The Hidden Dance of Phonemes and Visage: Unveiling the Enigmatic Link between Phonemes and Facial Features

Liao Qu [1,*], *Xianwei Zou* [1,*], *Xiang Li* [1,*], *Yandong Wen* [2], *Rita Singh* [1], *Bhiksha Raj* [1,3]

[1]Carnegie Mellon University
[2]Max Planck Institute
[3]Mohamed bin Zayed University of Artificial Intelligence

{liaoq,xianweiz,xl6,yandongw,rsingh,bhiksha}@andrew.cmu.edu

Abstract

This work unveils the enigmatic link between phonemes and facial features. Traditional studies on voice-face correlations typically involve using a long period of voice input, including generating face images from voices and reconstructing 3D face meshes from voices. However, in situations like voice-based crimes, the available voice evidence may be *short* and *limited*. Additionally, from a physiological perspective, each segment of speech - *phoneme* corresponds to different types of airflow and movements in the face. Therefore, it is advantageous to discover the hidden link between phonemes and face attributes. In this paper, we propose an analysis pipeline to help us explore the voice-face relationship in a fine-grained manner, i.e., phonemes *vs.* facial anthropometric measurements (AM). We build an estimator for each phoneme-AM pair and evaluate the correlation through hypothesis testing. Our results indicate that AMs are more predictable from vowels compared to consonants, particularly with plosives. Additionally, we observe that if a specific AM exhibits more movement during phoneme pronunciation, it is more predictable. Our findings support those in physiology regarding correlation and lay the groundwork for future research on speech-face multimodal learning.

Index Terms: voice-face correlation, phoneme

1. Introduction

The implicit relation between speech and anthropometry features has been extensively researched in recent years. Numerous voice profiling studies [1, 2, 3, 4, 5, 6] have shown that human voice carries a plethora of information about the speaker, making it possible to deduce biophysical characteristics of speakers, e.g., gender, age and health conditions, from their voice. However, in criminal profiling scenarios, the study of correlations between voice and face becomes essential. In voice-based crimes, such as hoax emergency calls and voice-based phishing, the officers seek to depict the facial features of the criminal merely from *short* voice evidence. "Mayday" can be an example of the audio samples obtained by officers. This motivates us to investigate the phoneme-level correlation between voice and face.

Several recent works have attempted to investigate the correlation between voice and face. Cognitive science studies [7, 8] suggests human has a strong capability to imagine the appearance of speakers based on their voice. To verify it, face reconstruction from voice, which aims to recover identity-fidelity faces from their corresponding voice recordings, is introduced by [9]. After that, great progresses [10, 11] has been achieved by using advanced Generative Adversarial Networks [12]. Going beyond, recent works [13, 14] attempt to recover 3D face geometry meshes from voice to avoid the impact of inevitable background area modeling in 2D images. However, all these approaches rely on a long period of voice and potentially neglect the advantage of exploring a more fine-grained voice-face correspondence.

Rethinking the human voice production mechanism, the voice is produced by either the vibration of the vocal cord or the resonance of the pulmonary airflow. For both of the mechanisms, the vocal track is highly enrolled. The vocal track can be assumed as a filter, reflecting the characteristics of human voice. With the tight bind of muscle and bone, the vocal track is also correlated with facial attributes. Specifically, each phoneme corresponds to a different vocal track status and also an accordingly facial movement. To construct an accurate voice-face correlation, we argue that phoneme-level voice-face modeling is vital.

To investigate and understand the voice-face correlation at a more fine-grained phoneme level, we propose an analysis pipeline that leverages a common feature extractor with a regression head to predict human anthropometric measurements (AM) from phoneme. Specifically, Human anthropometric measurements are a set of facial measurements summarized from cognitive science studies that can effectively represent the identity of a human. We decompose the audio recordings into phonemes and learn to predict AMs from phonemes. In this way, we can quantitatively analyze the relationship between each facial AM and phoneme pairs. In this paper, we aim to answer core two questions: 1) whether there exists any "enigmatic" link between phoneme and facial features and 2) whether those "enigmatic" links can be quantitatively described.

2. Related Works

Learning Human Attributes from Voice. There is a substantial body of research on inferring human attributes from a person's voice, including speaker identity [15, 16], age [1, 3], gender [4], and emotion status [17, 6]. In addition to predicting attributes directly related to voice, many studies have explored the implicit correlation between voice and facial features. One popular task is generating 2D face images from voice using GANs [12], which has been progressed in several recent works [10, 9, 11]. To avoid the impact of inevitable background area modeling in 2D images, recent work turns to the 3D domain: synthesizing 3D meshes from voices [13].

