Rethinking Voice-Face Correlation: A Geometry View

Xiang Li
CMU

Yandong Wen
MPI-IS

Muqiao Yang
CMU

Jinglu Wang
Microsoft

Rita Singh
CMU

Bhiksha Raj
CMU & MBZUAI

ABSTRACT

Previous works on voice-face matching and voice-guided face synthesis demonstrate strong correlations between voice and face, primarily relying on coarse semantic cues such as gender, age, and emotion. In this paper, we aim to investigate the capability of reconstructing the 3D facial shape from voice from a geometry perspective without any semantic information. We propose a voice-anthropometric measurement (AM)-face paradigm, which identifies predictable facial AMs from the voice and uses them to guide 3D face reconstruction. By leveraging AMs as a proxy to link the voice and face geometry, we can eliminate the influence of unpredictable AMs and make the face geometry tractable. Our approach is evaluated on a new dataset with ground-truth 3D face scans and corresponding voice recordings, and we find significant correlations between voice and specific parts of the face geometry, such as the nasal cavity and cranium. Our work offers a new perspective on voice-face correlation and can serve as a good empirical study for anthropometry science. Code: https://github.com/lxa9867/VAF.

CCS CONCEPTS

• **Computing methodologies** → *Appearance and texture representations.*

KEYWORDS

voice, face, vocal tract

ACM Reference Format:
Xiang Li, Yandong Wen, Muqiao Yang, Jinglu Wang, Rita Singh, and Bhiksha Raj. 2023. Rethinking Voice-Face Correlation: A Geometry View. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611779

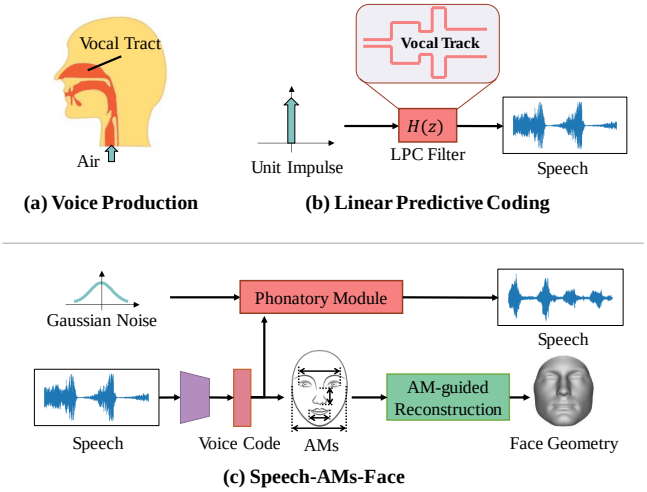


Figure 1: (a) Human voice production. (b) Linear predictive coding represents the voice by a unit impulse with a set of linear filters which can be interpreted as an estimation of the vocal tract. (c) Our voice-AM-Face pipeline first predicts and verifies predictable anthropometric measurements (AMs) and then utilizes AMs to guide 3D face reconstruction. A phonatory module is involved to obtain a better representation for AM prediction.

1 INTRODUCTION

The study of face-voice correlation has been extensively investigated in recent years. Previous works on voice-face matching [28, 45, 54], voice-guided face synthesis [9, 16, 19, 55], and voice-guided face modification have indicated a strong correlation between voice and face. The most intuitive and commonly used consensus encoded between voice and face is mainly based on semantics, such as gender, age and emotion. Most prior works aim to learn a semantic correspondence between voice and face and conduct crossmodal tasks by leveraging those consensuses. For example, for voice-guided face synthesis, the generated faces have reasonable appearances with proper gender, age and emotion status corresponding to the voice. Those semantic correlations are strong and easy to learn thus dominant previous models while a fundamental question we want to cast is, are there any other voice-face correlations except for those coarse semantics? Is reconstructing identity-fidelity 3D face from voice possible? In this paper, we aim to explore the voice-face correlations in a geometry view after constraining all those easily learned semantic biases.

There are several previous works investigating recovering face from voice. Most of them are from a 2D perspective [16, 19, 55], which utilize Generative Adversarial Network (GAN) [14, 26] to generate faces with voice as the condition. However, face recovering from voice is ill-posed. [29] found that the recovery mainly focuses on some semantics of the speaker. For example, attributes

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611779

such as ethnicity have weak or no function while gender and age tend to be recovered. Since those models mainly rely on semantics, the results are not identity-fidelity which means generated faces can look very different from the original ones. In addition, for a 2D face image, identity-unrelated factors like expressions, hairs, glasses, illumination, background, etc., are also involved in the recovery process leading to noisy and unstable outcomes. Different from 2D images, general 3D facial shape is represented by the 3D coordinates of a number of points on its surface called vertices [4] which inherently excludes the identity-unrelated factors. Moreover, since the topology of 3D facial shape is predefined and consistent across different faces, we can easily measure the reconstruction accuracy with distances between the predicted vertices and their ground truths.

Similar to our target, one recent work [48] attempts to recover 3D faces from voice while, due to the lack of ground-truth 3D face scans, they first generate 2D face images from voice and then reconstruct 3D faces guided by an off-the-shelf 3D face reconstruction model. The noise enrolled in the 2D-to-3D face reconstruction makes the result unconvincing. For example, any expression in the 2D face from the first stage will force the reconstructed 3D face to have the same expression. In this way, we consider the face is still determined by the first-stage 2D face image.

In our method, we aim to disable all previously used semantics, e.g., gender, age and emotion, and focus on the voice-face correlation from a pure geometry view. Before introducing our method, let us go back and understand how voice is generated by human beings. voice is produced by phonatory structures (Fig. 1 (a)), e.g., vocal tract and vocal cords. Specifically, when producing vowels, the vocal cords vibrate with no obstruction in the vocal tract. In contrast, for most consonants, the phonation purely depends on the vocal tract resonance with a pulmonic airflow. The vocal tract can be assumed as a filter that makes the phonemes versatile and personalized. With the phonation mechanism of human beings, as shown in Fig. 1 (b), Markel *et al* introduces linear predictive coding (LPC) [25] which models phonation as a unit impulse signal modified by a stack of tubes (vocal tract) and encodes personalized voice by vocal tract coefficients. The LPC yields a good physical model of the vocal tract with only voice inputs in an unsupervised manner. As the mouth and nose serve as the most important parts of the vocal tract, we hypothesize that their geometry should be encoded in the voice. With the tight bind of muscles and skeletons, other parts of face geometry may also be represented by voice.

