Cross-Modal Perceptionist: Can Face Geometry be Gleaned from Voices?

Cho-Ying Wu, Chin-Cheng Hsu, Ulrich Neumann
University of Southern California
{choyingw, chincheh, uneumann}@usc.edu

Abstract

This work digs into a root question in human perception: can face geometry be gleaned from one's voices? Previous works that study this question only adopt developments in image synthesis and convert voices into face images to show correlations, but working on the image domain unavoidably involves predicting attributes that voices cannot hint, including facial textures, hairstyles, and backgrounds. We instead investigate the ability to reconstruct 3D faces to concentrate on only geometry, which is much more physiologically grounded. We propose our analysis framework, Cross-Modal Perceptionist, under both supervised and unsupervised learning. First, we construct a dataset, Voxceleb-3D, which extends Voxceleb and includes paired voices and face meshes, making supervised learning possible. Second, we use a knowledge distillation mechanism to study whether face geometry can still be gleaned from voices without paired voices and 3D face data under limited availability of 3D face scans. We break down the core question into four parts and perform visual and numerical analyses as responses to the core question. Our findings echo those in physiology and neuroscience about the correlation between voices and facial structures. The work provides future human-centric cross-modal learning with explainable foundations. See our project page.

1. Introduction

This work studies to what extent voice can hint face geometry motivated by recent studies on voice-face matching and cross-modal learning [29, 53, 60]. Many physiological attributes are embedded in voices. For example, speech is produced by articulatory structures, such as vocal folds, facial muscles, and facial skeletons, which are all densely connected. Such a fact intuitively indicates potential correlations between voices and face shapes [19]. Experiments in cognitive science point out that audio cues are associated with visual cues in human perception– especially in recognizing a person's identity [4]. Recent neuroscience research further shows that two parallel processing of low-level auditory and visual cues are integrated in the cortex, where voice processing affects facial structural analysis for the perception purpose [58].

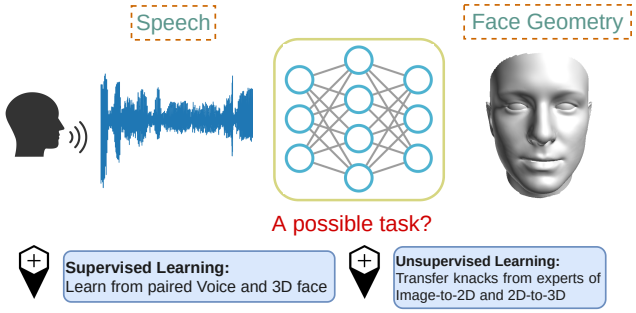

Figure 1. **Cross-Modal Perceptionist.** We study the correlations between voices and face geometry under both supervised and unsupervised learning settings. This work targets at more explainable human-centric cross-modal learning for biometric applications.

Traditional research in the voice domain focuses on utilizing voice inputs for predicting more conspicuous attributes which include speaker identity [6, 28, 42], age [15, 41, 47], gender [27], and emotion [52, 59]. A novel direction in recent development goes beyond predicting these attributes and tries to reconstruct *2D face images from voice* [8, 35, 54]. Their research is built on an observation that one can approximately envision how an unknown speaker looks when listening to the speaker's voice. Attempts towards validating this assumptive observation include the work [35] for image reconstruction and works [8, 54] using generative adversarial networks (GANs). They aim to output face images from only a speaker's voice.

However, face images from voices are inherently ill-posed: the task involves predicting extraneous attributes that voices cannot hint, including image backgrounds, hairstyles, headgears, or beards. These attributes are apparently that one can choose without changing voices. Similar concerns arise regarding the correlations between voices and facial textures or ethnicity. [35] demonstrates a t-SNE plot in which ethnicity is scattered across all samples, indicating its low correlations to voices. As a result, quantifying the differences between an output face image and a

reference is hard and less grounded.

Instead of producing face images, our analysis moves to the 3D domain with mesh representations and **predicts one's face geometry or skull structures from voices**, which is free from the above issues. Working on 3D meshes is less ambiguous than images because the former includes less noisy variations unrelated to a speaker's voice, such as stylistic variations, hairstyles, background, and facial textures. Moreover, meshes enable more straightforward quantification of differences between prediction and groundtruth in the Euclidean space– unlike the case in using face images, where sources of differences involve backgrounds and hairstyles.

From the perspective of 3D faces, much research attention has been paid to 3D reconstruction from monocular images [16, 46, 56, 62] or video sequences [12, 25] for 3D face animation or talking face synthesis. In contrast, we are the first to investigate the correlations between one's 3D face geometry and voices, and we focus on the analysis of the face geometry gleaned from one's voices. Our goal is to validate the correlations between voices and face geometry towards more explainable human-centric cross-modal learning with neuroscience support.

The analysis inevitably involves acquiring large-scale 3D face scans with paired voices, which is expensive and subject to privacy. To deal with this issue, we propose a novel *Voxceleb-3D* dataset that includes paired voices and 3D face models. Voxceleb-3D is inherited from two widely used datasets: Voxceleb [30]) and VGGFace [37], which include voice and face images of celebrities, respectively. The approach [63] we adopt to create Voxceleb-3D is inspired by 300W-LP-3D [62], the most-used 3D face dataset, and we will describe details in Sec.3.2.

Our analysis framework **Cross-Modal Perceptionist** (CMP), investigates the feasibility to predict face meshes using 3D Morphable Models (3DMM, Sec.3.1) from voices on the following two scenarios (Fig. 1). We first train neural networks directly from Voxceleb-3D in a *supervised learning* manner using the paired voices and 3DMM parameters (Sec.3.2). We further investigate an *unsupervised learning* setting to inspect whether face geometry can still be gleaned without paired voices and 3D faces, which is a more realistic scenario. In this case, we use *knowledge distillation* (KD) [20] to transfer knacks from the state-of-the-art method for 3D faces from images, SynergyNet [56], into our student network and jointly train speech-to-image and image-to-3D blocks (Sec.3.3).