Phoneme Pronunciation Mechanism. The human vocal tract can be considered as a series of resonance chambers that can be dynamically configured [18]. When the vocal cords vibrate, they convert the airflow from the lungs into acoustic energy in

[*] Equal contribution.

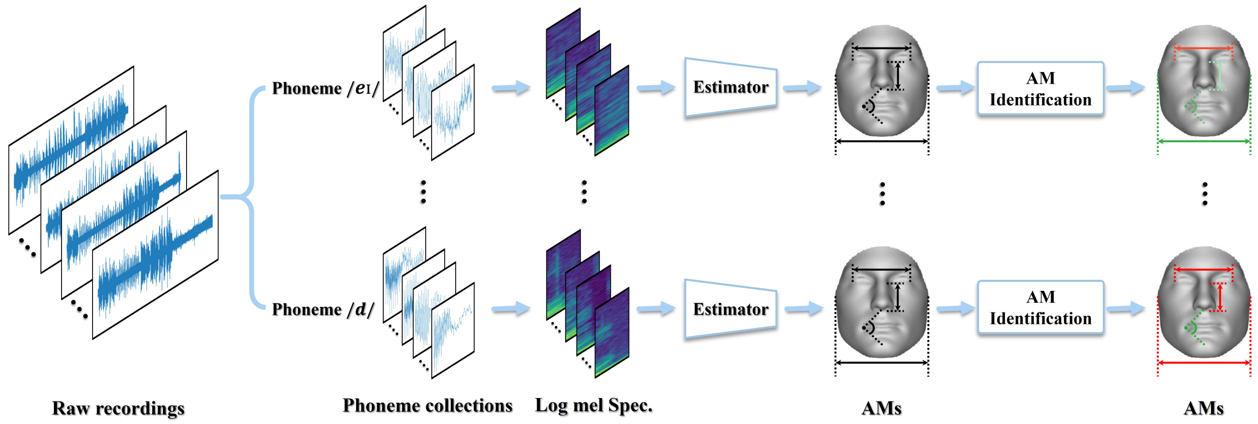

Figure 1: *Illustration of our framework. We convert phoneme clips into mel spectrograms and develop estimators for each phoneme-acoustic model (AM) pair. Hypothesis testing is used to determine the predictability of AMs from phonemes. Green denotes predictable AMs and red otherwise, where the color shade indicates the degree of predictability.*

the form of sound waves, producing *voice*. The shape and dimensions of the resonant chambers change as the movements of the vocal tract modify the acoustic signal, resulting in different patterns [19, 20]. To produce a specific pattern, the mouth and nose must form a corresponding shape. Each pattern corresponds to a unique compositional unit of speech, or a *phoneme*. When a speaker enunciates different phonemes, the vocal tract, mouse, nose, and other related facial structures act in concert. And each phoneme, therefore, carries some information about all these related features.

3. Methods

3.1. Overview

We aim to investigate the correlations between each phoneme and AM pair. As shown in Fig 1, we first transform the segmented phonemes into log mel spectrum to better capture information from the frequency domain. After that, an AM estimator is employed to predict each AM from phonemes. Finally, we use hypothesis testing to analyze the correlation between each phoneme-AM pair.

3.2. Notations

Our problem involves a set of paired voice recordings of phonemes and AMs, where we aim to predict each AM from different phonemes. We begin by segmenting the recordings into phonemes, which can be represented as $\boldsymbol{P} = p^{(1)}, p^{(2)}, \ldots, p^{(k)}$, where k denotes the total number of distinct phonemes. Similarly, the AMs can be represented as $AMs = m^{(1)}, m^{(2)}, \ldots, m^{(n)}$, where n represents the number of summarized AMs. We refer to the entire dataset as D.

To simplify the training and evaluation process, we divide D into three subsets. The first subset is the training set D_t, which is used for estimator learning. The second subset is the validation set D_{v_1}, which is used for estimator selection. Finally, the third subset is the validation set D_{v_2}, which is used for hypothesis testing and AM-phoneme pair selection.

