

# Capture, Learning, and Synthesis of 3D Speaking Styles

Daniel Cudeiro<sup>\*†</sup>    Timo Bolkart<sup>\*</sup>    Cassidy Laidlaw  
Anurag Ranjan    Michael J. Black  
Max Planck Institute for Intelligent Systems, Tübingen, Germany  
{tbolkart, claidlaw, aranjean, black}@tuebingen.mpg.de

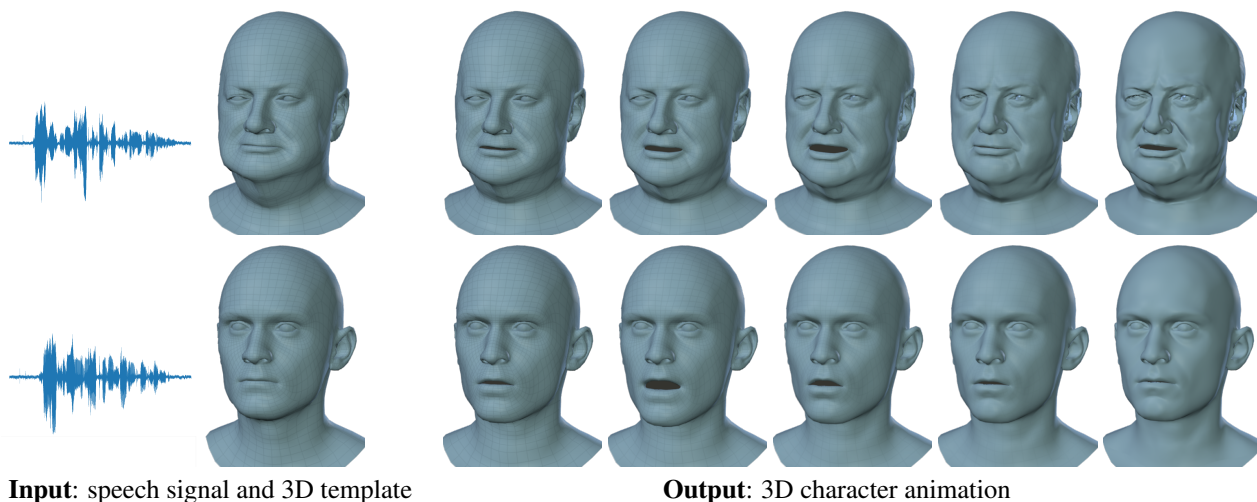


Figure 1: Given an arbitrary speech signal and a static 3D face mesh as input (left), our model, VOCA outputs a realistic 3D character animation (right). Top: Winston Churchill. Bottom: Actor from Karras et al. [33]. See [supplementary video](#).

## Abstract

Audio-driven 3D facial animation has been widely explored, but achieving realistic, human-like performance is still unsolved. This is due to the lack of available 3D datasets, models, and standard evaluation metrics. To address this, we introduce a unique 4D face dataset with about 29 minutes of 4D scans captured at 60 fps and synchronized audio from 12 speakers. We then train a neural network on our dataset that factors identity from facial motion. The learned model, VOCA (Voice Operated Character Animation) takes any speech signal as input—even speech in languages other than English—and realistically animates a wide range of adult faces. Conditioning on subject labels during training allows the model to learn a variety of realistic speaking styles. VOCA also provides animator controls to alter speaking style, identity-dependent facial shape, and pose (i.e. head, jaw, and eyeball rotations) during animation. To our knowledge, VOCA is the only realistic 3D facial animation model that is read-

ily applicable to unseen subjects without retargeting. This makes VOCA suitable for tasks like in-game video, virtual reality avatars, or any scenario in which the speaker, speech, or language is not known in advance. We make the dataset and model available for research purposes at <http://voca.is.tue.mpg.de>.

## 1. Introduction

Teaching computers to see and understand faces is critical for them to understand human behavior. There is an extensive literature on estimating 3D face shape, facial expressions, and facial motion from images and videos. Less attention has been paid to estimating 3D properties of faces from sound; however, many facial motions are caused directly by the production of speech. Understanding the correlation between speech and facial motion thus provides ad-

<sup>\*</sup> Equal contribution

<sup>†</sup> Deceased, December 5, 2018

ditional valuable information for analyzing humans, particularly if visual data are noisy, missing, or ambiguous. The relation between speech and facial motion has previously been used to separate audio-visual speech [22] and for audio-video driven facial animation [38]. Missing to date is a general and robust method that relates the speech of *any* person in *any* language to the 3D facial motion of *any* face shape. Here we present *VOCA (Voice Operated Character Animation)*, that takes a step towards this goal.

While speech-driven 3D facial animation has been widely studied, speaker-independent modeling remains a challenging, unsolved task for several reasons. First, speech signals and facial motion are strongly correlated but lie in two very different spaces; thus, non-linear regression functions are needed to relate the two. One can exploit deep neural networks to address this problem. However, this means that significant amounts of training data are needed. Second, there exists a many-to-many mapping between phonemes and facial motion. This poses an even greater challenge when training across people and styles. Third, because we are especially sensitive to faces, particularly realistic faces, the animation must be realistic to avoid falling into the Uncanny Valley [39]. Fourth, there is very limited training data relating speech to the 3D face shape of multiple speakers. Finally, while previous work has shown that models can be trained to create speaker-specific animations [14, 33], there are no generic methods that are speaker independent and that capture a variety of speaking styles.

**VOCASET:** To address this, we collected a new dataset of 4D face scans together with speech. The dataset has 12 subjects and 480 sequences of about 3-4 seconds each with sentences chosen from an array of standard protocols that maximize phonetic diversity. The 4D scans are captured at 60fps and we align a common face template mesh to all the scans, bringing them into correspondence. This dataset, called *VOCASET*, is unlike any existing public datasets. It allows training and testing of speech-to-animation models that can generalize to new data.