We design a set of metrics to measure the geometric fitness based on points, lines, and regions for both the supervised and the unsupervised scenarios. The evaluation attempts to show correlations between 3D faces and voices with straightforward neural network-based approaches. The analysis with CMP enables us to comprehend the correlations between face geometry and voices. Our research lays explainable foundations for human-centric cross-modal learning and biometric applications using voice-face correlations, such as security and surveillance when only voice is given.

Our goal is not to recover high-quality 3D face meshes from voices comparable to synthesis from visual modalities such as image or video inputs, but we try to answer the core question under our CMP framework: can face geometry be gleaned from voice? We break down the question into four parts and will answer them through experiments.

Q1. Is it feasible to predict visually reasonable face meshes from voice?

Q2. How stable is the mesh prediction from different utterances of the same person?

Q3. Compared with face meshes produced by cascading separately trained speech-to-image and image-to-3D-face methods, can the performance of a joint training flow, where mesh prediction is trained with voice information, improve? How much?

Q4. What is the major improvement that voice information can bring in the joint training flow?

Our contributions are summarized.

1. Towards explainable human-centric cross-modal learning, we are the first to study the correlations between face geometry and voices.
2. We devise an analysis framework, Cross-Modal Perceptionist, which studies both supervised and unsupervised approaches to learn face meshes from voices.
3. We show extensive analysis and discussion and answer to four breakdown questions to validate the correlations between voices and face shapes

2. Related Work

2.1. Audio: Learning Personal Traits from Voice

The human voice is embedded with a wide range of personal information and has long been exploited for recognizing personal traits, such as speaker identity [6, 28, 42], age [15, 41, 47], gender [27], and emotion status [52, 59]. Voices can also be used to monitor health conditions [3] or applied to other medical applications [18]. Most existing works focus on predicting personal traits that are more intuitively related to voice. Our work can be seen as a much more challenging task for learning implicit personal faces or skull structures from voices.

2.2. Visual: 2D/3D Face Synthesis

Face-related synthesis has been under much research in the past years. Generating 2D face images using GANs [1, 13, 22, 23, 32, 43] has been a prevalent task, and recent progress includes more realistic synthesis with diverse

styles. The task of face reenactment [11, 34, 50] focuses on transferring facial features from a source to a target. Some works focus on the 3D domain: synthesizing 3D face models from monocular images [16, 51, 56, 62], synthesizing 3D face motion from videos [12, 25] using 3DMM [10, 49], or implicit fields [57].

2.3. Audio-Visual Learning

Cross-Modal Face Matching [24, 29, 33, 53, 60] covers tasks where voices are used as queries to retrieve faces or vice versa. These tasks are inherently *selection* problems in which the best fit of a voice-face pair from the dataset is desired. Another similar task is cross-modal verification [31, 44, 48] that tells whether input faces and voices belong to the same person, which is a simply *classification* problem for paired inputs. Our work solves its root question and *explains* the success in voice-face matching or verification by verifying correlations between voices and face geometry.

Talking face synthesis targets at generating coherent and natural lip movements. Some works drive template images [17, 21, 61] or template face meshes [9] to talk by speech inputs. Some replace lip movements in a video with movements inferred from another video or speech [7, 55]. Their focuses are coherent lip movements and thus are different from our target at studying holistic facial structures.

Voice to Face is the closest task to our work. This task is introduced recently to synthesize face images from only voice inputs. [54] and [8] adopt GANs to generate face images from audio clips. [35] uses an encoder-decoder structure to reconstruct face images. However, the disadvantages are that 2D representations contain many variations, such as hairstyles, beards, backgrounds, and facial textures irrelevant to facial geometry, or the correlations lack physiological support. Besides, face reconstruction errors can be ambiguous because two images of the same person can contain different hairstyles and backgrounds.

Our analysis framework circumvents the issues raised by 2D face representations. 3D face models do not contain hairstyles, backgrounds, or texture variations. Geometric representation of meshes enables us to analyze the correlations between voices and 3D shapes and further directly measure gains and errors in the Euclidean space. In this way, we can focus on face geometry gleaned from voices.

3. Method

Our goal is to analyze how a person's voice relates to one's face geometry in the 3D space. Thus, we learn 3D face meshes using 3D Morphable Models (3DMM) from input speech and analyze the correlations under supervised and unsupervised learning settings. The supervised setting learns the correlation from a paired voice and 3D face dataset. The unsupervised learning studies a realistic case when such paired dataset is not available, is it still possible to predict face geometry from voice?

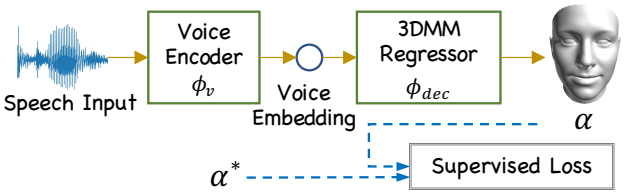

Figure 2. **Supervised learning framework.** Given a speech input, voice embedding is extracted by ϕ_v. ϕ_{dec} then estimates 3DMM parameters α for 3D face modeling. The supervision is computed with groundtruth α^*.

3.1. 3D Morphable Models (3DMM)

3DMM [10] is a popular method for 3D face modeling using principal component analysis (PCA). By estimating the weights of basis matrices, a 3D face can be constructed. We can decompose a face into two components: the average face and face shape variation. That is, for a face A,

$$A = \bar{A} + V\alpha, \tag{1}$$

where $\bar{A} \in \mathbb{R}^{3N}$ is the average face with N three-dimensional vertices, $V \in \mathbb{R}^{3N \times P}$ is a basis matrix for the face shape variation, $\alpha \in \mathbb{R}^P$ is the coefficients. Note that we can reshape A into $A_r \in \mathbb{R}^{3 \times N}$, a matrix representation suitable for 3D rotation and translation.