3.3. AM Estimator

We leverage an AM estimator E_{ij} to predict the j-th AM from the i-th phoneme $m^{(i)} = E_{ij}(p^{(j)})$ as an estimator that maps the j-th phoneme to the i-th AM. To begin, we transform each phoneme into a log mel spectrum, which is essentially an image. This is a classic regression problem, and therefore, we need a model with strong feature extraction capabilities to extract information from the image. We develop a modified version of the classical MNasNet model developed by Google AI. It is designed to be efficient, lightweight, and highly accurate for tasks such as image classification and object detection. [21]. Our modification retains the original structure, but with a few modifications to the input and output layers. Specifically, we change the input Conv2d module to accept only 1 channel, and the output Linear module to produce only 1 value. In addition, since this part is model-independent, other models such as ResNet [22] are also capable of achieving the same function.

3.4. Hypothesis Testing for Phoneme-AM Predictability

Once AMs are predicted from different phonemes, the next step is to determine whether a specific phoneme can actually predict an AM. To do this, we use hypothesis testing for each AM-phoneme pair separately. Firstly, we write the null hypothesis and the alternative hypothesis for the i-th AM and the j-th phoneme as

$$H_0 : \text{AM } m^{(i)} \text{ is not predictable from phoneme } p^{(j)}$$

$$H_1 : \text{AM } m^{(i)} \text{ is predictable from phoneme } p^{(j)}$$

To reject the null hypothesis H_0, we need to compare our estimator E_{ij} for the AM $m^{(i)}$ when using phoneme $p^{(j)}$ as input with a chance-level estimator C_{ij}. If the performance of E_{ij} is statistically significantly better than C_{ij}, we can reject H_0 and accept H_1. To estimate the chance level for phoneme $p^{(j)}$ in our training set D_t, we use the mean $m^{(i)}$ of all instances of that phoneme in the set. Specifically, we calculate a constant value C_{ij} as follows: $C_{ij} = \frac{1}{|D_t|} \sum_{m^{(i)} \in D_t} m^{(i)}$. We can express the hypotheses as:

$$H_0 : \mu(\varepsilon_{ij}/\varepsilon_{ij}^C) \geqslant 1$$
$$H_1 : \mu(\varepsilon_{ij}/\varepsilon_{ij}^C) < 1$$

Here, $\mu(\cdot)$ represents the mean function, and ε_{ij} and ε_{ij}^C are the mean squared errors (MSE) of the estimators E_{ij} and C_{ij} on the validation set D_{v_2}, respectively. We can compute them as follows:

$$\varepsilon_{ij} = \frac{1}{|D_{v_2}|} \sum_{m^{(i)} \in D_{v_2}} (\hat{m}^{(i)} - m^{(i)})^2$$

$$\varepsilon_{ij}^C = \frac{1}{|D_{v_2}|} \sum_{m^{(i)} \in D_{v_2}} (C_{ij} - m^{(i)})^2$$

To conduct repeated experiments, we need to train the estimators multiple times. In each iteration, we randomly split the dataset into D_t, D_{v_1}, and D_{v_2}. We then use the one-sided paired-sample t-test to test the hypothesis. The confidence interval (CI) bounds are:

$$CI_l = \mu\left(\frac{\varepsilon_{ij}}{\varepsilon_{ij}^C}\right) - t_{1-\alpha,\nu} \cdot \frac{\sigma(\varepsilon_{ij}/\varepsilon_{ij}^C)}{\sqrt{N}}$$

$$CI_u = \mu\left(\frac{\varepsilon_{ij}}{\varepsilon_{ij}^C}\right) + t_{1-\alpha,\nu} \cdot \frac{\sigma(\varepsilon_{ij}/\varepsilon_{ij}^C)}{\sqrt{N}}$$

Here, $\sigma(\cdot)$ represents the standard deviation function, N represents the number of experiments, α represents the significance level, and $\nu = N - 1$ represents the degree of freedom. For this project, we set $N = 10$, and we choose $\alpha = 0.05$ to obtain statistically significant results. We can read the value of $t_{1-\alpha,\nu}$ directly from the t-distribution table. To test the hypothesis, if the CI upper bound $CI_u < 1$, we can infer that we successfully reject H_0 and accept H_1, meaning that the AM $m^{(i)}$ is predictable from phoneme $p^{(j)}$. On the contrary, if $CI_u \geq 1$, we cannot reject H_0, indicating that the result is not statistically significant.

4. Experiments

4.1. Dataset

We conducted experiments on a private audio-visual dataset D. The dataset contains 1,026 individuals' paired voice recordings and scanned 3D facial shapes. Each recording is a raw speech speaking out general phonemes and sentences with a length of 1-2 minutes. Each facial data consists of 6790 3D-coordinate points collected from one person.