Though voice and face geometry should have some correlations, we have no idea about which part of the face voice can represent. Constructing uncorrelated relations will lead to random results and raise the model instability. To tackle this problem, we introduce the voice-anthropometric measurement (AM)-face paradigm. Previous studies have shown that anthropometric measurements like the dimensions of nasal cavities [43] or cranium [49, 50] directly influence the speaker's voices. In our voice-AM-face paradigm, we first summarize a set of AMs from anthropometry literature [10, 11, 33, 39, 56], then identify predictable AMs and use them to guide the 3D face reconstruction by conducting AM-guided optimization. By leveraging AMs as a proxy to link the voice and face geometry, we can eliminate the influence of unpredictable AMs and make the face geometry tractable. In addition, the analysis of AMs also brings a new view to understanding voice-face correlation in a fine-grained fashion.

Inspired by LPC which learns the shape of the vocal tract by producing voice, we utilize a phonatory module to facilitate voice representation learning for face geometry. Similar to the auto-regressive impulse-by-filter model used in LPC, recently introduced denoising diffusion probabilistic models [18] share a similar structure, which samples a random noise with auto-regressive updating to form the final result. Based on the structure similarity, we choose the diffusion model as our phonatory module.

With the predicted AMs, we reconstruct the facial shapes by an optimization-based method, which first projects the 3D facial shapes into a low-dimensional linear space [4]. By adjusting the coefficients in low-dimensional space, we obtain different re-projected 3D facial shapes. Though this paper mainly focuses on understanding the relationship between the 3D facial shape and voice from a scientific angle, this technique has its potential applications. For example, the identity-fidelity facial shape can be used for criminal profiling scenarios, such as hoax calls and voice-based phishing.

In this paper, we try to answer two core questions - (1) Is there a correlation between face geometry and voice? (2) If so, which part of the face can be represented by the voice? To fulfill our target, we collect a large-scale dataset containing ground-truth 3D face scans and corresponding voice recordings from 1026 speaker identities. A voice-AM-face paradigm equipped with a phonatory module is proposed for analyzing the voice-face correlation. Our contributions can be summarized as follows.

- We propose a voice-AM-face paradigm and a corresponding voice-face dataset for tractable 3D face deduction from voice.
- We investigate voice-face correlation in a fine-grained manner by statistically verifying which part of the face can be reflected by the voice. The results can serve as a good reference to support future voice-face research, such as voice-face verification.
- We leverage voice production as a proxy task to learn face geometry representation and verify that voice production is highly related to 3D facial shapes.

2 RELATED WORKS

2.1 Voice-face Correlation

The human voice contains rich information that can be used to recognize personality traits, such as speaker identity [6, 24, 34], gender [22], age [15, 30, 40], and emotion status [44, 52]. Voices can also be used for monitoring health conditions [1] and other medical applications [17]. Most existing works in this area focus on predicting personality traits that are intuitively related to voice. Such personality traits may have essential correlations between the human voice and their faces [46].

Cross-modal voice-face matching [28, 45, 54] and cross-modal verification [27, 38, 42] are tasks where voices are used as queries to retrieve faces or vice versa, which have received increasing attention in recent years. Voice-guided face synthesis is another related task, which aims to generate coherent and natural lip movements, and includes methods that drive template images [16, 19, 55] or template face meshes [9] to talk by speech inputs, or replace lip

movements in a video with movements inferred from another video or speech [8, 47].

Unlike the existing work in related fields that are more focused on semantic correlations between voice and face, our work investigates the voice-face correlation from a geometry view by studying holistic facial structures. There has been recent work that seeks to understand the correlations between voice and facial geometry by first recovering 2D faces from voice and then reconstructing 3D faces from the 2D representations [48]. However, during this process, it is still inevitable that the semantic correlations are encoded in the 2D face and affect the 2D-to-3D face reconstruction. Instead, we aim to model our voice-face correlation from a pure geometry view without the influence of any semantics.

2.2 Phonation and Anthropometry

The human voice is generated by phonatory structures, and the phonation of different phonemes may be dependent on different physiological structures. By utilizing such properties, it has been proven to be informative and helpful in various tasks, including automatic speech recognition [12], speech enhancement [51], and emotion recognition [13]. Beyond those language-related usages, human attributes are also predictable from voice. There is a substantial body of research on inferring human attributes from a person's voice, including speaker identity [7, 35], age [3, 31], gender [23], face [32], and emotion status [44, 53].

To explicitly describe the correspondence between vocal and facial features, anthropometric measurements have been used in a wide range of applications to associate with voice production [10, 11, 33, 39, 41, 56]. In a broad sense, AMs may cover various body parameters and characteristics, including skeletal proportions, race, height, body size, etc. These characteristics may influence the phonation of voice by the differences in the placement of the glottis, length of vocal cords, etc.

In this work, we summarize a large set of AMs that is highly associated with voice-face correlation. Meanwhile, we also identify the predictable AMs to guide the 3D facial shape reconstruction. The results can serve as a good reference to support future voice-face research.

3 METHOD

In this section, we first introduce the task formulation and then demonstrate our method in detail.

3.1 Formulation

We aim to reconstruct any speaker's 3D facial shape from their voice recordings. Given a set of paired voice recordings and 3D facial shapes $\{(v_i, f_i)\}$ from different individuals, where v_i is a voice recording spoken by the i-th person and f_i is a 3D facial shape scanned from the speaker of v_i. The goal is to reconstruct the 3D facial shape f of any speaker from their voice recording v. In our method, we introduce anthropometric measurements (AMs) $m = \{m^{(1)}, \cdots, m^{(k)}\}$ computed from f as a proxy, where K is a positive integer and m^k ($k \in [1, K]$) denotes the k-th AM. Accordingly, the overall dataset is denoted as $\mathcal{D} = \{(v_i, f_i, m_i)\}$. To statistically analyze the results, we construct an additional validation set for empirically validating the dependency. Specifically, the dataset $\mathcal{D}$ is split into a training set $\mathcal{D}_t$ for model learning, a validation set $\mathcal{D}_{v1}$ for model selection, a validation set $\mathcal{D}_{v2}$ for AM selection, and an evaluation set $\mathcal{D}_e$ for evaluating the reconstructed 3D facial shapes. All splits have no overlap.

3.2 Pipeline Overview

As shown in Fig. 2, the proposed method has three main components - facial AM prediction, AM-guided reconstruction and an auxiliary phonatory module. On one hand, we predict the AMs that are potentially correlated with voice production from anthropometry literature [10, 11, 33, 39, 56]. An estimator $\mathcal{E}$ is trained with uncertainty learning with a voice code e. On the other hand, inspired by the voice production mechanism, we introduce a phonatory module as a constraint to facilitate the training of AM prediction. In particular, a diffusion-based voice generation module is involved as the phonatory module which aims to imitate the voice identity conditioning on the voice code e. After that, we select the AMs predictable from voice for hypothesis testing. The null hypothesis is made for each AM and states the AM is unpredictable from voice. We can successfully reject the corresponding null hypothesis if any AM estimation is better than chance on a held-out validation set with statistical significance. The final 3D facial shapes can be reconstructed by a fitting process [5] based on the predictable AMs. This is conducted by adjusting a set of coefficients in low-dimensional space, such that the differences between the AMs of the generated 3D facial shape and the predicted AMs are minimized. Intuitively, if there are more predictable AMs spanning different locations of a face, the reconstruction can be more indistinguishable.