**VOCA:** Given such data, we train a deep neural network model, called *VOCA* (Figure 2), that generalizes to new speakers (see Figure 1). Recent work using deep networks has shown impressive results for the problem of regressing *speaker-dependent* facial animation from speech [33]. Their work, however, captures the idiosyncrasies of an individual, making it inappropriate for generalization across characters. While deep learning is advancing the field quickly, even the best recent methods rely on some manual processes or focus only on the mouth [53], making them inappropriate for truly automatic full facial animation.

The key problem with prior work is that facial motion and facial identity are confounded. Our key insight is to factor identity from facial motions and then learn a model relating speech to only the motions. Conditioning on subject

labels during training allows us to combine data from many subjects in the training process, which enables the model both to generalize to new subjects not seen during training and to synthesize different speaker styles. Integrating DeepSpeech [29] for audio feature extraction makes *VOCA* robust w.r.t. different audio sources and noise. Building on top of the expressive FLAME head model [37] allows us i) to model motions of the full face (i.e. including the neck), ii) to animate a wide range of adult faces, as FLAME can be used to reconstruct subject-specific templates from a scan or image, and iii) to edit identity-dependent shape and head pose during animation. *VOCA* and *VOCASET* are available for research purposes [58].

## 2. Related work

Facial animation has received significant attention in the literature. Related work in this area can be grouped into three categories: speech-based, text-based, and video- or performance-based.

**Speech-driven facial animation:** Due to the abundance of images and videos, many methods that attempt to realistically animate faces use monocular video [7, 8, 16, 23, 51, 59, 62]. Bregler et al. [8] transcribe speech with a Hidden Markov Model (HMM) into phonetic labels and animate the mouth region in videos with an exemplar-based video warping. Brand [7] uses a mix of Linear Predictive Coding (LPC) and RASTA-PLP [30] audio features and an HMM to output a sequence of facial motion vectors. Ezzat et al. [23] perform Principal Component Analysis (PCA) on all images and use an example-based mapping between phonemes and trajectories of mouth shape and mouth texture parameters in the PCA space. Xie and Liu [62] model facial animation with a dynamic Bayesian network-based model. Wang et al. [59] use an HMM to learn a mapping between Mel-frequency Cepstral Coefficients (MFCC) and PCA model parameters. Zhang et al. [67] combine the HMM-based method of [59] trained on audio and visual data of one actor with a deep neural network based encoder trained from hundreds of hours of speaker independent speech data to compute an embedding of the MFCC audio features. Shimba et al. [48] use a deep Long Short-term Memory (LSTM) network to regress active appearance model (AAM) parameters from MFCC features. Chen et al. [16] correlate audio and image motion to synthesize lip motion of arbitrary identities.

Suwajanakorn et al. [51] use an Recurrent Neural Network (RNN) for synthesizing photorealistic mouth texture animations using audio from 1.9 million frames from Obama’s weekly addresses. However, their method does not generalize to unseen faces or viewpoints. In contrast to this, *VOCA* is trained across subjects sharing a common topology, which makes it possible to animate new faces from previously unseen viewpoints. Pham et al. [42] regress

global transformation and blendshape coefficients [13] from MFCC audio features using an LSTM network. While their model is trained across subjects—similar to VOCA—they rely on model parameters regressed from 2D videos rather than using 3D scans, which limits their quality.

A few methods use multi-view motion capture data [10, 14] or high-resolution 3D scans [33]. Busso et al. [10] synthesize rigid head motion in expressive speech sequences. Cao et al. [14] segment the audio into phonemes and use an example-based graph method to select a matching mouth animation. Karras et al. [33] propose a convolutional model for mapping LPC audio features to 3D vertex displacements. However, their model is subject specific, and animating a new face would require 3D capture and processing of thousands of frames of subject data. Our model, VOCA factors identity from facial motion and is trained across subjects, which allows animation of a wide range of adult faces.

Several works also aim at animating artist designed character rigs [20, 21, 31, 32, 46, 52, 53, 54, 70]. Taylor et al. [53] propose a deep-learning based speech-driven facial animation model using a sliding window approach on transcribed phoneme sequences that outperforms previous LSTM based methods [24, 25]. While these models are similar to VOCA in that they animate a generic face from audio, our focus is animating a realistic face mesh, for which we train our model on high-resolution face-scans.

**Text-driven facial animation:** Some methods aim to animate faces directly from text. Sako et al. [45] use a hidden Markov model to animate lips in images from text. Anderson et al. [4] use an extended hidden Markov text-to-speech model to drive a subject-specific active appearance model (AAM). In a follow-up, they extend this approach to animate the face of an actress in 3D. While our focus is not to animate faces from text, this is possible by animating our model with the output of a text-to-speech (TTS) system (e.g. [56]), similar to Karras et al. [33].

**Performance-based facial animation:** Most methods to animate digital avatars are based on visual data. Alexander et al. [3], Wu et al. [61], and Laine et al. [35] build subject-specific face-rigs from high-resolution face scans and animate these rigs with video-based animation systems.

Several methods build personalized face-rigs using generic face models from monocular videos to transfer and reenact facial performance between videos. Tensor-based multilinear face models [5, 11, 12, 19, 57, 63] and linear models [55] are widely used to build personalized face-rigs. Cao et al. [11, 12] use a regression-based face tracker to animate the face-rig and digital avatars, while Thies et al. [55] use a landmark-based face tracker and deformation transfer [50] to reenact monocular videos.

Other methods that animate virtual avatars rely on RGB-D videos or 4D sequences to track and retarget facial performance. Li et al. [36] and Weise et al. [60] capture example-

based rigs in an offline calibration procedure to build personalized face-rigs, Bouaziz et al. [6] use a generic identity model. Liu et al. [38] combine audio and video to robustly animate a generic face model from RGB-D video. Li et al. [37] capture facial performance with a high-resolution scanner and animate static face meshes using an articulated generic head model. In contrast to these methods, our approach solely relies on audio to animate digital avatars.

