shape is adjusted to the new audio but general upper face motion like eye closures are maintained.

5. Limitations

While we have demonstrated state-of-the-art performance for audio driven facial animation both in terms of lip-sync as well as naturalness of the generated motion of the entire face, our approach is subject to a few limitations that can be addressed in follow-up work: (1) Our approach relies on audio inputs that extend 100ms beyond the respective visual frame. This leads to an inherent latency of 100ms and prevents the use of our approach for online applications. Please note, this 'look ahead' is beneficial to achieve highest quality lip-sync, *e.g.*, for sounds like '$/p/$' the lip closure can be modeled better. (2) Besides latency, our approach can not be run in real-time on low-cost commodity hardware such as a laptop CPU or virtual reality devices. We believe that the computational cost can be drastically improved by further research. (3) If the face tracker fails to track certain parts of the face, *e.g.*, if hair overlaps and occludes the eyebrows or eyes, we can not correctly learn the correlation of their motion to the audio signal.

6. Conclusion

We have presented a generic method for generating 3D facial animation from audio input alone. A novel categorical latent space in combination with a cross-modality loss enables autoregressive generation of highly realistic animation. Our approach has demonstrated highly accurate lip motion, while also synthesizing plausible motion of uncorrelated regions of the face. It outperforms several baselines and obtains state-of-the-art quality. We hope that our approach will be a stepping stone towards VR telepresence applications, as head mounted capture devices become smaller, thus complicating the regression of accurate lip and tongue motion from oblique camera viewpoints.

References

[1] Amazon sumerian, 2018. https://aws.amazon.com/sumerian/. 2

[2] Matthew Brand. Voice puppetry. In *ACM Transaction on Graphics*, 1999. 1

[3] Matthew Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28, 1999. 2

[4] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997. 1, 2

[5] Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. In *ACM Transaction on Graphics*, 2005. 3

[6] Hang Chu, Shugao Ma, Fernando De la Torre, Sanja Fidler, and Yaser Sheikh. Expressive telepresence via modular codec avatars. In *European Conf. on Computer Vision*, pages 330–345, 2020. 3

[7] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *British Machine Vision Conference*, 2017. 2

[8] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, learning, and synthesis of 3d speaking styles. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2019. 2, 3, 5, 7

[9] José Mario De Martino, Léo Pini Magalhães, and Fábio Violaro. Facial animation based on context-dependent visemes. *Computers & Graphics*, 30(6):971–980, 2006. 2

[10] Zhigang Deng, Pei-Ying Chiang, Pamela Fox, and Ulrich Neumann. Animating blendshape faces by cross-mapping motion capture data. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 43–48, 2006. 3

[11] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. JALI: An animator-centric viseme model for expressive lip synchronization. *ACM Transaction on Graphics*, 35(4), 2016. 2

[12] Sefik Emre Eskimez, Ross K. Maddox, Chenliang Xu, and Zhiyao Duan. Generating talking face landmarks from speech. In *Int. Conf. on Latent Variable Analysis and Signal Separation*, pages 372–381, 2018. 3

[13] Tony Ezzat and Tomaso Poggio. Miketalk: A talking facial display based on morphing visemes. In *Computer Animation*, pages 96–102, 1998. 2

[14] Tony Ezzat and Tomaso Poggio. Visual speech synthesis by morphing visemes. *International Journal on Computer Vision*, 38(1):45–57, 2000. 2

[15] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transaction on Graphics*, 35(3), 2016. 3

[16] David Greenwook, Iain Matthews, and Stephen D. Laycock. Joint learning of facial expression and head pose from speech. In *Interspeech*, pages 2484–2488, 2018. 3

[17] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 7

[18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Int. Conf. on Learning Representations*, 2017. 4

[19] Sam Johnson, Colin Lea, and Ronit Kassis. Tech note: Enhancing oculus lipsync with deep learning, 2018. https://developer.oculus.com/blog/tech-note-enhancing-oculus-lipsync-with\-deep-learning/. 2

[20] Gregor A Kalberer, Pascal Müller, and Luc Van Gool. Speech animation using viseme space. In *Int. Symposium on Vision, Modeling, and Visualization*, pages 463–470, 2002. 2

[21] Gregor A Kalberer and Luc Van Gool. Face animation based on observed 3d speech dynamics. In *Computer Animation*, 2001. 2

[22] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transaction on Graphics*, 36(4), 2017. 1, 2, 3, 5

[23] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transaction on Graphics*, 37(4), 2018. 3

[24] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. End-to-end learning for 3d facial animation from speech. In *ACM Int. Conf. on Multimodal Interaction*, page 361–365, 2018. 3

[25] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and C V Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Conf. on Multimedia*, page 484–492, 2020. 2

[26] Alexander Richard, Colin Lea, Shugao Ma, Juergen Gall, Fernando de la Torre, and Yaser Sheikh. Audio- and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 41–50, 2021. 1, 2, 3, 4, 5

[27] Yuge Shi, Narayanaswamy Siddharth, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, 2019. 5

[28] Yang Song, Jingwen Zhu, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *Int. Joint Conf. on Artificial Intelligence*, 2019. 2

[29] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Transaction on Graphics*, 36(4), 2017. 2

[30] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transaction on Graphics*, 36(4), 2017. 2, 3, 5

[31] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *ECCV 2020*, 2020. 2

[32] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A

generative model for raw audio. In *ISCA Speech Synthesis Workshop*, page 125, 2016. 3

[33] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, 2016. 3, 5

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 3

[35] Ashish Verma, Nitendra Rajput, and L Venkata Subramaniam. Using viseme based acoustic models for speech driven lip synthesis. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2003. 2

[36] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. In *British Machine Vision Conference*, 2018. 2

[37] Lijuan Wang, Wei Han, and Frank K Soong. High quality lip-sync animation for 3d photo-realistic talking head. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 4529–4532, 2012. 3

[38] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 5

[39] Shih-En Wei, jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Purdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. VR facial animation via multiview image translation. *ACM Transaction on Graphics*, 38(4), 2019. 3

[40] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *European Conf. on Computer Vision*, pages 670–686, 2018. 2

[41] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conf. on Artificial Intelligence*, pages 9299–9306, 2019. 2

[42] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transaction on Graphics*, 37(4), 2018. 2