4.2. Data Processing and Training

Phoneme segmentation. To identify predictable AMs and their corresponding phonemes, the first step is to extract individual phonemes from the dataset. However, due to the large amount of data and the complexity of distinguishing phoneme intervals, manually segmenting phonemes can be laborious, difficult, and imprecise.

To improve the accuracy of phoneme segmentation, we employ state-of-the-art phoneme segmentation approaches. Specifically, we use the Wav2Vec2-Large-XLSR-53 model [23] developed by FAIR, which learns powerful speech representations from more than 50.000 hours of unlabeled speech. This model is trained using a contrastive task on masked latent speech representations and can learn a quantization of the latent shared across languages. It is then fine-tuned on multi-lingual labeled common voice data. Since our data primarily contains standard English pronunciations, this model can provide relatively high segmentation accuracy.

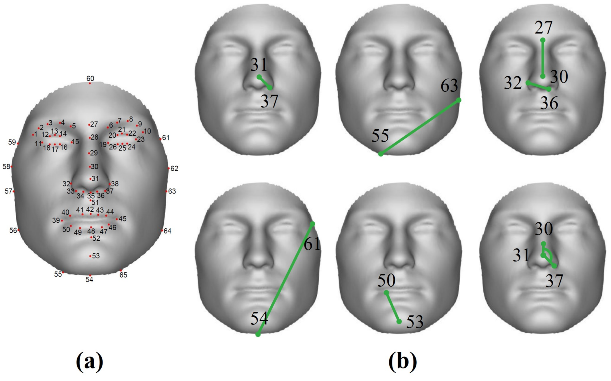

Figure 2: *(a) The selected landmarks. (b) The visualization of the 6 most predictable AMs. They are arranged in descending order from left to right and top to bottom. Numbers in the face denote the index of landmarks.*

We adopt the `wav2vec2-xlsr-53-espeak-cv-ft` in huggingface [1] in our experiments. After splitting, we choose the most frequently used phonemes which have number of samples $\geq$ 5000. The detailed list is provided in the label of Fig. 3. For each phoneme recording, we follow [11] and perform 64-dimensional log mel-spectrograms using an analysis window of 25ms, with a hop of 10ms between frames. We perform normalization by mean and variance of each mel-frequency bin.

AM summarization. We summarize the most commonly used AMs [24, 25, 26, 27, 28], including distances, proportions, and angles in Table 1. The selected landmark is shown in Fig. 2 (a). These AMs are more robust than 3D coordinate representations. This is attributed to the complete elimination of variations induced by spatial misalignment, thus rendering them more reliable and resistant to perturbations. The ground truth AMs are normalized to have a mean of zero and a variance of one.

Table 1: *The summarized AMs.*

	distance	
31-37	32-36	40-42
39-43	33-35	50-53
2-7	30-53	59-53
55-63	54-61	
	proportion	
31-37 / 27-30	32-36 / 27-30	31-37 / 59-53
32-36 / 59-53	54-64 / 31-37	56-62 / 31-37
	angle	
31-30-37	31-29-37	29-30-34

Training details. For each phoneme-AM pair, we conduct 10 repeated experiments to ensure statistical significance. In

[1] *https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft*

each experiment, we randomly sample 5000 data samples and randomly split them into the $D_t/D_{v_1}/D_{v_2}$ set in the ratio of $70\%/10\%/20\%$. We follow the typical settings of Adam [29] for optimization of the estimator. The loss function we use is the mean squared error loss. The size of the mini-batch and learning rate is set to 128 and 0.0001, respectively.

4.3. Results

4.3.1. Analysis of phonemes

For each phoneme, we calculate the average $1 - CI_u$ result with every AMs. As can be seen from Fig. 3, /iː/ got the highest avg. $1 - CI_u$ value 0.199, and /b/ got the lowest value -0.06. When $1 - CI_u$ is lower than 0, AMs are averagely unpredictable from the phoneme. The three phonemes with the lowest and negative values are /t/, /b/ and /d/, which are all plosive consonants. During the pronunciation of plosive consonants, we complete stoppage of airflow followed by a sudden release of air through trivial mouse open and close, and there is minimal movement of the facial muscles and structures. Consequently, the prediction of any acoustic model based solely on such phonemes is challenging. On the contrary, most vowels achieve good performance in the test set, and all the top 6 phonemes belong to vowels with $1 - CI_u > 0.10$. Compared with consonants, there is no constriction of airflow in the vocal tract when pronouncing vowels. In order to produce specific vowels, 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. This, therefore, can make the model better capture the hidden correlation when predicting AMs.