**3D face datasets:** Several 3D face datasets have been released that focus on the analysis of static 3D facial shape and expression (e.g. [13, 47, 65]) or dynamic facial expressions (e.g. [2, 15, 18, 44, 64, 68, 69]). Most of these datasets focus on emotional expressions and only a few datasets capture facial dynamics caused by speech. The recently published 4DFAB dataset [17] contains 4D captures of 180 subjects, but with only nine word utterances per subject and lower mesh quality than VOCASET.

The B3D(AC)<sup>2</sup> dataset [26] contains a large set of audio-4D scan pairs of 40 spoken English sentences. In contrast, VOCASET contains 255 unique sentences in total. To enable training on both a large number of sentences and subjects, some sentences are shared across subjects and some sentences are spoken by only one subject. The visible artifacts present in the raw B3D(AC)<sup>2</sup> scans (i.e. holes and capture noise) mean that subtle facial motions may be lost; also, the registered template only covers the face, ignoring speech-related motions in the neck region. VOCASET, in comparison, provides higher-quality 3D scans as well as alignments of the entire head, including the neck.

### 3. Preliminaries

Our goal for VOCA is to generalize well to arbitrary subjects not seen during training. Generalization across subjects involves both (i) generalization across different speakers in terms of the audio (variations in accent, speed, audio source, noise, environment, etc.) and (ii) generalization across different facial shapes and motion.

**DeepSpeech:** To gain robustness to different audio sources, regardless of noise, recording artifacts, or language, we integrate DeepSpeech [29] into our model. DeepSpeech [29] is an end-to-end deep learning model for Automatic Speech Recognition (ASR). DeepSpeech uses a simple architecture consisting of five layers of hidden units, of which the first three layers are non-recurrent fully connected layers with ReLU activations. The fourth layer is a bi-directional RNN, and the final layer is a fully connected layer with ReLU activation. The final layer of the network is fed to a softmax function whose output is a probability distribution over characters. The TensorFlow implementation provided by Mozilla [40] slightly differs from the original paper in two ways: (i) the RNN units are replaced by LSTM cells and (ii) 26 MFCC audio features are used instead of directly performing inference on the spectrogram.

| Type            | Kernel | Stride | Output  | Activation |
|-----------------|--------|--------|---------|------------|
| DeepSpeech      | -      | -      | 16x1x29 | -          |
| Identity concat | -      | -      | 16x1x37 | -          |
| Convolution     | 3x1    | 2x1    | 8x1x32  | ReLU       |
| Convolution     | 3x1    | 2x1    | 4x1x32  | ReLU       |
| Convolution     | 3x1    | 2x1    | 2x1x64  | ReLU       |
| Convolution     | 3x1    | 2x1    | 1x1x64  | ReLU       |
| Identity concat | -      | -      | 72      | -          |
| Fully connected | -      | -      | 128     | tanh       |
| Fully connected | -      | -      | 50      | linear     |
| Fully connected | -      | -      | 5023x3  | linear     |

Table 1: Model architecture.

Please see [40] for more details.

**FLAME:** Facial shape and head motion vary greatly across subjects. Furthermore, different people have different speaking styles. The large variability in facial shape, motion, and speaking style motivates using a common learning space. We address this problem by incorporating FLAME, a publicly available statistical head model, as part of our animation pipeline. FLAME uses linear transformations to describe identity and expression dependent shape variations, and standard linear blend skinning (LBS) to model neck, jaw, and eyeball rotations. Given a template  $\mathbf{T} \in \mathbb{R}^{3N}$  in the “zero pose”, identity, pose, and expression blendshapes are modeled as vertex offsets from  $\mathbf{T}$ . For more details we refer the reader to [37].

## 4. VOCA

This section describes the model architecture and provides details on how the input audio is processed.

**Overview:** VOCA receives as input a subject-specific template  $\mathbf{T}$  and the raw audio signal, from which we extract features using DeepSpeech [29]. The desired output is the target 3D mesh. VOCA acts as an encoder-decoder network (see Figure 2 and Table 1) where the encoder learns to transform audio features to a low-dimensional embedding and the decoder maps this embedding into a high-dimensional space of 3D vertex displacements

**Speech feature extraction:** Given an input audio clip of length  $T$  seconds, we use DeepSpeech to extract speech features. The outputs are unnormalized log probabilities of characters for frames of length 0.02 s (50 frames per second); thus, it is an array of size  $50T \times D$ , where  $D$  is the number of characters in the alphabet plus one for a blank label. We resample the output to 60 fps using linear interpolation. In order to incorporate temporal information, we convert the audio frames to overlapping windows of size  $W \times D$ , where  $W$  is the window size. The output is a three-dimensional array of dimensions  $60T \times W \times D$ .

**Encoder:** The encoder is composed of four convolutional layers and two fully connected layers. The speech features and the final convolutional layer are conditioned

on the subject labels to learn subject-specific styles when trained across multiple subjects. For eight training subjects, each subject  $j$  is encoded as an one-hot-vector  $I_j = (\delta_{ij})_{1 \leq i \leq 8}$ . This vector is concatenated to each  $D$ -dimensional speech feature vector (i.e. resulting in windows of dimension  $W \times (D + 8)$ ), and concatenated to the output of the final convolution layer.

To learn temporal features and reduce the dimensionality of the input, each convolutional layer uses a kernel of dimension  $3 \times 1$  and stride  $2 \times 1$ . As the features extracted using DeepSpeech do not have any spatial correlation, we reshape the input window to have dimensions  $W \times 1 \times (D + 8)$  and perform 1D convolutions over the temporal dimension. To avoid overfitting, we keep the number of parameters small and only learn 32 filters for the first two, and 64 filters for the last two convolutional layers.

The concatenation of the final convolutional layer with the subject encoding is followed by two fully connected layers. The first has 128 units and a hyperbolic tangent activation function; the second is a linear layer with 50 units.

**Decoder:** The decoder of VOCA is a fully connected layer with linear activation function, outputting the  $5023 \times 3$  dimensional array of vertex displacements from  $\mathbf{T}$ . The weights of the layer are initialized by 50 PCA components computed over the vertex displacements of the training data; the bias is initialized with zeros.