3.3 Facial AM Prediction

In this section, we illustrate our method to predict facial AMs from voice.

AM summarization. There is a large body of literature on anthropometry. Extensive studies show that many AMs of human faces can be associated with voice production [10, 11, 33, 39, 56]. We summarize the most commonly used AMs as shown in Fig. 3 (the complete list of AMs is available in the appendix). The chosen AMs are categorized as proportion, angles and distance of a set of face landmarks. Those intra-face features are more robust than 3D coordinate representations as the variations resulting from spatial misalignment are completely eliminated.

Uncertainty-aware AM estimation. The AM prediction is conducted by an estimator trained with an uncertainty-aware scheme. Let $F_k(v; \mathcal{E}_k, \omega_k) : v \mapsto \mathbb{R}$ be an estimator that maps voice recording v into the k-th predicted AMs, where $\mathcal{E}_k$ and ω_k are the learnable parameters. As this is a regression problem, we leverage

$$\{\mathcal{E}_k^*, \omega_k^*\} = \arg\min_{\mathcal{E}_k, \omega_k} \frac{1}{|\mathcal{D}_t|} \sum_{(v, m^{(k)}) \in \mathcal{D}_t} (F_k(v; \mathcal{E}_k, \omega_k) - m^{(k)})^2 \quad (1)$$

as the training objective for the k-th AM. $|\mathcal{D}_t|$ is the number of the triplets (voice, face and AMs) in dataset $\mathcal{D}_t$. By incorporating uncertainty into the estimator learning, the prediction becomes a random variable rather than a single value. We leverage a Gaussian distribution to the prediction. The estimator $F_k(v; \mathcal{E}_k, \omega_k)$ maps v into the mean of the i-th predicted AM. Similarly, we define

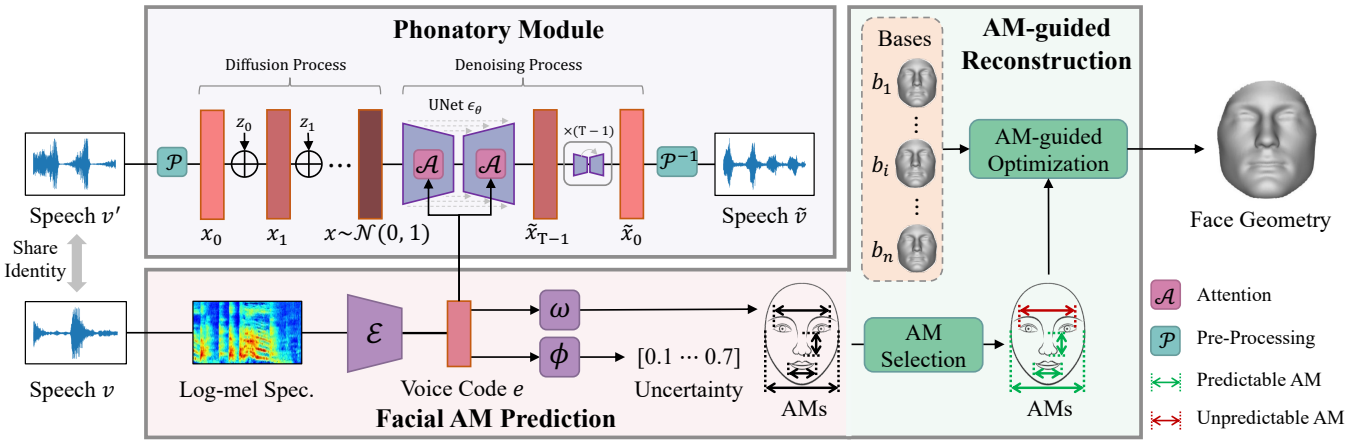
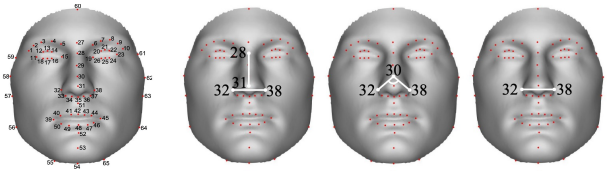


Figure 2: Illustration of our analysis pipeline for voice-face correlation. We randomly pick two voice recordings with shared speaker identity as v and v'. We then analyze the relationship between each AM and voice by predicting each AM from voice with an estimator and an intermediate voice code e. The optional phonatory module equips a diffusion-based voice generation model with a voice code e as a condition to conduct voice style cloning to help us understand the relationship between face geometry and voice characteristics, which serves as an additional constraint to enforce the estimator learn voice identity. We analyze and select AMs with hypothesis testing. The statically significantly predictable AMs are utilized for 3D facial shape reconstruction for further analysis.



(a) Landmarks (b) Proportion (c) Angle (d) Distance

Figure 3: Examples of summarized AMs. We summarized three types of AM: proportion, angle and distance. Those AMs are computed from the predefined landmark on the 3D face representation.

an uncertainty estimator $G_L(v;\mathcal{E}_k,\phi_k) : v \mapsto \mathbb{R}^+ \cup \{0\}$ that v into the variance of the k-th predicted AM. Again, $\mathcal{E}_k$ and ϕ_k are the learnable parameters. The predicted AM and its ground truth become $\mathcal{N}(F_k(v), G_k(v))$ and $\mathcal{N}(m^{(0)}, 0)$ respectively [20]. Given two random variables, a more reasonable learning objective is to minimize their KL divergence.

$$\{\mathcal{E}_k^*, \omega_k^*, \phi_k^*\} = \arg\min_{\mathcal{E}_k,\omega_k,\phi_k} \frac{1}{|\mathcal{D}_t|} \sum_{(v,m^{(k)})\in\mathcal{D}_t} \frac{(F_k(v;\mathcal{E}_k,\omega_k) - m^{(k)})^2}{G_k(v;\mathcal{E}_k,\phi_k)} + \ln G_k(v;\mathcal{E}_k,\phi_k) \tag{2}$$

For a fixed $(F_k(v;\mathcal{E}_k,\omega_k) - m^{(k)})^2$, there is an optimal variance $G_k(v;\mathcal{E}_k,\phi_k) = (F_k(v;\mathcal{E}_k,\omega_k) - m^{(k)})^2$ such that the loss function is minimized. Thereby the uncertainty estimator G_k is learned to produce a small variance if the prediction error is small and vice versa. On the contrary, a smaller variance indicates that the predicted AM is more likely to yield a small prediction error,i.e., close to the ground truth. In this way, we can choose to trust the predicted AMs when the predicted variances are small, and defer the voice recordings to human experts otherwise. An extreme case is $G_k(v) \equiv 1$ where the uncertainty learning objective degrades to the regular regression model.