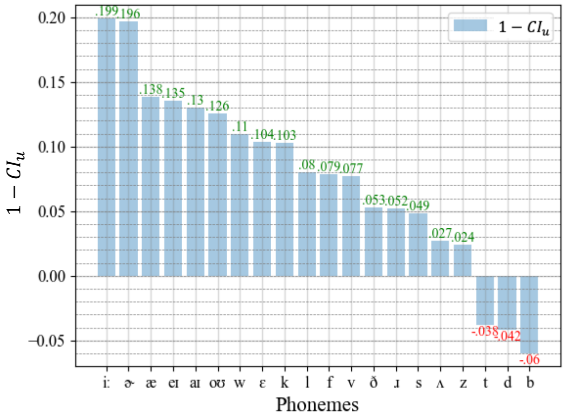

Figure 3: *Phonemes in descending order by avg.* $1 - CI_u$.

4.3.2. Analysis of AMs

Similarly, for each AM, we also calculated the average $1-CI_u$ results with all phonemes. To intuitively locate the most predictable AMs (with the highest avg. $1 - CI_u$) on the 3D face, we visualize them in Fig. 2 (b). Most of the predictable AMs are around the nose and mouth. On the contrary, AMs around the eyes are less predictable. This is consistent with the fact that the nose and mouth shapes (distances, proportions, and angles) affect the pronunciation of phonemes. Other than the nose and mouth, the jaw is another region frequently occurring in the most predictable AMs. Since the jaw is another region that exhibits frequent movement during pronunciation, we hypothesize that for a specific AM, if it is more frequently moved during the pronunciation of phonemes, the AM is generally more predictable. We further verify this hypothesis in the next section.

4.3.3. Relationship between phonemes and AMs

Table 2: *Detailed results of phoneme-AM pairs.*

AMs	/ɛ/	/ð/	/f/	/iː/	/v/	/w/	/æ/
39-43	0.10	-0.04	0.08	0.23	0.11	0.18	0.18
31-30-37	0.10	0.04	0.19	0.11	0.10	0.21	-0.09
50-53	0.05	0.09	-0.07	0.21	0.06	0.11	0.21
2-7	0.02	0.08	0.05	0.04	-0.03	0.08	0.09

We investigate the detailed relationship between phoneme and AM pairs to verify our hypothesis. As shown in Table 2, we list 4 typical AMs paired with 7 phonemes, where 39-43 is an oblique distance of the lip, 31-30-37 is an angle in the nose, 50-53 is the distance between the lip and jaw, and 2-7 is the distance between eyebrow. In the case of AM 2-7, no matter pairing with any phoneme, the value is relatively low (all $1 - CI_u$ values close to 0). During the pronunciation process of any phonemes, the movement of this particular region is very limited. Therefore, phonemes can barely carry information about AMs in this part. However, for 39-43, it shows that /iː/, /w/, and /æ/ have the highest value. When pronouncing these three phonemes, the mouth usually grins in order to control the output airflow. And the distance between facial landmarks 39 and 43 could slightly influence the airflow from a physical perspective, therefore the phoneme produced may have subtle differences. In contrast, when pronouncing /f/ and /ð/, this AM barely moves. Thus, hardly can this AM influences the output airflow. The results verify that it is less predictable for these phoneme-AM pairs. Similarly, the phenomenon also occurs in other AMs like 50-53. It is more predictable when pairing with phonemes than needing mouse opening. All these experiments verify that for a specific AM, if it is more frequently moved during the pronunciation of phonemes, then the AM would be more predictable.

5. Conclusions

In this work, we delve deeply into a fundamental question: whether there exists any “enigmatic” link between phoneme and facial features. If so, whether those “enigmatic” links can be quantitatively described? As a forerunner in this field, we design a phoneme-AMs paradigm, which enables us to explore the speech-face relationship in a fine-grained manner. Hypothesis testing is utilized to verify whether an AM is predictable from a phoneme. Experiments show that AMs are averagely more predictable with vowels compared with consonants, especially plosives, and are consistent with the physiological explanation. On the other hand, most of the predictable AMs are around the nose, mouth, and jaw. Results also verify that for a specific AM, if it moves more frequently during phoneme pronunciation, the AM will be more predictable since the phonemes might carry this hidden information during pronunciation. We hope our work lays a foundation for this field. In the future, we would like to scale up the range of phonemes and AMs to discover more hidden relationships. Moreover, we are investigating models that make use of the found hidden correlation knowledge in scenarios like 3D face reconstruction.