We set $N = 53490$ vertices following BFM [40], a particular form of 3DMM. Per the dimensionality of shape variation basis, we choose $P = 50$ following SynergyNet [56], the state-of-the-art 3D face reconstruction methods from *images* using BFM. There are 12 additional pose parameters in SynergyNet used to align reconstructed 3D faces to its 2D image inputs: a rotation matrix $R \in \mathbb{R}^{3 \times 3}$ and a translation vector $t \in \mathbb{R}^3$, i.e., $A_p = RA_r + t$. In our analysis, we only use these pose parameters for visualizing how well a predicted face mesh fits a 2D shape outline.

3.2. Supervised Learning with Voice/Mesh Pairs

We first describe the supervised learning setting, illustrated in Fig. 2. Given a paired speech sequence and 3DMM parameters for an identity, we build an encoder-decoder structure first to extract voice embedding $v \in \mathbb{R}^{64}$ from a mel-spectrogram [14], which is a commonly used time-frequency representation for speech, of the input speech. Following [54], the voice encoder ϕ_v is pretrained on the large-scale speaker recognition task. Then, we train a decoder ϕ_{dec} to estimate 3DMM parameters, α. We use groundtruth 3DMM parameters to supervise the training with L_2 loss.

$$\mathcal{L}_{reg} = \|\alpha - \alpha^*\|^2 \tag{2}$$

where α^* is groundtruth 3DMM parameters.

In addition, we adopt the triplet loss on the estimated 3DMM parameters α. The triplet loss minimizes the difference of pairwise relations between (anchor, positive) and (anchor, negative) pairs with a soft margin.

$$\mathcal{L}_{tri} = max\{\|\alpha - \alpha_p\|_2 - \|\alpha - \alpha_n\|_2 + 1, 0\}, \quad (3)$$

where α plays as an anchor, α_p is a positive sample for the anchor, representing the same identity but regressed from different images, and α_n, coming from a different identity, is a negative sample for the anchor. The triplet loss aims at coherent 3DMM parameters for the anchor and positive samples due to the same identities and simultaneously contrasting to the negative sample due to a different identity. The overall loss function is $\mathcal{L}_{sup} = \mathcal{L}_{reg} + \mathcal{L}_{tri}$.

The challenge of this supervised learning problem is how to obtain α^*. Most large voice datasets, such as Voxceleb [30], only contain speech for celebrities, and most large face datasets, such as VGGFace [37], only consist of publicly scraped face images. We first follow [54] to fetch the intersection of voice and image data from Voxceleb and VGGFace. Then, we propose to fit 3D faces from 2D to create a novel dataset, **Voxceleb-3D**, using an optimization-based approach adopted by 300W-LP-3D [62], the most-used 3D face dataset. In detail, we use an off-the-shelf 3D landmark detector [5] to extract facial landmarks from collected face images and then optimize 3DMM parameters to fit in the extracted landmarks. Our Voxceleb-3D contains paired voice and 3D face data to fulfill our supervised learning.

3.3. Unsupervised Learning with KD

Obtaining real 3D face scans is very expensive and limited by privacy, and the workaround of optimization-based 3DMM fitting with facial landmarks is time-consuming. An unsupervised framework may serve real-world scenarios. As a result, we propose an unsupervised framework with knowledge distillation. By leveraging a well-pretrained expert, it helps to validate whether face geometry can still be gleaned with neither real 3D face scans nor optimized 3DMM parameters.

Our unsupervised framework, illustrated in Fig. 3, has two stages: (1) synthesizing 2D face images from voices with GAN and (2) 3D face modeling from synthesized face images. The motivation is that we first use the GAN to generate 2D faces from voices to obtain the speaker's appearance. However, 2D images contain variations of backgrounds, textures, hairstyles that are irrelevant to voice. Thus, the second-stage image-to-3D-face module disentangles geometry from other variations.

Synthesizing face images from voices with GANs. Previous research develops a GAN-based speech-to-image framework [54]. A voice encoder ϕ_v extracts voice embeddings from input speech. Then a generator ϕ_g synthesizes face images from the voice embeddings, and a discriminator ϕ_{dis} decides whether the synthesis is indistinguishable

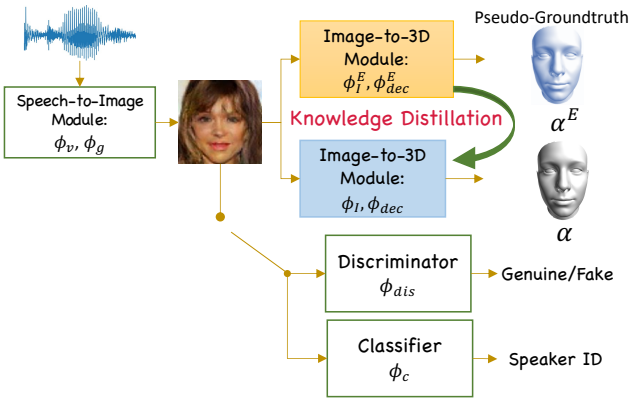

Figure 3. **Unsupervised learning with KD.** The unsupervised framework contains a GAN for face image synthesis with voice encoder ϕ_v, generator ϕ_g, discriminator ϕ_{dis}, and classifier ϕ_c. Then, knowledge distillation is used to achieve unsupervised learning, where information of image-to-3D-face mapping distilled from the expert network (yellow block) is exploited to train the student network (blue block). 2D face is a latent representation in this fashion. Beside using pseudo-groundtruth α^E to train the student, we also distill knowledge at intermediate layers using conditional probability distributions.

from a real face image. Last, a face classifier ϕ_c learns to predict the identity of an incoming face, ensuring that the generator produces face images that are truly close to the identity in interest. Here we overload notations of ϕ_v and other components introduced later for 3D face modeling in both Sec.3.2 and 3.3 due to the same functionalities.