Temporal aggregation. In practice, following the convention of voice understanding, the long voice recording v is fed into the network in the form of multiple short segments $\{v^{(1)}, \cdots, v^{(L)}\}$. We obtain a sequence of means and variances of the predicted AM. During training, we compute the loss for each segment individually and average them as the training loss. While during evaluation, the predicted AM and its uncertainty are given by aggregating the predictions among all segments. Assuming the short segments from a long recording are class-conditionally independent, the formulations of aggregation are

$$\hat{m}^{(k)} = \sum_{l=1}^{L} \frac{w^{(k)}}{G_k(v^{(l)})} \cdot F_k(v^{(l)}), \qquad \frac{1}{w^{(k)}} = \sum_{l=1}^{L} \frac{1}{G_k(v^{(l)})} \tag{3}$$

where $\hat{m}^{(k)}$ is the aggregated mean and also the predicted k-th AM. However, the aggregated variance $w^{(k)}$ is not used as the uncertainty of the predicted k-th AM since the conditional independence assumption does not always hold in cases such as noises, silences, the computed aggregated variance will be biased by the number of voice segments in the long recording. So we calibrate the uncertainty as $\hat{w}^{(k)} = L \cdot w^{(k)}$.

Predictable AM identification. We have collected a number of AMs and trained estimators for predicting them. However, only a few of the AMs are actually predictable from voice, which we had

anticipated while designing the task. To identify those AMs, we use hypothesis testing to them. Formally, we can write the null and alternative hypotheses for the k-th AM as

$$H_0 : \text{the AM } m^{(k)} \text{ is NOT predictable from voice}$$
$$H_1 : \text{the AM } m^{(k)} \text{ is predictable from voice}$$

In order to reject H_0, we only need to find a counterexample to show that voice is indeed useful in predicting AM $m^{(k)}$. An effective example is to compare the estimators with and without the voice input. If there exists a learned estimator $F_k(v)$ performing better than the chance-level estimator C_k without using voice input and the results are statistically significant, we can successfully reject H_0 and accept H_1. Here the chance-level estimator for the k-th AM is a constant $C_k = \frac{1}{|\mathcal{D}_t|}\sum_{m^{(k)}\in\mathcal{D}_t} m^{(k)}$, which is the mean $m^{(k)}$ of the training set $\mathcal{D}_t$. So the null and alternative hypothesis can be rewritten as

$$H_0 : \mu(\epsilon_k/\epsilon_k^C) \le 1$$
$$H_1 : \mu(\epsilon_k/\epsilon_k^C) \ge 1$$

where ϵ_k and ϵ_k^C are the mean square errors of estimators with and without voice inputs on validation set $\mathcal{D}_{v2}$, respectively. The formulations of ϵ_k and ϵ_k^C are given as $\epsilon_k = \frac{1}{|\mathcal{D}_{v2}|}\sum_{m^{(k)}\in\mathcal{D}_{v2}}(\hat{m}^{(k)} - m^{(k)})^2$ and $\epsilon_k^C = \frac{1}{|\mathcal{D}_{v2}|}\sum_{m^{(k)}\in\mathcal{D}_{v2}}(C_k - m^{(k)})^2$. Since the true variance of ϵ_k/ϵ_k^C is unknown, the type of hypothesis testing is one-sided paired-sample t-test. The upper bound of the confidence interval (CI) is given by

$$CI_u = \mu(\epsilon_k/\epsilon_k^C) + t_{1-\alpha,\nu}\cdot\frac{\sigma(\epsilon_k/\epsilon_k^C)}{\sqrt{N}} \tag{4}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are the functions for computing mean and standard deviation respectively. N is the number of the repeated experiments and we set $N = 100$ here. α and $\nu = N - 1$ are the significance level and the degree of freedom respectively. For the purpose of this section, we adopt the significance level of 5% and then we can read $t_{0.95,N-1}$ from t-distribution table. Now we can determine whether to reject H_0 and accept H_1,i.e., the AM $m^{(k)}$ is predictable from voice. According to the experimental results, the probability that the aforementioned decision is correct is higher than 95%,i.e., statistically significant. In contrast, $CI_u \ge 1$ implies that we fail to reject H_0, for the current experimental results are not statistically significant enough. Note that failing to reject H_0 does not imply we accept H_0.

We emphasize that it is necessary to compute ϵ_k^C and ϵ_k on $\mathcal{D}_{v2}$ rather than $\mathcal{D}_t$ or $\mathcal{D}_{v1}$. This is because our estimators are trained on $\mathcal{D}_t$ and selected by the errors on $\mathcal{D}_{v1}$, we can easily get significantly lower ϵ_k and ϵ_k^C on these splits.

Optional phonatory module. Inspired by linear predictive coding (LPC) [25] which leverages voice producing to learn vocal tract geometry, we aim to facilitate face geometry capture by learning characteristics of voice. We enroll a phonatory module serving as an additional constraint when predicting facial AMs. In particular, we leverage a diffusion-based [18] voice generation method to model the time-domain speech signals. As shown in Fig. 2, the diffusion model converts the noise distribution to a speech $\tilde{v}$ controlled by the voice code e extracted from speech v. During training speech v' which shares speaker identity with v is fed to the diffusion model as ground-truth. Please note that the phonatory module only serves as an additional training constraint and is not applied during inference. Let $x_0, \cdots, x_T$ be a sequence of variables with the same dimension where t is the index for diffusion time steps. Then the diffusion process transforms x_0 into a Gaussian noise x_T through a chain of Markov transitions with a set of variance schedule $\beta_1, \cdots, \beta_T$. Specifically, each transformation is performed according to the Markov transition probability $q(x_t|x_{t-1}, e)$ assumed to be independent of the style code e as

$$q(x_t|x_{t-1}, e) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{5}$$

Unlike the diffusion process, the denoising process aims to recover the speech signal from Gaussian noise which is defined as a conditional distribution $p_\theta(x_{0:T-1}|x_T, c)$. Through the reverse transitions $p_\theta(x_{0:T-1}|x_T, c)$, the variables are gradually restored to a speech signal with style code condition. The phonatory module actually models a distribution $q(x_0|c)$. By applying the parameterization trick [21], we obtain the additional training constraint as

$$\{\mathcal{E}^*, \theta^*\} = \arg\min_{\mathcal{E},\theta} = \mathbb{E}_{x_0,\epsilon,t}\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, e)\|_1 \tag{6}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{t'=1}^{t}\alpha_{t'}$. As shown in Fig. 2, the θ is a Net [37] with cross-attention [36]. Since the phonatory model is only utilized as an auxiliary constraint during training, we omit the inference details to obtain $\tilde{v}$ here.