**Animation control:** During inference, changing the eight-dimensional one-hot-vector alters the output speaking style. The output of VOCA is an expressed 3D face in “zero pose” with the same mesh topology as the FLAME face model [37]. VOCA’s compatibility with FLAME allows alteration of the identity-dependent facial shape by adding weighted shape blendshapes from FLAME. The face expression and pose (i.e. head, jaw, and eyeball rotations) can also be changed using the blendweights, joints, and pose blendshapes provided by FLAME.

## 5. Model training

In this section we describe training relevant details.

**Training set-up:** We start from a large dataset of audio-4D scan pairs, denoted as  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^F$ . Here  $\mathbf{x}_i \in \mathbb{R}^{W \times D}$  is the input audio window centered at the  $i$ th video frame,  $\mathbf{y}_i \in \mathbb{R}^{N \times 3}$ . Further, let  $\mathbf{f}_i \in \mathbb{R}^{N \times 3}$  denote the output of VOCA for  $\mathbf{x}_i$ .

For training, we split the captured data into a training set (eight subjects), a validation set (two subjects), and a test set (two subjects). The training set consists of all 40 sentences of the eight subjects, i.e. in total 320 sentences. For validation and test data, we only select the 20 unique sentences that are not shared with any other subject, i.e. 40 sentences for validation and testing, respectively. Note that our training, validation, and test sets for all experiments are fully disjoint, i.e. no overlap of subjects or sentences exists.



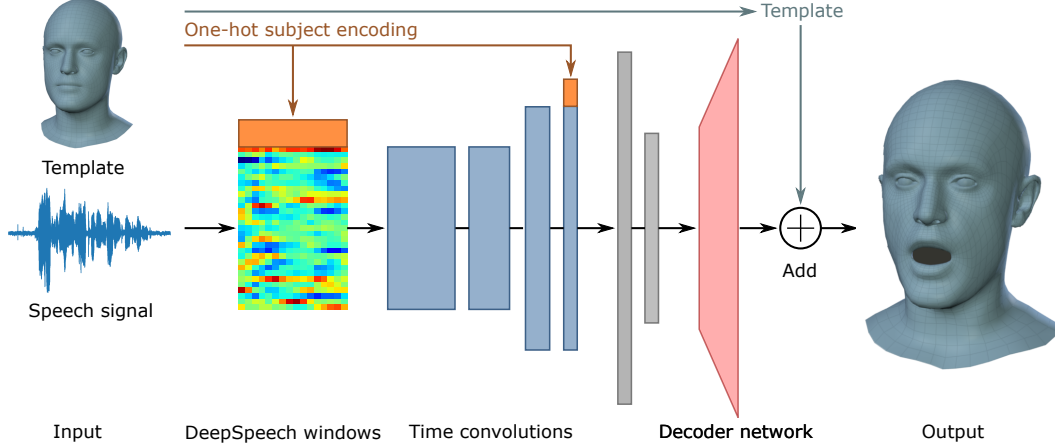


Figure 2: VOCA network architecture.

**Loss function:** Our training loss function consists of two terms, a position term and a velocity term. The position term  $E_p = \|\mathbf{y}_i - \mathbf{f}_i\|_F^2$  computes the distance between the predicted outputs and the training vertices. This position term encourages the model to match the ground truth performance. The velocity term  $E_v = \|(\mathbf{y}_i - \mathbf{y}_{i-1}) - (\mathbf{f}_i - \mathbf{f}_{i-1})\|_F^2$  uses backward finite differences. It computes the distance between the differences of consecutive frames between predicted outputs and training vertices. This velocity term induces temporal stability.

**Training parameters:** We perform hyperparameter tuning on the held-out validation set. We train VOCA for 50 epochs with a constant learning rate of  $1e - 4$ . The weights for the position and velocity terms are 1.0 and 10.0, respectively. During training, we use batch normalization with a batch size of 64. We use a window size of  $W = 16$  with  $D = 29$  speech features.

**Implementation details:** VOCA is implemented in Python using TensorFlow [1], and trained using Adam [34]. Training one epoch takes about ten minutes on a single NVIDIA Tesla K20. We use a pre-trained DeepSpeech model [40] which is kept fixed during training.

## 6. VOCASET

This section introduces VOCASET and describes the capture setup and data processing.

**VOCASET:** Our dataset contains a collection of audio-4D scan pairs captured from 6 female and 6 male subjects. For each subject, we collect 40 sequences of a sentence spoken in English, each of length three to five seconds. The sentences were taken from an array of standard protocols and were selected to maximize phonetic diversity using the method described in [27]. In particular, each subject spoke 27 sentences from the TIMIT corpus [28], three pangrams used by [33], and 10 questions from the Stanford Question

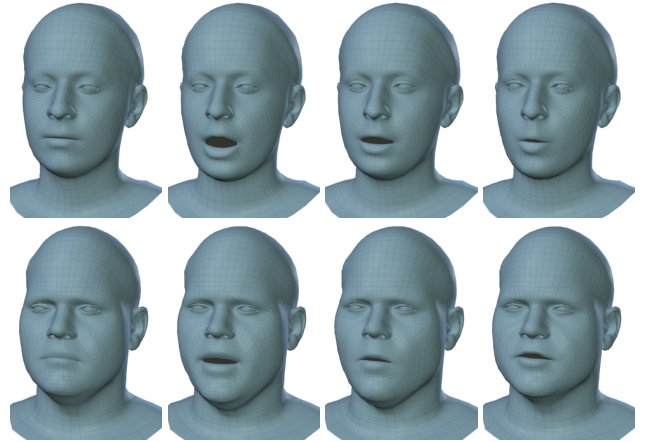


Figure 3: Sample meshes of two VOCASET subjects.

Answering Dataset (SQuAD) [43]. The recorded sequences are distributed such that five sentences are shared across all subjects, 15 sentences are spoken by three to five subjects (50 unique sentences), and 20 sentences are spoken only by one or two subjects (200 unique sentences). We make VOCASET available to the research community.