In detail, given a speech input S, its corresponding speaker ID id, and real face images I_r for the speaker, the image synthesized from the generator is $I_f = \phi_g(\phi_v(S))$. The loss formulation is divided into two parts: real and fake images. For real images, the discriminator learns to assign them to "real" (r) and the classifier learns to assign them to id. The loss for real images is $\mathcal{L}_r = \mathcal{L}_d(\phi_{dis}(I_r), r) + \mathcal{L}_c(\phi_c(I_r), id)$ showing the discriminator and classifier losses respectively. For fake images, after producing I_f from ϕ_g, the discriminator learns to assign them to "fake" ($\bar{r}$) and the classifier also learns to assign them to id. The loss counterpart for fake images is $\mathcal{L}_f = \mathcal{L}_d(\phi_{dis}(I_f), \bar{r}) + \mathcal{L}_c(\phi_c(I_f), id)$.

3D face modeling from synthesized images. After image synthesis by GAN, we build a network to estimate 3DMM parameters from fake images. The parameter estimation consists of an encoder ϕ_I and an decoder ϕ_{dec} to obtain 3DMM parameters $\alpha = \phi_{dec}(\phi_I(I_f))$. 3D face meshes are then reconstructed by Eq. 1.

Knowledge distillation for unsupervised learning

To fulfill the unsupervised training, we distill the knowledge of image-to-3D-face reconstruction from a pretrained expert network. The expert, consisting of encoder ϕ_I^E and decoder ϕ_{dec}^E, reconstructs 3D face models from syn-

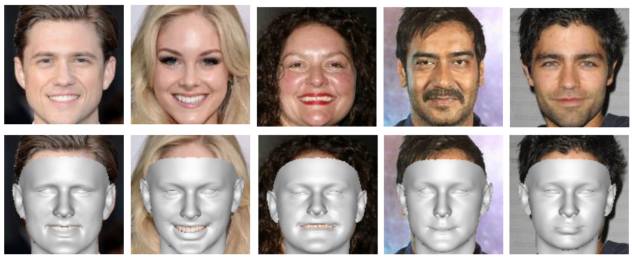

Figure 4. **Samples of face meshes in Voxceleb-3D.** We overlay the 3D faces with associated images to show how well 3D meshes fit in 2D face outlines.

thesized face images and produces pseudo-groundtruth of 3DMM parameters α^E. α^E is used to train the student network by L_2 loss:

$$\mathcal{L}_{p-gt} = \|\alpha^E - \alpha\|^2. \tag{4}$$

This KD strategy *circumvents the needs of paired voice and 3D face data* and helps us achieve unsupervised learning.

In addition to pseudo-groundtruth, we also distill knowledge at intermediate layers and minimize their distribution divergence between the expert and the student. We measure the distributions in the feature spaces by the extracted image embedding $z^E \in \mathbb{R}^{B\times\nu}$ and $z \in \mathbb{R}^{B\times\nu}$ of the expert and the student network. We maintain the batch dimension B and collapse the rest to ν. Then as in [38], we calculate the conditional probability z between feature points as follows.

$$z_{i|j} = \frac{K(z_i, z_j)}{\sum_{k,k\neq j} K(z_k, z_j)}, z^E_{i|j} = \frac{K(z^E_i, z^E_j)}{\sum_{k,k\neq j} K(z^E_k, z^E_j)}, \tag{5}$$

where $K(\cdot,\cdot)$ is scaled and shifted cosine similarity whose outputs lie in [0,1]. Kullback-Leibler (KL) divergence is then used to minimize the two conditional distributions.

$$\mathcal{L}_{div} = \sum_i \sum_{j\neq i} z^E_{j|i} \log\left(\frac{z^E_{j|i}}{z_{j|i}}\right). \tag{6}$$

The KD loss is $\mathcal{L}_{KD} = \mathcal{L}_{p-gt} + \mathcal{L}_{div}$. The overall unsupervised learning loss is combined with GAN loss and also triplet loss in Eq.3.

$$\mathcal{L}_{unsuper} = \mathcal{L}_f + \mathcal{L}_r + \mathcal{L}_{KD} + \mathcal{L}_{tri}. \tag{7}$$

4. Experiments and Results

Datasets. We use our created Voxceleb-3D dataset described in Sec. 3.2. There are about 150K utterances and 140K frontal face images from 1225 subjects. The train/test split for Voxceleb-3D is the same as [54]: Names starting with A-E are used for testing, and the others are for training. We manually pick the best-fit 3D face models for each identity as reference models for evaluations. We display samples of face meshes in Fig. 4.

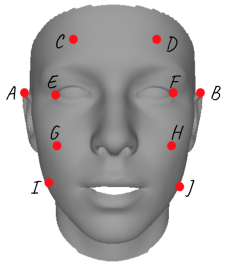

Figure 5. **Distance illustration for our ARE metric.** $\overline{AB}$: ear-to-ear distance. $\overline{CD}$: forehead width. $\overline{EF}$: outer-interocular distance. $\overline{GH}$: midline distance. $\overline{IJ}$: cheek-to-cheek distance.

Data Processing and Training. We follow [54] and extract 64-dimensional log mel-spectrograms with a window size of 25 ms, and perform normalization by mean and variance of each frequency bin for each utterance. In the unsupervised setting, we adopt SynergyNet [56] as the expert. Face images from the generator are 64×64, and we bilinearly upsample them to 120 ×120 to fit the input size of the expert for 3D face reconstruction from images. Our framework is implemented in PyTorch [39]. We use Adam optimizer [26] and set the learning rate to 2×10^{-4}, batch size to 64, and a total number of training steps to 50,000, which consumes about 16 hours to train on a machine with a GeForce RTX 2080 GPU.

To train with triplet loss, for each sample in a batch, we further uniformly sample one utterance of the same person as the positive sample and sample the other one of the different person as the negative sample. We illustrate the network architectures in the supplementary

Metrics. We design several metrics to evaluate 3D face deformation based on α. Here we introduce a line-based metric, ARE, and present point-based and region-based metrics using iterative closet point registration and facial landmarks in the supplementary.