3.4 AM-Guided 3D Facial Shape Reconstruction

To reconstruct the 3D facial shape, we first need to predict AMs of the voice recordings in $\mathcal{D}_e$ first. Subsequently, we generate the 3D facial shapes based on the predicted AMs by an optimization-based method. To do so, we first project the 3D facial shapes into a low-dimensional linear space [5]. By adjusting the coefficients in low-dimensional space, we obtain different re-projected 3D facial shapes. The learning objective is to find a set of coefficients, such that the differences between the AMs of the re-projected 3D facial shape and the predicted AMs are minimized. Specifically, we construct a big matrix $B = [b_1, b_2, \cdots] \in \mathbb{R}^{3T\times|\mathcal{D}_t|}$ where each column $b_i \in \mathbb{R}^{3T\times 1}$ is a long vector obtained by flattening a 3D facial shape $f_i \in \mathbb{R}^{T\times 3}$. T is the number of vertices on 3D faces. Since $3T \gg |\mathcal{D}_t|$, we compute the project matrix $P \in \mathbb{R}^{3T\times d}(d \ll 3T)$ using eigenfaces [5] on B. Now any flattened 3D facial shape b can be approximated by re-projecting a low-dimensional vector $\beta \in \mathbb{R}^{b\times 1}$ in the form of $P\beta$. We define the computation of AM as $Q_k(b) : b \mapsto \mathbb{R}$, which maps any flattened 3D facial shape b into the k-th AM of b. Since $Q_k(\cdot)$ computes a distance, a proportion, or an angle of the 3D facial shape, it is a differentiable function. The optimization objective is given below.

$$\beta^* = \arg\min_{\beta} \lambda\|\beta\|_2^2 + \sum_{k=1}^{K}(Q_k(P\beta) - \hat{m}^{(k)})^2 \cdot z^{(k)} \tag{7}$$

where λ is the loss weight balancing two terms. The reconstructed 3D facial shape is given by $\hat{b} = P\beta^*$.

4 EXPERIMENTS

In this section, we elaborate on the dataset setting, implementation details and experimental results.

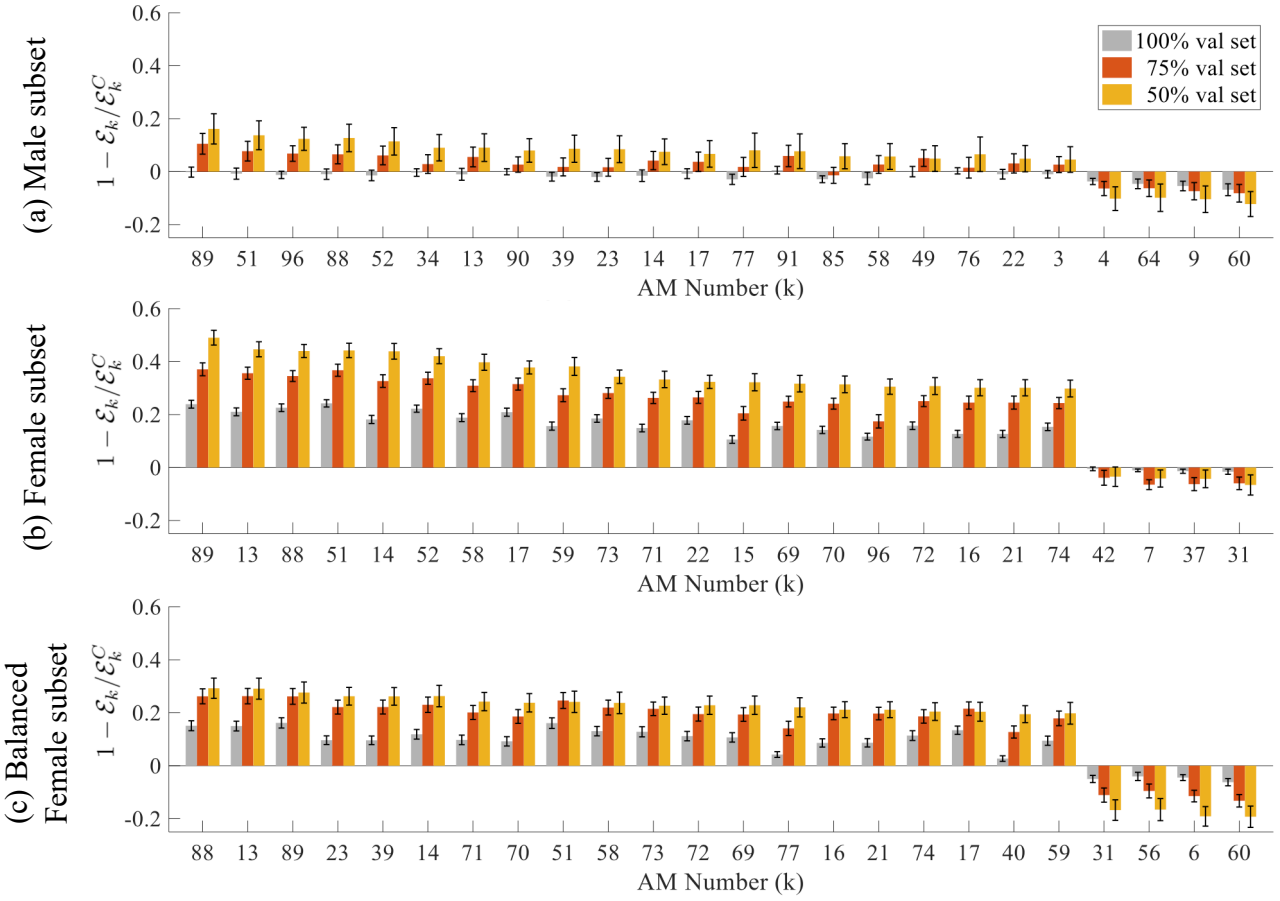


Figure 4: The normalized errors and CIs of 24 AMs on (a) male subset, (b) female subset, and (c) a smaller female subset. If $1 - CI_u > 0$, the AM is predictable else unpredictable.

4.1 Dataset

We perform experiments on a private audiovisual dataset $\mathcal{D}$. The dataset consists of paired voice recordings and scanned 3D facial shapes from 1,026 people, with 364 males and 662 females. The scanned 3D face is stored in the mesh format with 6790 points for each face. The voice recordings are about 2 minutes long for each speaker. We reduce the influencing factors to the voice and face by (1) asking participants to speak a set of specified sentences, (2) asking participants to speak without emotion, (3) control the age of participants (roughly 18-28 years old). In addition, to prevent the models from taking the gender shortcuts, we split the dataset $\mathcal{D}$ by gender, and experiments are individually performed on male and female subsets. For each subset, we adopt 7/1/1/1 splitting for $\mathcal{D}_t/\mathcal{D}_{v1}/\mathcal{D}_{v2}/\mathcal{D}_e$. In training, the voice recordings are randomly trimmed to segments of 6 to 8 seconds, while we use the entire recordings in testing. The ground truth AMs are normalized to zero mean and unit variance. For voice features, we extract 64-dimensional log Mel-spectrograms using an analysis window of 25ms, with the hop of 10ms between frames. We perform mean and variance normalization of each Mel-frequency bin.