**Capture setup:** We use a multi-camera active stereo system (3dMD LLC, Atlanta) to capture high-quality 3D head scans and audio. The capture system consists of six pairs of gray-scale stereo cameras, six color cameras, five speckle pattern projectors, and six white light LED panels. The system captures 3D meshes at 60fps, each with about 120K vertices. The color images are used to generate UV texture maps for each scan. The audio, synchronized with the scanner, is captured with a sample rate of 22 kHz.

**Data processing:** The raw 3D head scans are registered with a sequential alignment method as described in [37] using the publicly available generic FLAME model. The

image-based landmark prediction method of [9] is used during alignment to add robustness while tracking fast facial motions. After alignment, each mesh consists of 5023 3D vertices. For all scans, we measure the absolute distance between each scan vertex and the closest point in the FLAME alignment surface: median (0.09mm), mean (0.13mm), and standard deviation (0.14mm). Thus, the alignments faithfully represent the raw data.

All meshes are then unposed; i.e. effects of global rotation, translation, and head rotation around the neck are removed. After unposing, all meshes are in “zero pose”. For each sequence, the neck boundary and the ears are automatically fixed, and the region around the eyes is smoothed using Gaussian filtering to remove capture noise. Note that no smoothing is applied to the mouth region so as to preserve subtle motions. Figure 3 shows sample alignments of two VOCASET subjects. The supplementary video shows sequences of all subjects.

## 7. Experiments

Quantitative metrics, such as the norm on the prediction error, are not suitable for evaluating animation quality. This is because facial visemes form many-to-many mappings with speech utterances. A wide range of plausible facial motions exists for the same speech sequence, which makes quantitative evaluation intractable. Instead, we perform perceptual and qualitative evaluations. Further, our trained model is available for research purposes for direct comparisons [58].

### 7.1. Perceptual evaluation

**User study:** We conduct three Amazon Mechanical Turk (AMT) blind user studies: i) a binary comparison between held-out test sequences and our model conditioned on all training subjects, ii) an ablation study to assess the effectiveness of the DeepSpeech features, and iii) a study to investigate the correlation between style, content, and identity. All experiments are performed on sequences and subjects fully disjoint from our training and validation set.

For binary comparisons, two videos with the same animated subject and audio clip are shown side by side. For each video pair, the participant is asked to choose the talking head that moves more naturally and in accordance with the audio. To avoid any selection bias, the order (left/right) of all methods for comparison is random for each pair.

Style comparisons are used to evaluate the learned speaking styles. Here, Turkers see three videos: one reference and two predictions. The task is to determine which of the two predictions is more similar to the reference video.

To ensure the quality of the study and remove potential outliers, we require Turkers to pass a simple qualification test before they are allowed to submit HITs. The qualification task is a simplified version of the following user study,

where we show three comparisons with an obvious answer, i.e. one ground-truth sequence and one sequence with completely mismatched video and audio.

**Comparison to recorded performance:** We compare captured and processed test sequences with VOCA predictions conditioned on all eight speaker styles. In total, Turkers (400 HITs) perceived the recorded performance more natural ( $83 \pm 9\%$ ) than the predictions ( $17 \pm 9\%$ ), across all conditions. While VOCA results in realistic facial motion for the unseen subjects, it is unable to synthesize the idiosyncrasies of these subjects. As such, these subtle subject-specific details make the recorded sequences look more natural than the predictions.

**Speech feature ablation:** We replace the DeepSpeech features by Mel-filterbank energy features (fbank) and train a model for 50 epochs (the same as for VOCA). Turkers (400 HITs) perceived the performance of VOCA with DeepSpeech more natural ( $78 \pm 16\%$ ) than with fbank features ( $22 \pm 16\%$ ) across all conditions. That indicates that VOCA with DeepSpeech features generalizes better to unseen audio sequences than with fbank features.

**Style comparisons:** Speech-driven facial performance varies greatly across subjects. However, it is difficult to separate between style (facial motion of a subject), identity (facial shape of a subject), and content (the words being said), and how these different factors influence perception. The goal of this user study is to evaluate the speech-driven facial motion independently from identity-dependent face shape in order to understand if people can recognize the styles learned by our model.

To accomplish this, we subtract the personalized template (neutral face) from all sequences to obtain “displacements”, then add these displacements to a single common template (randomly sampled from the FLAME shape space). Then, for several reference sequences from the training data, we compare two VOCA predictions (on audio from the test set): one conditioned on the reference subject and one conditioned on another randomly selected subject. We ask Turkers to select which predicted sequence is more similar in speaking style to the reference.

To explore the influence of content, we perform the experiment twice, once where the reference video and the predictions share the same sentence (spoken by different subjects) and once with different sentences. Figure 4 shows the results for this experiment. Results varied greatly across conditions. For some conditions, Turkers could consistently pick the sequence with the matching style (e.g. conditions 3, 4, and 5); for others, their choices were no better or worse than chance. The impact of the content was not significant for most conditions. More research is needed to understand which factors are important for people to recognize different speaking styles, and to develop new models that more efficiently disentangle facial shape and motion.

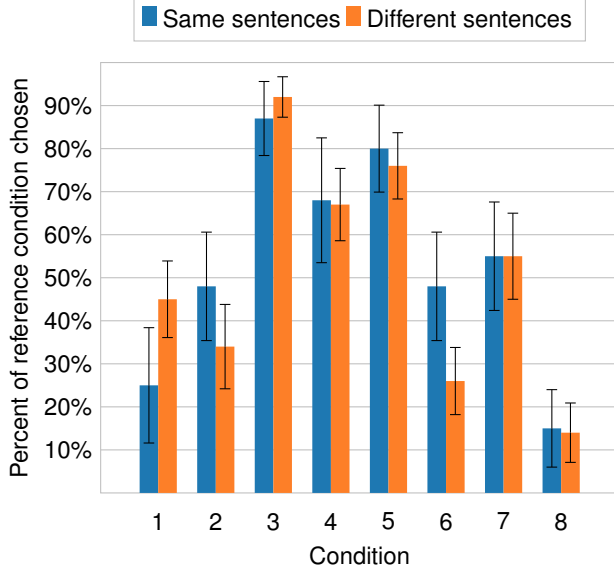


Figure 4: AMT study of styles. The bars show the percentage of Turkers choosing the reference condition when the same sentence was being shown for reference and prediction, and with difference sentences.