Absolute Ratio Error (**ARE**, line-based): Distances between facial points are commonly used as measures related to aesthetics or surgical purposes [2, 36, 45]. We pick point pairs (shown in Fig. 5) that are most representative for evaluation and calculate the distance ratios to outer-interocular distance (OICD). For example, ear ratio (ER) is $\overline{AB}/\overline{EF}$, and the same for forehead ratio (FR), midline ratio (MR), and cheek ratio (CR). We evaluate our models by the absolute ratio error (ARE) between the predicted and the reference face meshes because these ratios can capture face deformation. As an example, ARE of ER is $|\text{ER} - \text{ER}^*|$, where * denotes the ratios of reference models.

Baseline. We build a straightforward baseline by directly cascading two separately pretrained methods without joint training: the GAN-based speech-to-image block [54] and SynergyNet [56] for image-to-3D-face block (illustrated in Fig. 6) to produce 3D meshes from voices as the baseline framework. In addition, 3DDFA-V2 [16] is another method

Figure 6. **Baseline framework.** The baseline is a direct cascade of two pre-trained state-of-the-art modules: speech-to-image [54] and image-to-3D-face modeling [16, 56].

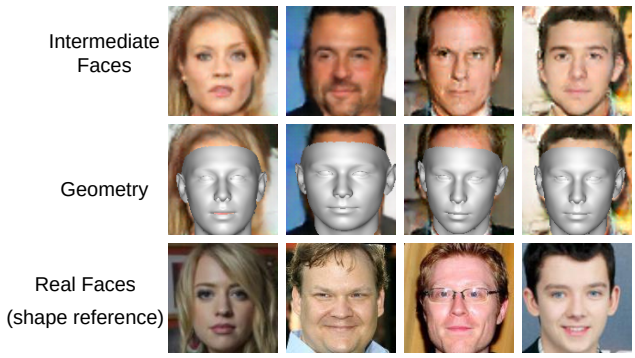

Figure 7. **Evidence for positive response to Q1**. Our unsupervised framework predicts intermediate 2D images and 3D meshes. This answers to Q1 that 3D face models exhibiting similar *face shapes* to the references can be predicted from only voice inputs.

for 3D face modeling from monocular images using BFM and holds a close performance to SynergyNet. Thus, we experiment with combinations of speech-to-image block + 3DDFA-V2 (Base-1) and speech-to-image block + SynergyNet (Base-2). Aside from network-based approaches, we also devise simple oracles that use mean shapes of labels, such as male/female, as predictions, and provide the results as references in the supplementary.

4.1. Analysis

We attempt to answer Q1-Q4 raised in Sec.1 in this section and respond to each respective question in A1-A4. In A1-A3, we show predictions using our unsupervised learning setting since the by-product intermediate images help explain 3D mesh prediction for better comprehension of the mechanism. We show visuals from the supervised version in the supplementary.

A1: Meshes and intermediate images. In Fig. 7, we display intermediate 2D images, 3D meshes, and real faces. Note that the real faces should only be treated for identification purposes in terms of **face shapes** because those images include backgrounds or hairstyle variations that differ from references. Our end targets are the 3D face meshes that are free from these factors. Prediction from our framework generates wider meshes in Column 2 and thinner meshes in Column 3 and 4, which reflect the real face wideness. All the generated 3D meshes fit in 2D facial outlines well.

These results exemplify the ability to convert voices into plausible 3D face meshes. Although meshes are rough compared with 3D synthesis from images or videos modalities, the results conform to our intuitions that when an unheard speech comes, one can roughly envision whether the speaker's face is overall wider or thinner. However, we cannot picture subtle details, such as bumps or wrinkles on faces. The same trends can be observed in a vast result collection in Fig. 8. The results are not cherry-picked.

A2: Prediction coherence of the same speaker. To address Q2, we showcase in Fig. 9 and 10 for the coherence of the predicted face shapes from different utterances of the same speakers. The 2D predictions exhibit *face shape and outline consistency*, though they are still plagued by stylistic variations that are geometrically unrelated to our task. This not only confirms the ability to produce coherent face meshes but also underlines why predicting face meshes from voices is regarded as less noisy than face image synthesis.

A3: Gain from cross-modal joint training. For Q3, we compare results from our unsupervised framework against those from the baseline in Fig. 11. Joint training for the speech-to-image and image-to-3D sub-networks attain higher and more stable image synthesis quality, which benefits 3D mesh prediction. In contrast, those from the baseline (Base-2) include more artifacts. This justifies our CMP's cross-modal joint training strategy, which lets networks learn to predict 3D faces with voice input at the training, improves over the baseline that is separately trained.

To this end, we understand that voices can help 3D face prediction and produce visually reasonable meshes that are close to real face shapes.

A3-Quantification+A4. We numerically compare supervised and unsupervised settings of our analysis framework, CMP, against the baseline (Fig. 6) using the ARE proposed in Sec.4. Both supervised and unsupervised settings improve the line-based ARE over the baseline around *20%*, as exhibited in Table 1. The results show that cross-modal joint training achieves better results than the direct cascade of pretrained blocks. These improvements reveal underlying correlations between voices and face shapes such that training face mesh prediction with joined voice information is helpful. Among all metrics, ear ratio (ER) has the most prominent improvements, indicating that the best indicative attribute voice can hint is the head width, and thus it answers Q4. This analysis aligns with the findings in Sec.4.1 that voice can indicate wider/thinner faces, which corresponds to our intuition that we can roughly envision a speaker's face width from voices. Through this study, we quantify the improvements of cross-modal learning from voice inputs, and the findings echo human perception intuitively.