4.2 Implementation Details

We leverage a backbone $\mathcal{E}$ to learn voice code e which is a simple convolutional neural network. The detailed network structure is presented in the supplementary materials. F_k and G_k share the backbone's learnable parameters but have individual parameters for their heads. We use a single layer fully-connected network for each head. For the variance head, we add an exponential activation to the last layer of G_k for non-negative positive output. We follow the typical settings of stochastic gradient descent (SGD) for optimization. Minibatch size is 64. The momentum, learning rate, and weight decay values are 0.9, 0.1, and 0.0005, respectively. The training is completed at 5k iterations. Since the phonatory module requires a long training procedure, we first train it with the voice code encoder $\mathcal{E}$ for 60k steps on our training set $\mathcal{D}_t$. We follow the training setting in [18] to train the phonatory module. The other parameter setting follows [18]. We directly normalized the voice signal as input to the network instead of first converting it to Log-Mel spectrum. To ensure statistical significance, we perform N = 100 repeated experiments to compute the CI_u. For the experiments

Phonation Module	100% $\hat{w}$	75% $\hat{w}$	50% $\hat{w}$
✓	0.953± 0.009	0.909±0.024	0.842±0.030
✗	0.952±0.014	0.927±0.030	0.879±0.041

Table 1: Effect of the phonatory module. We measure the normalized mean squared error between predicted and ground-truth AM among all AMs with different confidence thresholds.

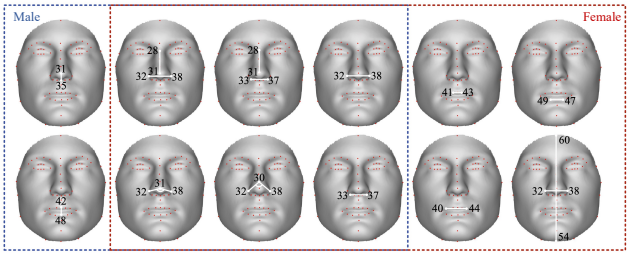


Figure 5: Visualization of the predictable AMs. Blue box: male, Red box: female.

at phoneme level, we leverage Wav2Vec [2] to cut the long voice recordings into phonemes.

4.3 Predictable AM Analysis

For AM prediction, the estimation models are trained on $\mathcal{D}_t$ and selected based on their performance on $\mathcal{D}_{v1}$ (hyperparameter tuning). For AM selection, the predictable AMs are selected based on the upper bound of the CI (CI_u) on $\mathcal{D}_{v2}$. The performance can be evaluated by the mean error of each AM and its CI.

Fig. 4 shows the results, including 20 AMs with highest $1 - CI_u$ and 4 AMs with lowest $1 - CI_u$. The gray bars are the results on the entire validation set $\mathcal{D}_{v2}$, while the red and yellow ones are the results of 75% and 50% voice samples with lowest uncertainty $\hat{w}$ on $\mathcal{D}_{v2}$, respectively. The self-constructed female subset has the same size as the male subset. Higher $1 - CI_u$ indicates better results and the normalized error of 0 indicates the chance-level performance. As suggested by our hypothesis testing formulation, the AMs with $1 - CI_u > 0$ are considered predictable from voice. In this sense, we have discovered a number of predictable female AMs (see the gray bars and their CI_u in Fig. 4 (b)). By filtering out the voice samples with high uncertainties, we achieve even higher $1 - CI_u$ (see the red and yellow bars and their CIs). The improved performance indicates that more AMs are discovered as predictable from voice. The complete results of all AMs are given in the appendix. The results empirically demonstrate that the information of 3D facial shape is indeed encoded in the voices and can be discovered by our analysis pipeline.

To intuitively locate the predictable AMs on the 3D face, we visualize them in Fig. 5. We clearly observe that most of the predictable AMs are around nose and mouth, and many of them are shared between male and female subsets. This is consistent with the fact that nose and mouth shapes affect pronunciation.

We also notice that the performance of female subset is much better than that of the male subset. To investigate whether the improvements come from the larger data scale (364 males *v.s.* 662 females), we perform another set of repeated experiments on a self-constructed female subset, which has the same size as the male subset,i.e., 364 females. Surprisingly, the results on the new subset are still better than those on the male subset, as shown in Fig. 4 (c). This is possible because the female subjects have higher nasalance scores on the nasal sentences [43] among other things, which provides useful information for predicting the AMs around the nose. Here we note that our experiments have revealed that measurements around the nose are highly correlated to voice. More investigations are left for future work.

On the other hand, some AMs have not been shown to be predictable from voice. This observation suggests that voices may only associate with a few specific regions of the 3D facial shape, like the nose and mouth. For the AMs with higher errors than chance level, we do not claim they are not predictable from voice. Instead, we fail to demonstrate their predictability based on our current empirical results. The possible reasons include imperfect modeling, limited data, data noise, etc.

Phonatory Module	Predictable	Unpredictable
✓	0.628±0.021	0.990±0.032
✗	0.730±0.048	1.002±0.031

Table 2: Effect of phonatory module for predictable and unpredictable AMs. We measure the normalized mean squared error between predicted and ground-truth AM among all predictable and unpredictable AMs. Interestingly, we find phonatory module only improves predictable AMs.

4.4 Effect of Phonatory Module

As presented in Table 1, it is evident that utilizing the phonatory module during training enhances the accuracy of predicted AMs. Our evaluation involved computing the normalized error across all AMs with various confidence thresholds. Although the models with and without the phonatory module exhibited a marginal difference in error when evaluating all the data, the ones trained with the phonatory module showed a clear improvement in error when considering more confident samples.

Furthermore, we conducted an error evaluation for predictable and unpredictable AMs as depicted in Table 2. We observed that utilizing the phonatory module resulted in a 0.102-point decrease in normalized error for predictable AMs, highlighting its effectiveness in improving the prediction performance. Interestingly, the phonatory module did not have any apparent effect on unpredictable AMs. Overall, the results indicate that utilizing the phonatory module during training is beneficial for predicting AMs, particularly for predictable ones.

4.5 Phoneme-level Analysis

We also experiment with the voice-face correlation at the phoneme level. For this experiment, we train and evaluate estimators by taking one phoneme as input each time. We computed the average $1 - CI_u$ value for each phoneme across all AMs, as shown in Fig. 7. Our results indicate that /iː/ had the highest average $1 - CI_u$ value of 0.199, while /b/ had the lowest value of -0.06. When the $1 - CI_u$ value is less than 0, it suggests that AMs are generally unpredictable from the corresponding phoneme.