## 7.2. Qualitative evaluation

**Generalization across subjects:** Factoring identity from facial motion allows us to animate a wide range of adult faces. To show the generalization capabilities of VOCA, we select, align and pose-normalize multiple neutral scans from the BU-3DFE database [66], with large shape variations. Figure 5 shows the static template (left) and some VOCA animation frames, driven by the same audio sequence.

**Generalization across languages:** The video shows the VOCA output for different languages. This indicates that VOCA can generalize to non-English sentences.

**Speaker styles:** Conditioning on different subjects during inference results in different speaking styles. Stylistic differences include variation in lip articulation. Figure 6 shows the distance between lower and upper lip as a function of time for VOCA predictions for a random audio sequence and different conditions. This indicates that the convex combination of styles provides a wide range of different mouth amplitudes.

We generate new intermediate speaking styles by convex combinations of conditions. Due to the linearity of the decoder, performing this convex combination in the 3D vertex space or in the 50-dimensional encoding space is equivalent. The supplementary video shows that combining styles offers animation control to synthesize a range of varying speaking styles. This is potentially useful for matching the speaking performance of a subject not seen during training.

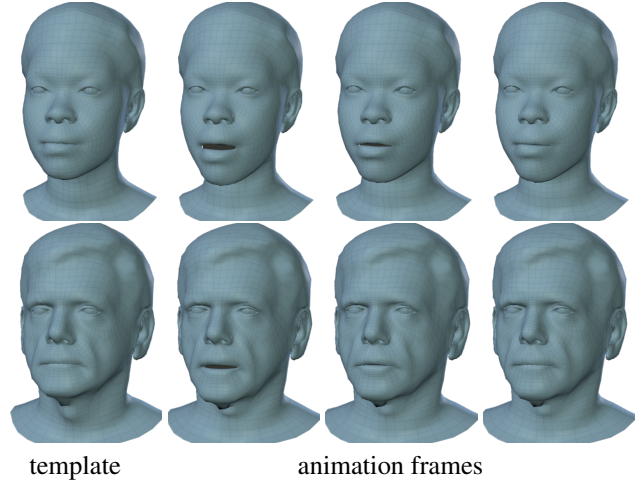


Figure 5: VOCA generalizes across face shapes. Each row shows the template of a subject selected from the static BU-3DFE face database [66] (left), and three randomly selected animation frames, driven by the same audio input (right).

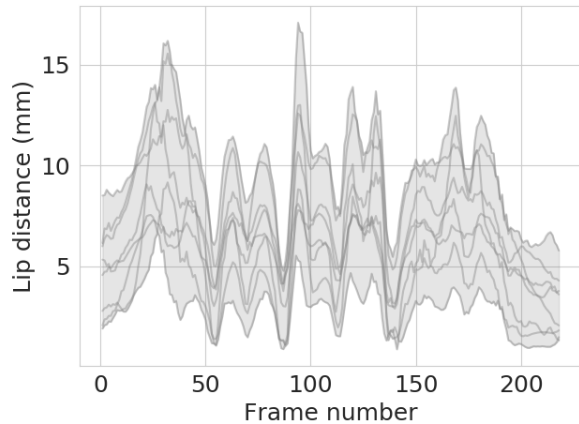


Figure 6: Distance between lower and upper lip for VOCA predictions conditioned on different subjects. The shaded region represents the space of convex combinations of the different conditions.

In the future, this could be estimated from video.

**Robustness to noise:** To demonstrate robustness to noise, we combine a speech signal with different levels of noise and use the noisy signal as VOCA input. As a noise source, we use a realistic street noise sequence [49] added with negative gain of 36dB (low), 24dB (medium), 18dB (slightly high), and 12dB (high). Only the high noise level leads to a damped facial motion, but despite the noise, the facial animations remain plausible.

**Comparison to Karras et al. [33]:** We compare VOCA to Karras et al. [33], the state-of-the-art in realistic subject-specific audio-driven facial animation. The results are

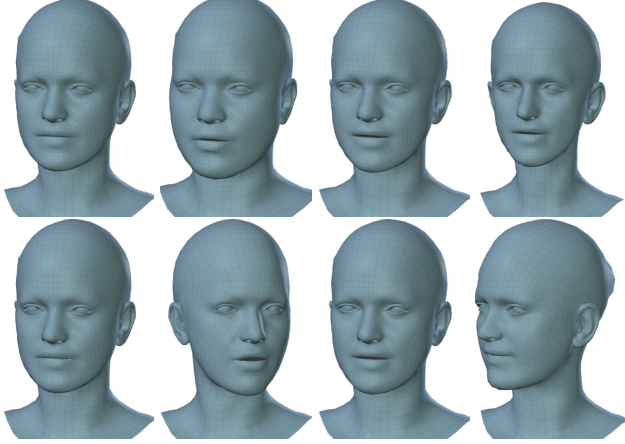


Figure 7: Animation control. Top: varying the first identity shape components to plus two (second column) and minus two (last column) standard deviations. Bottom: varying the head pose to minus 30 degrees (second column) and plus 30 degrees (last column).

shown in the supplementary video. For comparison, the authors provided us with a static mesh, to which we aligned the FLAME topology. We then use eight audio sequences from their supplementary video (including singing, spoken Chinese, an excerpt of a Barack Obama speech, and different sequences of the actor), to animate their static mesh. The supplementary video shows that, while their model produces more natural and detailed results, we can still reproduce similar facial animation without using any of their subject-specific training data. Further, Karras et al. use professional actors capable of simulating emotional speech. This enables them to add more realism in the upper face by modeling motions (i.e. eyes and eyebrows) that are more correlated with emotions than speech.