4.2. Subjective Evaluations

We further conduct subjective preference tests over the outputs to quantify the difference of preference. The test was divided into three sections, considering *images*, *3D*

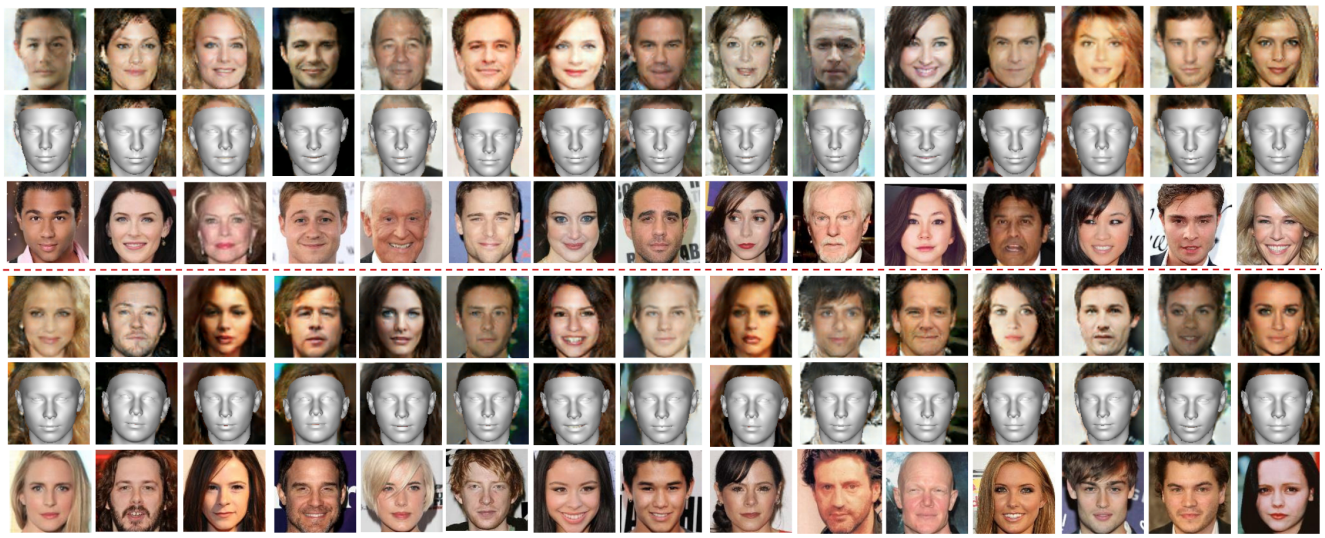

Figure 8. **A collection of results supports our positive response to Q1.** This figure extends Fig.7. Top to down for two row chunks: predicted intermediate face images, predicted 3D models, real faces for references.

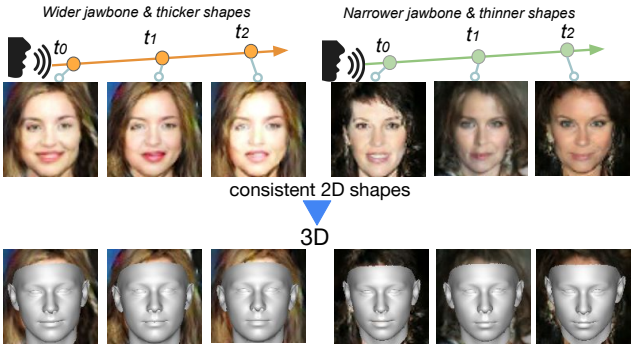

Figure 9. **Illustration for our positive response to Q2.** Consistent intermediate images and 3D faces can be predicted from the same speaker with different time-step utterances.

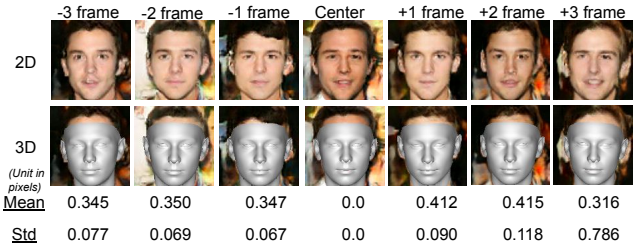

Figure 10. **Shape variation statistics in response to Q2.** Mean and std of per-vertex variation w.r.t. the center frame are shown, calculated in frontal pose. 3D shapes recovered from different utterances are consistent with only sub-pixel differences.

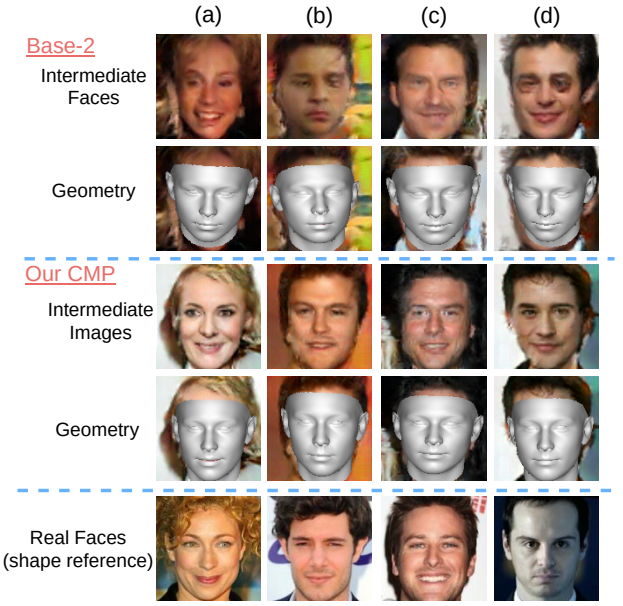

Figure 11. **Comparison of intermediate images and meshes in response to Q3.** The cross-modal joint training strategy in our unsupervised CMP produces better-quality images than the baseline. More reliable images as latent representations from our CMP can facilitate the mesh prediction. We include real faces for *face shape* references.

models, and *joint materials*. Though we favor face meshes over images because the former are free from irrelevant textures or backgrounds, we included intermediate images from our unsupervised setting in the test and asked subjects to focus on face shapes since better-outlined shapes on images lead to better-shaped meshes, as indicated in Fig. 11.

Evaluation design. Thirty questions were included in the test, and 154 subjects with no prior knowledge of our work were invited to the test. In the first section, each of the ten questions consisted of three images– a reference face image, a face image from our unsupervised CMP, and a face image generated from the Base-2 ([54]+ [56]). The order of the generated images was randomized. The subjects were asked to select the face image "whose shape is geometrically more similar to the reference face?". In the

Table 1. **ARE metric study.** Compared with baseline in Fig. 6, results from CMP show that cross-modal joint training with voice input can obtain around 20% improvements. We also highlight the largest improvement, ER, that answers to Q4.