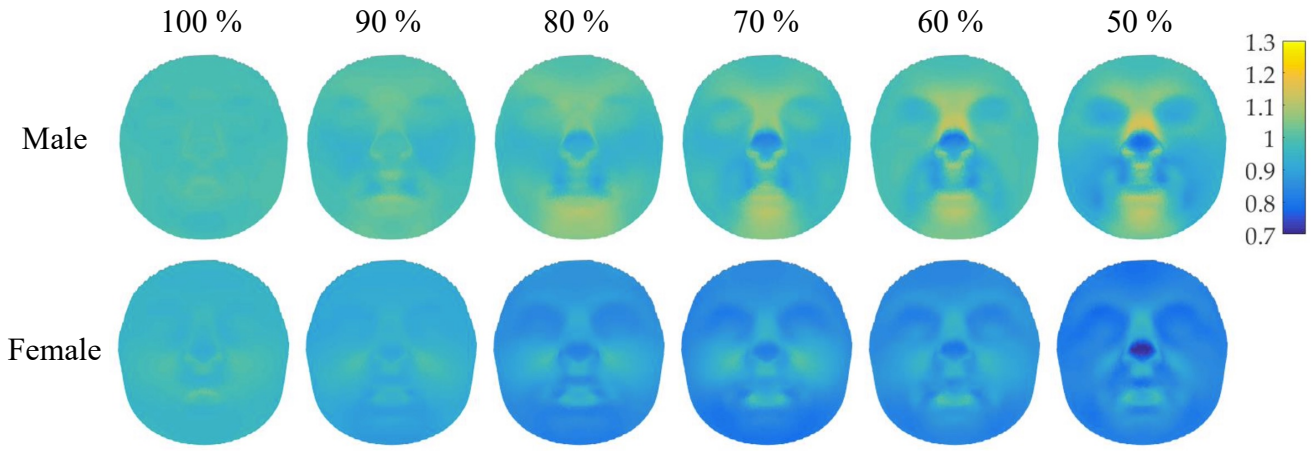


Figure 6: Error maps of the reconstructed 3D facial shapes for the male and female subsets. From left to right: the error maps corresponding to 100% (i.e. the entire test set) to 50% of the test set.

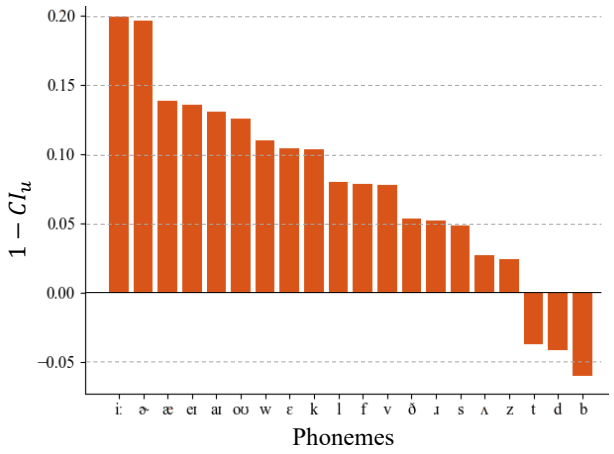


Figure 7: Phonemes with corresponding averaged CI_u in decreasing order.

We observed that the three phonemes with the lowest and negative $1 - CI_u$ values were /t/, /b/, and /d/, all of which are plosive consonants. During the pronunciation of plosive consonants, there is a complete stoppage of airflow followed by a sudden release of air through minimal mouth opening and closing. As a result, there is minimal movement of the facial muscles and structures, making it challenging for the model to predict AMs based solely on these phonemes.

In contrast, most vowels achieved good performance in the test set, with all of the top 6 phonemes belonging to vowels with $1 - CI_u > 0.10$. Compared to consonants, the production of vowels does not involve constriction of airflow in the vocal tract. Instead, the facial muscles have relatively greater movement during the pronunciation of these phonemes, such as jaw movement due to mouth opening or lip spreading. Thus, vowel phonemes may carry more information about facial features, making it easier for the model to capture the hidden correlation when predicting AMs.

4.6 3D Facial Shape Reconstruction

In Section 4.3, we have discovered a number of predictable AMs, from which we choose 10 AMs with the highest $1 - CI_u$ for the subsequent reconstructions on male and female subsets.

To evaluate the performance, we compute the per-vertex errors between the reconstructed 3D facial shape and their ground truths. We also filter out a portion of voice samples with the highest uncertainties and evaluate the errors in the remaining data. The filter-out rate is from 0% to 50%, as shown from left to right in Fig. 6.

Unsurprisingly, we achieve the lowest errors around the nose region for male and female subsets, consistent with the AM estimations. Moreover, the reconstruction errors decrease significantly by filtering out the voice samples with the highest uncertainties. This indicates that the learned uncertainty is effectively associated with the reconstruction quality and allows the system to decide whether to trust the model or not.

5 CONCLUSION

In conclusion, this paper presents a novel approach to exploring the voice-face correlation by focusing on the geometric aspects of the face rather than relying on semantic cues such as gender, age, and emotion. The proposed voice-anthropometric measurement (AM)-face paradigm identifies predictable facial AMs from the voice to guide 3D face reconstruction, which results in significant correlations between voice and specific parts of the face geometry, such as the nasal cavity and cranium. This approach not only eliminates the influence of unpredictable AMs but also offers a new perspective on voice-face correlation, which can be valuable for anthropometry science. The results of this study open up possibilities for future research in this area, such as developing more accurate voice-guided face synthesis techniques and a better understanding of the relationship between voice and facial geometry.

REFERENCES

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

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[3] Mohamad Hasan Bahari, Mitchell McLaren, Hugo Van hamme, and David A. van Leeuwen. 2012. Age Estimation from Telephone Speech using i-vectors. In *Interspeech*.

[4] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.

[5] Volker Blanz and Thomas Vetter. 2003. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on pattern analysis and machine intelligence* 25, 9 (2003), 1063–1074.

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

[7] R. H. C. Bull, Harriet Rathborn, and Brian R. Clifford. 1983. The Voice-Recognition Accuracy of Blind Listeners. *Perception* 12 (1983), 223 – 226.

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

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

[10] Leslie G Farkas, Otto G Eiben, Stefan Sivkov, Bryan Tompson, Marko J Katic, and Christopher R Forrest. 2004. Anthropometric measurements of the facial framework in adulthood: age-related changes in eight age categories in 600 healthy white North Americans of European ancestry from 16 to 90 years of age. *Journal of Craniofacial Surgery* 15, 2 (2004), 288–298.