6. References

[1] M. H. Bahari, M. McLaren, H. V. hamme, and D. A. van Leeuwen, "Age estimation from telephone speech using i-vectors," in *Interspeech*, 2012.

[2] S. McGilloway, R. Cowie, and E. Douglas-Cowie, "Automatic recognition of emotion from voice: a rough benchmark," 2000.

[3] P. H. Ptacek and E. K. Sander, "Age recognition from voice." *Journal of speech and hearing research*, vol. 9 2, pp. 273–7, 1966.

[4] S. Li, D. Raj, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation," in *Interspeech*, 2019.

[5] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks," *Multimedia Tools and Applications*, vol. 81, pp. 3535 – 3552, 2022.

[6] Z. Zhang, B. Wu, and B. Schuller, "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)*, pp. 6705–6709, 2019.

[7] P. Belin, S. Fecteau, and C. Bédard, "Thinking the voice: neural correlates of voice perception," *Trends in Cognitive Sciences*, vol. 8, pp. 129–135, 2004.

[8] W. J. Hardcastle and J. Laver, "The handbook of phonetic sciences," *Language*, vol. 75, p. 152, 1999.

[9] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7531–7540, 2019.

[10] H.-S. Choi, C. Park, and K. Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," *ArXiv*, vol. abs/2004.05830, 2020.

[11] Y. Wen, B. Raj, and R. Singh, "Face reconstruction from voice using generative adversarial networks," in *Neural Information Processing Systems*, 2019.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[13] C.-Y. Wu, C.-C. Hsu, and U. Neumann, "Cross-modal perceptionist: Can face geometry be gleaned from voices?" *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 442–10 451, 2022.

[14] C.-Y. Wu, K. Xu, C.-C. Hsu, and U. Neumann, "Voice2mesh: Cross-modal 3d face model generation from voices," *ArXiv*, vol. abs/2104.10299, 2021.

[15] R. H. C. Bull, H. Rathborn, and B. R. Clifford, "The voice-recognition accuracy of blind listeners," *Perception*, vol. 12, pp. 223 – 226, 1983.

[16] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, 2018.

[17] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5150–5154, 2017.

[18] T. Riede, E. Bronson, H. Hatzikirou, and K. Zuberbühler, "Vocal production mechanisms in a non-human primate: morphological data and a model," *Journal of Human Evolution*, vol. 48, no. 1, pp. 85–96, 2005.

[19] R. Singh, B. Raj, and D. Gençaga, "Forensic anthropometry from voice: An articulatory-phonetic approach," *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1375–1380, 2016.

[20] M. T. Ghiselin, P. Ekman, and H. E. Gruber, "Darwin and facial expression: A century of research in review.@@@darwin on man: A psychological study of scientific creativity." *Systematic Biology*, vol. 23, p. 562, 1974.

[21] M. Tan, B. Chen, R. Pang, V. Vasudevan, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2815–2823, 2018.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[23] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *ArXiv*, vol. abs/2109.11680, 2022.

[24] D. Ghafourzadeh, C. Rahgoshay, S. Fallahdoust, A. Beauchamp, A. Aubame, T. Popa, and E. Paquette, "Part-based 3d face morphable model with anthropometric local control," in *Graphics Interface*, 2019.

[25] Z. Shan, R. T. C. Hsung, C. Zhang, J. Ji, W. S. Choi, W. Wang, Y. Yang, M. Gu, and B. S. Khambay, "Anthropometric accuracy of three-dimensional average faces compared to conventional facial measurements," *Scientific Reports*, vol. 11, 2021.

[26] Z. Zhuang, D. Landsittel, S. M. Benson, R. J. Roberge, and R. Shaffer, "Facial anthropometric differences among gender, ethnicity, and age groups." *The Annals of occupational hygiene*, vol. 54 4, pp. 391–402, 2010.

[27] Y. Wen, "Reconstruction of human faces from voice," Ph.D. dissertation, Carnegie Mellon University, 2022.

[28] L. G. Farkas, O. G. Eiben, S. T. Sivkov, B. Tompson, M. Katić, and C. R. Forrest, "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*, vol. 15, pp. 288–298, 2004.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.