ARE	Base-1	Base-2	CMP-supervised	CMP-unsupervised
ER	0.0319	0.0311	**0.0152**	**0.0181**
FR	0.0184	0.0173	0.0186	0.0169
MR	0.0177	0.0173	0.0169	0.0174
CR	0.0562	0.0551	0.0457	0.0480
Mean	0.0311	0.0302	**0.0241**	**0.0251**
Gain	-	0%	**-20.2%**	**-16.9%**

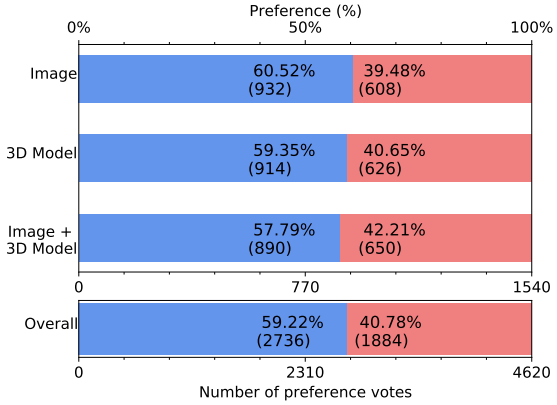

	Image	3D Model	Image + 3D Model	Overall
p-value	$\sim 10^{-16}$	$\sim 10^{-14}$	$\sim 10^{-10}$	$\sim 10^{-16}$

Figure 12. **Results of subjective preference tests.** The blue bars are the preference for our method, while the red bars are the preference for the baseline method. The percentages are labeled on the bar, and the total number of votes is enclosed in the parentheses. The x-axis on the bottom labels the total number of responses, and that on the top denotes the percentage. The p-values of the statistical significance tests are provided under the bar. $\sim$ shows the value's order of magnitude.

second section (10 questions), a similar design was laid out, but 3D face models from Base-2 and our CMP were used instead of images. Finally, in the third section (10 questions), each of the two options comprised a face image and a 3D face model; the subject was asked to jointly consider over the two materials: "overall, whose shape geometrically fits the given reference image better?"

Statistical significance test. Fig. 12 summarizes our subjective evaluation. We conduct a statistical significance test with the following formulation. A subject's response to a question is considered as a Bernoulli random variable with a parameter p. The null hypothesis ($\mathcal{H}_0$) assumes $p \leq 0.5$, meaning that the subjects do not prefer our model. The alternative hypothesis $\mathcal{H}_1$ assumes $p > 0.5$, meaning that the subjects prefer our model. For each section, there are 154 subjects and ten responses per subject. For a significance level $\gamma = 0.001$, let $b_{n,p}(\gamma)$ denote the quantile of order γ for the binomial distribution with parameters p and n. We can decide whether the subjects prefer our model by

$$\begin{aligned} &\text{Reject } \mathcal{H}_0 \text{ versus } \mathcal{H}_1 \Leftrightarrow np \geq b_{n,p}(1-\gamma) \\ &\mathcal{H}_0 : p \leq 0.5, \mathcal{H}_1 : p > 0.5. \end{aligned} \tag{8}$$

As shown in Fig. 12, np is well above the threshold $b_{n=1540,p=0.5}(1-\gamma) = 831$, rejecting $\mathcal{H}_0$ and suggesting that the subjects significantly prefer our model over the baseline. The single-sided p-values are displayed under the bar chart. A lower p-value means stronger rejection of $\mathcal{H}_0$. The p-values from our tests are much lower than the level 0.001, showing high statistical significance. In conclusion, the hypothesis test verifies that the subjects indeed favor the predictions from our method.

5. Conclusion and Discussion

In this work, we investigate a root question in human perception: can face geometry be gleaned from voices? We first point out shortcomings in previous studies in which 2D faces are predicted: such synthesis contains variations in hairstyles, backgrounds, and facial textures with controversial correlations to voices. We instead focus on 3D faces whose correlations to voices have been supported by neuroscience and cognitive science studies. As a pioneering work toward this direction, we innovate a way to construct Voxceleb-3D that includes paired voices and 3D face models, devise and test baseline methods and oracles, and propose a set of evaluation metrics. Our proposed main framework, CMP, learns 3DMM parameters from voices under both supervised and unsupervised settings. Based on CMP, we answer the core question with a four-part breakdown by detailed analyses and subjective evaluations. We conclude that 3D faces can be roughly reconstructed from voices. Our study is far from complete, but hopefully, it lays a foundation for speech and 3D cross-modal studies in the future.

Ethical statement. There are arguably implicit factors, such as voices after smoking and drinking might be different. The data of Voxceleb contains speech from interviews, where interviewees usually speak in normal voices. More implicit and subtle factors such as drug use or health conditions might affect voices, but it needs clinical studies and should be validated from physiological views. The results shown in this work only aim to point out the correlation between voice and face (skull) structure exist and do not make assumptions on race/ethnic origin, and this work does not indicate the relation between race and voice or race and face structure. As mentioned in Introduction, the correlation between race/ethnicity cannot be easily resolved. Besides, the reconstructed meshes do not contain skin color, facial textures, or hairstyles that can explicitly correspond to one's true identity, and thus anonymity can be preserved.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *CVPR*, pages 4432–4441, 2019. 2

[2] Mohammed Aleem Abdullah. Inner canthal distance and geometric progression as a predictor of maxillary central incisor width. *The Journal of prosthetic dentistry*, 88(1):16–20, 2002. 5

[3] Zulfiqar Ali, Ghulam Muhammad, and Mohammed F Alhamid. An automatic health monitoring system for patients suffering from voice complications in smart cities. *IEEE Access*, 5:3900–3908, 2017. 2

[4] Pascal Belin, Shirley Fecteau, and Catherine Bedard. Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, 8(3):129–135, 2004. 1