[11] Donya Ghafourzadeh, Cyrus Rahgoshay, Sahel Fallahdoust, Adeline Aubame, Andre Beauchamp, Tiberiu Popa, and Eric Paquette. 2019. Part-based 3D face morphable model with anthropometric local control. (2019).

[12] Prasanta Kumar Ghosh and Shrikanth Narayanan. 2011. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America* 130, 4 (2011), EL251–EL257.

[13] Patrick Gomez and Brigitta Danuser. 2007. Relationships between musical structure and psychophysiological measures of emotion. *Emotion* 7, 2 (2007), 377.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[15] Joanna Grzybowska and Stanislaw Kacprzak. 2016. Speaker Age Classification and Regression Using i-Vectors.. In *INTERSPEECH*. 1402–1406.

[16] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *ICCV*.

[17] Jing Han, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. 2021. Exploring Automatic COVID-19 Diagnosis via voice and symptoms from Crowdsourced Data. In *ICASSP*. IEEE.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

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

[20] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* 30 (2017).

[21] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[22] Sheng Li, Dabre Raj, Xugang Lu, Peng Shen, Tatsuya Kawahara, and Hisashi Kawai. 2019. Improving Transformer-Based Speech Recognition Systems with Compressed Structure and Speech Attributes Augmentation.. In *INTERSPEECH*. 4400–4404.

[23] Sheng Li, Dabre Raj, Xugang Lu, Peng Shen, Tatsuya Kawahara, and Hisashi Kawai. 2019. Improving Transformer-Based Speech Recognition Systems with Compressed Structure and Speech Attributes Augmentation. In *Interspeech*.

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

[25] John D Markel, Augustine H Gray, and Augustine H Gray. 1976. Linear prediction of speech: Communication and cybernetics. (1976).

[26] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[27] Shah Nawaz, Muhammad Saad Saeed, Pietro Morerio, Arif Mahmood, Ignazio Gallo, Muhammad Haroon Yousaf, and Alessio Del Bue. 2021. Cross-Modal Speaker Verification and Recognition: A Multilingual Perspective. In *CVPRW*.

[28] Hailong Ning, Xiangtao Zheng, Xiaoqiang Lu, and Yuan Yuan. 2021. Disentangled Representation Learning for Cross-modal Biometric Matching. *TMM* (2021).

[29] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. 2019. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7539–7548.

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

[31] Paul H. Ptacek and Eric K. Sander. 1966. Age recognition from voice. *Journal of speech and hearing research* 9 2 (1966), 273–7.

[32] Liao Qu, Xianwei Zou, Xiang Li, Yandong Wen, Rita Singh, and Bhiksha Raj. 2023. The Hidden Dance of Phonemes and Visage: Unveiling the Enigmatic Link between Phonemes and Facial Features. *arXiv preprint arXiv:2307.13953* (2023).

[33] Narayanan Ramanathan and Rama Chellappa. 2006. Modeling age progression in young faces. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 1. IEEE, 387–394.

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

[35] Mirco Ravanelli and Yoshua Bengio. 2018. Speaker Recognition from Raw Waveform with SincNet. *2018 IEEE Spoken Language Technology Workshop (SLT)* (2018), 1021–1028.

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[38] Leda Sarı, Kritika Singh, Jiatong Zhou, Lorenzo Torresani, Nayan Singhal, and Yatharth Saraf. 2021. A Multi-View Approach to Audio-Visual Speaker Verification. In *ICASSP*.

[39] Zhiyi Shan, Richard Tai-Chiu Hsung, Congyi Zhang, Juanjuan Ji, Wing Shan Choi, Wenping Wang, Yanqi Yang, Min Gu, and Balvinder S Khambay. 2021. Anthropometric accuracy of three-dimensional average faces compared to conventional facial measurements. *Scientific Reports* 11, 1 (2021), 1–12.

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

[41] Rita Singh, Bhiksha Raj, and Deniz Gencaga. 2016. Forensic anthropometry from voice: an articulatory-phonetic approach. In *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 1375–1380.

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

[43] Tomáš Vampola, Jaromír Horáček, Vojtěch Radolf, Jan G Švec, and Anne-Maria Laukkanen. 2020. Influence of nasal cavities on voice quality: Computer simulations and experiments. *The Journal of the Acoustical Society of America* 148, 5 (2020), 3218–3231.

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

[45] Peisong Wen, Qianqian Xu, Yangbangyan Jiang, Zhiyong Yang, Yuan He, and Qingming Huang. 2021. Seeking the Shape of Sound: An Adaptive Framework for Learning Voice-Face Association. In *CVPR*. 16347–16356.

[46] Yandong Wen, Bhiksha Raj, and Rita Singh. 2019. Face Reconstruction from Voice using Generative Adversarial Networks. In *NeurIPS*, Vol. 32.

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

[48] Cho-Ying Wu, Chin-Cheng Hsu, and Ulrich Neumann. 2022. Cross-Modal Perceptionist: Can Face Geometry be Gleaned from Voices?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10452–10461.

[49] Marzena Wyganowska-Swikatkowska, Iwona Kowalkowska, Grazyna Flicinska-Pamfil, Mikolaj Dabrowski, Przemyslaw Kopczynski, and Bozena Wiskirska-Woznica. 2017. Vocal training in an anthropometrical aspect. *Logopedics Phoniatrics Vocology* 42, 4 (2017), 178–186.

[50] Marzena Wyganowska-Swikatkowska, Iwona Kowalkowska, Katarzyna Mehr, and Mikolaj Dkabrowski. 2013. An anthropometric analysis of the head and face in vocal students. *Folia Phoniatrica et Logopaedica* 65, 3 (2013), 136–142.

[51] Muqiao Yang, Joseph Konan, David Bick, Yunyang Zeng, Shuo Han, Anurag Kumar, Shinji Watanabe, and Bhiksha Raj. 2023. PAAPLoss: A Phonetic-Aligned Acoustic Parameter Loss for Speech Enhancement. *Proc. of ICASSP* (2023).

[52] Zixing Zhang, Bingwen Wu, and Björn Schuller. 2019. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP*.

IEEE, 6705–6709.
[53] Zixing Zhang, Bingwen Wu, and Björn Schuller. 2019. Attention-augmented End-to-end Multi-task Learning for Emotion Prediction from Speech. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), 6705–6709.
[54] Aihua Zheng, Menglan Hu, Bo Jiang, Yan Huang, Yan Yan, and Bin Luo. 2021. Adversarial-metric learning for audio-visual cross-modal matching. *TMM* (2021).
[55] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, Vol. 33. 9299–9306.
[56] Ziqing Zhuang, Douglas Landsittel, Stacey Benson, Raymond Roberge, and Ronald Shaffer. 2010. Facial anthropometric differences among gender, ethnicity, and age groups. *Annals of occupational hygiene* 54, 4 (2010), 391–402.