**Animation control:** Figure 7 demonstrates the possibility of changing the identity dependent shape (top) and head pose (bottom) during animation. Both rows are driven by the same audio sequence. Despite the varying shape or pose, the facial animation looks realistic.

## 8. Discussion

While VOCA can be used to realistically animate a wide range of adults faces from speech, it still lacks some of the details needed for conversational realism. Upper face motions (i.e. eyes and eyebrows) are not strongly correlated with the audio [33]. The causal factor is emotion, which is absent in our data due the inherent difficulty of simulating emotional speech in a controlled capture environment. Thus, VOCA learns the causal facial motions from speech, which are mostly present in the lower face.

Non-verbal communication cues, such as head motion,

are weakly correlated with the audio signal and hence are not modeled well by audio-driven techniques. VOCA offers animators and developers the possibility to include head motion, but does not infer it from data. A speech independent model for head motion could be used to simulate realistic results. Application specific techniques, such as dyadic interactions between animated assistants and humans require attention mechanisms that consider spatial features, such as eye tracking. Learning richer conversation models with expressive bodies [41] is future research.

Conditioning on subject labels is one of the key aspects of VOCA that allows training across subjects. This allows a user to alter the speaking style during inference. Using data from more subjects to increase the number of different speaking styles remains a task for future work. Further experiments on mitigating or amplifying different speaking styles, or combining characteristics of different subjects also remain for future work.

## 9. Conclusion

We have presented VOCA, a simple and generic speech-driven facial animation framework that works across a range of identities. Given an arbitrary speech signal and a static character mesh, VOCA fully automatically outputs a realistic character animation. VOCA leverages recent advances in speech processing and 3D face modeling in order to be subject independent. We train our model on a self-captured multi-subject 4D face dataset (VOCASET). The key insights of VOCA are to factor identity from facial motion, which allows us to animate a wide range of adult faces, and to condition on subject labels, which enables us to train VOCA across multiple subjects, and to synthesize different speaker styles during test time. VOCA generalizes well across various speech sources, languages, and 3D face templates. We provide optional animation control parameters to vary the speaking style and to alter the identity dependent shape and head pose during animation. The dataset, trained model, and code are available for research purposes [58].

## Acknowledgement

We thank T. Alexiadis and J. Márquez for the data acquisition, B. Pellkofer for hardware support, A. Quiros-Ramires for support with MTurk, A. Osman for support with Tensorflow, and S. Pujades for help with finalizing the paper. We further thank Karras et al. for providing us with a static face mesh for comparison.

MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI. MJB has financial interests in Amazon and Meshcapade GmbH.

**In loving memory of Daniel Cudeiro.**



## References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. [5](#)
- [2] T. Alashkar, B. Ben Amor, M. Daoudi, and S. Berretti. A 3D dynamic database for unconstrained face recognition. In *International Conference and Exhibition on 3D Body Scanning Technologies*, 2014. [3](#)
- [3] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. The digital emily project: Photoreal facial modeling and animation. In *SIGGRAPH 2009 Courses*, pages 12:1–12:15, 2009. [3](#)
- [4] R. Anderson, B. Stenger, V. Wan, and R. Cipolla. Expressive visual text-to-speech using active appearance models. In *Conference on Computer Vision and Pattern Recognition*, pages 3382–3389, 2013. [3](#)
- [5] T. Bolkart and S. Wuhler. A groupwise multilinear correspondence optimization for 3D faces. In *International Conference on Computer Vision*, pages 3604–3612, 2015. [3](#)
- [6] S. Bouaziz, Y. Wang, and M. Pauly. Online modeling for real-time facial animation. *Transactions on Graphics*, 32(4):40, 2013. [3](#)
- [7] M. Brand. Voice puppetry. In *SIGGRAPH*, pages 21–28, 1999. [2](#)
- [8] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *SIGGRAPH*, pages 353–360, 1997. [2](#)
- [9] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*, pages 1021–1030, 2017. [6](#)
- [10] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *Transactions on Audio, Speech, and Language Processing*, 15(3):1075–1086, 2007. [3](#)
- [11] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 34(4):46:1–46:9, 2015. [3](#)
- [12] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 33(4):43:1–43:10, 2014. [3](#)
- [13] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Faceware-house: A 3D facial expression database for visual computing. *Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. [3](#)
- [14] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *Transactions on Graphics*, 24(4):1283–1302, 2005. [2](#), [3](#)
- [15] Y. Chang, M. Vieira, M. Turk, and L. Velho. Automatic 3D facial expression analysis in videos. In *Analysis and Modelling of Faces and Gestures*, pages 293–307, 2005. [3](#)
- [16] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *European Conference on Computer Vision*, pages 538–553, 2018. [2](#)
- [17] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou. 4DFAB: A large scale 4d database for facial expression analysis and biometric applications. In *CVPR*, 2018. [3](#)
- [18] D. Cosker, E. Krumhuber, and A. Hilton. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In *International Conference on Computer Vision*, pages 2296–2303, 2011. [3](#)
- [19] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik, and H. Pfister. Video face replacement. *Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 30(6):130:1–10, 2011. [3](#)
- [20] C. Ding, L. Xie, and P. Zhu. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications*, 74(22):9871–9888, 2015. [3](#)
- [21] P. Edwards, C. Landreth, E. Fiume, and K. Singh. JALI: An animator-centric viseme model for expressive lip synchronization. *Transactions on Graphics (Proc. SIGGRAPH)*, 35(4):127:1–127:11, 2016. [3](#)
- [22] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *Transactions on Graphics*, 37(4):112:1–112:11, 2018. [2](#)
- [23] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. *Transactions on Graphics (Proc. SIGGRAPH)*, 21(3):388–398, 2002. [2](#)
- [24] B. Fan, L. Wang, F. K. Soong, and L. Xie. Photo-real talking head with deep bidirectional LSTM. In *International Conference on Acoustics, Speech and Signal Processing*, pages 4884–4888, 2015. [3](#)
- [25] B. Fan, L. Xie, S. Yang, L. Wang, and F. K. Soong. A deep bidirectional LSTM approach for video-realistic talking head. *Multimedia Tools and Applications*, 75(9):5287–5309, 2016. [3](#)
- [26] Gabrielle Fanelli, Jürgen Gall, Harald Romsdorfer, Thibaut Weise, and Luc van Gool. A 3D audio-visual corpus of affective communication. *IEEE MultiMedia*, 12(6):591 – 598, 2010. [3](#)
- [27] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The DARPA speech recognition research database: Specifications and status. In *DARPA Speech Recognition Workshop*, 1986. [5](#)
- [28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993. [5](#)
- [29] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. [2](#), [3](#), [4](#)
- [30] H. Hermansky and N. Morgan. RASTA processing of speech. *Transactions on Speech and Audio Processing*, 2(4):578–589, 1994. [2](#)
- [31] P. Hong, Z. Wen, and T. S. Huang. Real-time speech-driven face animation with expressions using neural networks. *Transactions on Neural Networks*, 13(4):916–927, 2002. [3](#)