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *CVPR*, pages 1021–1030, 2017. 4

[6] Ray Bull, Harriet Rathborn, and Brian R Clifford. The voice-recognition accuracy of blind listeners. *Perception*, 12(2):223–226, 1983. 1, 2

[7] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *ECCV*, pages 520–535, 2018. 3

[8] Hyeong-Seok Choi, Changdae Park, and Kyogu Lee. From inference to generation: End-to-end fully self-supervised generation of human face from speech. *ICLR*, 2020. 1, 3

[9] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, pages 10101–10111, 2019. 3

[10] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 3

[11] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *CVPR*, pages 4217–4224, 2014. 3

[12] 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 Transactions on Graphics (TOG)*, 35(3):1–15, 2016. 2, 3

[13] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *NeurIPS*, 2014. 2

[14] Karlheinz Gröchenig. *Foundations of time-frequency analysis*. Springer Science & Business Media, 2001. 3

[15] Joanna Grzybowska and Stanislaw Kacprzak. Speaker age classification and regression using i-vectors. In *INTERSPEECH*, pages 1402–1406, 2016. 1, 2

[16] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020. 2, 3, 5, 6

[17] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, 2021. 3

[18] Jing Han, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. Exploring automatic covid-19 diagnosis via voice and symptoms from crowdsourced data. In *ICASSP*. IEEE, 2021. 2

[19] Jonathan Harrington. Acoustic phonetics. *The handbook of phonetic sciences*, 2:81–129, 2010. 1

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[21] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision (IJCV)*, 127(11):1767–1779, 2019. 3

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[24] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. In *ACCV*, pages 276–292. Springer, 2018. 3

[25] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2, 3

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[27] Sheng Li, Dabre Raj, Xugang Lu, Peng Shen, Tatsuya Kawahara, and Hisashi Kawai. Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation. In *INTERSPEECH*, pages 4400–4404, 2019. 1, 2

[28] Corrina Maguinness, Claudia Roswandowitz, and Katharina von Kriegstein. Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, 116:179–193, 2018. 1, 2

[29] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*, pages 8427–8436, 2018. 1, 3

[30] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. *INTERSPEECH*, pages 2616–2620, 2017. 2, 4

[31] Shah Nawaz, Muhammad Saad Saeed, Pietro Morerio, Arif Mahmood, Ignazio Gallo, Muhammad Haroon Yousaf, and Alessio Del Bue. Cross-modal speaker verification and recognition: A multilingual perspective. In *CVPRW*, 2021. 3

[32] Weili Nie, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit Patel, and Animashree Anandkumar. Semi-supervised stylegan for disentanglement learning. In *ICML*, 2020. 2

[33] Hailong Ning, Xiangtao Zheng, Xiaoqiang Lu, and Yuan Yuan. Disentangled representation learning for cross-modal biometric matching. *TMM*, 2021. 3

[34] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *CVPR*, pages 7184–7193, 2019. 3

[35] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *CVPR*, pages 7539–7548, 2019. 1, 3

[36] Pamela M Pallett, Stephen Link, and Kang Lee. New "golden" ratios for facial beauty. *Vision research*, 50(2):149–154, 2010. 5

[37] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015. 2, 4

[38] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2020. 5

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 5

[40] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. IEEE, 2009. 3

[41] Paul H Ptacek and Eric K Sander. Age recognition from voice. *Journal of speech and hearing Research*, 9(2):273–277, 1966. 1, 2

[42] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE, 2018. 1, 2

[43] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 2

[44] Leda Sarı, Kritika Singh, Jiatong Zhou, Lorenzo Torresani, Nayan Singhal, and Yatharth Saraf. A multi-view approach to audio-visual speaker verification. In *ICASSP*, 2021. 3

[45] David Sarver and Ronald S Jacobson. The aesthetic dentofacial analysis. *Clinics in plastic surgery*, 34(3):369–394, 2007. 5

[46] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *ECCV*, 2020. 2

[47] Rita Singh, Joseph Keshet, Deniz Gencaga, and Bhiksha Raj. The relationship of voice onset time and voice offset time to physical age. In *ICASSP*, pages 5390–5394. IEEE, 2016. 1, 2

[48] Ruijie Tao, Rohan Kumar Das, and Haizhou Li. Audio-visual speaker recognition with a cross-modal discriminative network. In *INTERSPEECH*, 2020. 3

[49] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *CVPR*, pages 3361–3371, 2021. 3

[50] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016. 3

[51] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, pages 7346–7355, 2018. 3

[52] Zhong-Qiu Wang and Ivan Tashev. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In *ICASSP*, pages 5150–5154. IEEE, 2017. 1, 2

[53] Peisong Wen, Qianqian Xu, Yangbangyan Jiang, Zhiyong Yang, Yuan He, and Qingming Huang. Seeking the shape of sound: An adaptive framework for learning voice-face association. In *CVPR*, pages 16347–16356, 2021. 1, 3

[54] Yandong Wen, Bhiksha Raj, and Rita Singh. Face reconstruction from voice using generative adversarial networks. In *NeurIPS*, volume 32, 2019. 1, 3, 4, 5, 6, 7

[55] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, pages 670–686, 2018. 3

[56] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. *3DV*, 2021. 2, 3, 5, 6, 7

[57] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *CVPR*, pages 12803–12813, 2021. 3

[58] Andrew W Young, Sascha Frühholz, and Stefan R Schweinberger. Face and voice perception: understanding commonalities and differences. *Trends in cognitive sciences*, 24(5):398–410, 2020. 1

[59] Zixing Zhang, Bingwen Wu, and Björn Schuller. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP*, pages 6705–6709. IEEE, 2019. 1, 2

[60] Aihua Zheng, Menglan Hu, Bo Jiang, Yan Huang, Yan Yan, and Bin Luo. Adversarial-metric learning for audio-visual cross-modal matching. *TMM*, 2021. 1, 3

[61] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, volume 33, pages 9299–9306, 2019. 3

[62] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016. 2, 3, 4

[63] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, June 2015. 2