- [32] P. Kakumanu, R. Gutierrez-Osuna, A. Esposito, R. Bryll, A. Goshtasby, and O. Garcia. Speech driven facial animation. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5, 2001. 3
- [33] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *Transactions on Graphics (Proc. SIGGRAPH)*, 36(4):94, 2017. 1, 2, 3, 5, 7, 8
- [34] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [35] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen. Production-level facial performance capture using deep convolutional neural networks. In *SIGGRAPH / Eurographics Symposium on Computer Animation*, pages 10:1–10:10, 2017. 3
- [36] H. Li, T. Weise, and M. Pauly. Example-based facial rigging. 29(4):32, 2010. 3
- [37] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6), 2017. 2, 3, 4, 5
- [38] Y. Liu, F. Xu, J. Chai, X. Tong, L. Wang, and Q. Huo. Video-audio driven real-time facial animation. *Transactions on Graphics (Proc. SIGGRAPH Asia)*, 34(6):182:1–182:10, 2015. 2, 3
- [39] M. Mori. Bukimi no tani [the uncanny valley]. *Energy*, 7:33–35, 1970. 2
- [40] Mozilla. Project DeepSpeech. <https://github.com/mozilla/DeepSpeech>, 2017. 3, 4, 5
- [41] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Computer Vision and Pattern Recognition*, 2019. 8
- [42] H. Pham and V. Pavlovic. Speech-driven 3D facial animation with implicit emotional awareness: A deep learning approach. In *Conference on Computer Vision and Pattern Recognition Workshop*, 2017. 2
- [43] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. 2016. 5
- [44] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision*, pages 725–741, 2018. 3
- [45] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. HMM-based text-to-audio-visual speech synthesis. In *International Conference on Spoken Language Processing*, pages 25–28, 2000. 3
- [46] G. Salvi, J. Beskow, S. Al Moubayed, and B. Granström. Synface—speech-driven facial animation for virtual speech-reading support. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009(1), 2009. 3
- [47] A. Savran, N. Alyüz, H. Dibeklioglu, O. Celiktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3D face analysis. In *Biometrics and Identity Management*, pages 47–56, 2008. 3
- [48] T. Shimba, R. Sakurai, H. Yamazoe, and J.-H. Lee. Talking heads synthesis from audio with deep neural networks. In *System Integration*, pages 100–105, 2015. 2
- [49] soundible.com. <http://soundible.com/tags-city.html>. 7
- [50] R.W. Sumner and J. Popović. Deformation transfer for triangle meshes. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 23(3):399–405, 2004. 3
- [51] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. *Transactions on Graphics (Proc. SIGGRAPH)*, 36(4):95, 2017. 2
- [52] S. Taylor, A. Kato, B. Milner, and I. Matthews. Audio-to-visual speech conversion using deep neural networks. 2016. 3
- [53] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews. A deep learning approach for generalized speech animation. *Transactions on Graphics*, 36(4):93, 2017. 2, 3
- [54] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews. Dynamic units of visual speech. In *SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284, 2012. 3
- [55] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobald, and M. Nießner. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Computer Vision and Pattern Recognition*, 2016. 3
- [56] A. van den Oord, S. Dieleman, H. Zen, . Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and . Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. 3
- [57] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 24(3):426–433, 2005. 3
- [58] VOCA. <http://voca.is.tue.mpg.de>, 2019. 2, 6, 8
- [59] L. Wang, W. Han, F. Soong, and Q. Huo. Text driven 3d photo-realistic talking head. In *Conference of the International Speech Communication Association, INTERSPEECH*, pages 3307–3308, 2011. 2
- [60] T. Weise, S. Bouaziz, H. Li, and M. Pauly. Real-time performance-based facial animation. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 30(4):77:1–77:10, 2011. 3
- [61] C. Wu, D. Bradley, M. Gross, and T. Beeler. An anatomically-constrained local deformation model for monocular face capture. *Transactions on Graphics (Proc. SIGGRAPH)*, 2016. 3
- [62] L. Xie and Z.-Q. Liu. Realistic mouth-synching for speech-driven talking face using articulatory modelling. *Transactions on Multimedia*, 9(3):500–510, 2007. 2
- [63] F. Yang, L. Bourdev, J. Wang, E. Shechtman, and D. Metaxas. Facial expression editing in video using a temporally-smooth factorization. In *Conference on Computer Vision and Pattern Recognition*, pages 861–868, 2012. 3
- [64] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3D dynamic facial expression database. In *International Conference on Automatic Face and Gesture Recognition*, pages 1–6, 2008. 3

- [65] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006. 3
- [66] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006. 7
- [67] X. Zhang, L. Wang, G. Li, F. Seide, and F. K. Soong. A new language independent, photo-realistic talking head driven by voice only. In *INTERSPEECH*, pages 2743–2747, 2013. 2
- [68] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692 – 706, 2014. 3
- [69] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, pages 3438–3446, 2016. 3
- [70] Y. Zhou, Y. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh. Visemenet: Audio-driven animator-centric speech animation. *Transactions on Graphics*, 37(4):161:1–161:10, 2018. 3