What Does Your Face Sound Like? 3D Face Shape towards Voice

Zhihan Yang[1], Zhiyong Wu[1*], Ying Shan[2], Jia Jia[3*]

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[2]Applied Research Center (ARC), Tencent PCG, Shenzhen 518054, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
zhihan-y21@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, jjia@tsinghua.edu.cn

Abstract

Face-based speech synthesis provides a practical solution to generate voices from human faces. However, directly using 2D face images leads to the problems of uninterpretability and entanglement. In this paper, to address the issues, we introduce 3D face shape which (1) has an anatomical relationship between voice characteristics, partaking in the "bone conduction" of human timbre production, and (2) is naturally independent of irrelevant factors by excluding the blending process. We devise a three-stage framework to generate speech from 3D face shapes. Fully considering timbre production in anatomical and acquired terms, our framework incorporates three additional relevant attributes including face texture, facial features, and demographics. Experiments and subjective tests demonstrate our method can generate utterances matching faces well, with good audio quality and voice diversity. We also explore and visualize how the voice changes with the face. Case studies show that our method upgrades the face-voice inference to personalized custom-made voice creating, revealing a promising prospect in virtual human and dubbing applications.

Introduction

With the development of talking head techniques, virtual humans have shown tremendous potential in virtual anchors, virtual idols, and other fields. TTS (text-to-speech) systems are usually employed to produce a voice for a virtual human. Current methods of TTS can generate a natural voice comparable to the voice of a real person. However, it is required to record a large-scale real human voice corpus and train the model for days to generate a proper voice matching the virtual character. If it is possible to automatically infer the voice characteristic of a virtual human and then synthesize the speech with it, it will be easy to adapt TTS models without time-consuming training to produce personalized voices.

Therefore, it is necessary to explore the prospect of speech synthesis with faces. Though lots of previous works pay attention to face-voice correlations, only a few of them spend effort on face-voice generation. Among them, Face2Speech (Goto et al. 2020) replaces the speaker embedding with the face embedding in multi-speaker TTS. FaceVC (Lu et al. 2021) introduces a three-stage training strategy to fit face embedding to speaker embedding in voice conversion systems. However, all of their works only consider 2D face images involving the face pose, expression, lighting, and other factors irrelevant to voice characteristics. Besides, since face images are hardly editable, it is also difficult to modify them to get different voices and explain how and to what extent voice is influenced by face.

In this paper, we aim to exploit face-voice inference by introducing 3D face shapes to explain the face-voice correlation and predict and manipulate the voice afterward, because face shape always figures personality and matters to dubbing work (Mirza and Osindero 2014). For example, a superman with a large chin and prominent brow ridges usually has a powerful and deep voice. 3D face-voice inference is innovative, reasonable and meaningful. Anatomically, it is the skull shape, but not the 2D face image, that influences the voice. Because "bone conduction" is an important factor of timbre production, such as cheekbones and chins (Maurer and Landis 1990). The bones also influence the growth of vocal organs. Technically, the 3D approach (Deng et al. 2019) disentangles the speaker's identity from irrelevant factors, such as lighting, pose, and expression. The entangling problem is not discussed and addressed in existing papers. By introducing the 3D approach, unrelated factors are excluded.

To fully explore cause of formation of human voice, we also consider more factors influencing voice characteristics besides face shape. For example, wheatish skin is related to good health and loose skin is related to aging; gold-framed glasses and beards also make different impressions. Although these factors are not necessarily associated with voice characteristics in real life, voice actors are required to balance all appearance factors to choose a proper voice type for a character (Smith et al. 2016). Therefore, besides the 3D face shape, we supplement three additional voice-related factors: face texture, facial features, and demographics. The face texture represents soft tissues from the aspect of anatomy; the facial features, such as glasses and beards, are related to personality and speaking habits; demographics, including gender and age, also matter in voice characteristics.

We propose a multi-modal framework to address this task, taking these factors as input, outputting the corresponding voice characteristics, and synthesizing the speech at the end.

*Corresponding authors: Z. Wu , J. Jia

First, we employ a 3D face constructor to restore face shape and face texture from a video. A texture encoder compresses face texture into appearance features. We utilize an attribute extractor to derive the facial features and demographic labels. All features are concatenated and sent to a voice encoder to predict the eigenvalue of speaker embedding, which contains speaker information compressed by a speaker encoder. The predicted eigenvalue is restored to speaker embedding by an inverse PCA (Principal Component Analysis) process and sent to a multi-speaker text-to-speech model to generate the speech with proper voice characteristics.

We implement our framework on two datasets and conduct experiments. Objective comparison and subjective evaluations suggest that the voice our model produced qualifies for matching the speaker's appearance, along with high audio quality and voice diversity, outperforming the method considering 2D face image merely.

Further case studies provide a possible approach to voice editing by face interpolation and principal component editing. Visualizations reveal how faces influence voices. For example, face width causes different effects on pitch in speech of males and females.

We summarize our contributions as follows:

- We propose a task generating speech based on 3D face shape, making face-voice relation anatomically explainable and controllable from deep learning models.
- We establish a practicable pipeline from 3D face shape and other additional inputs to speaker embedding, and finally to speech. Our generated audio achieves a good level of audio quality, voice diversity, and face-voice matching degree.
- We propose two possible approaches to voice editing and visualizations demonstrate how faces influence voices.

Related Work

Face Voice Correlation

Faces and voices are highly related to personal identities. The voice is one kind of 'auditory face' on the concept. In cognitive science experiments, participants can predict voice characteristics by giving the faces of speakers (Kamachi et al. 2003; Lachs and Pisoni 2004), and the matching accuracy is significantly above chosen level, suggesting that faces and voices offer overlapping or complementary information about a person (Smith et al. 2016), such as gender, age, height, and weight. In neuroscience research, speaker cognition can result from the information sharing between auditory voice and visual face regions (Kriegstein et al. 2005).

Speech-face-associated learning has been widely studied. Face height and head length can predict the vocal tract structures (Vorperian et al. 1999), and deep learning methods also indicate the relationship between face and voice, such as predicting faces from voices (Oh et al. 2019) and matching faces and voices (Horiguchi, Kanda, and Nagamatsu 2018; Mavica and Barenholtz 2013).

3D Face Model Reconstruction

Current 3D face reconstruction can recover fine face geometric shapes from 2D face images. Pretty many methods need to reconstruct 3D faces from different inputs, from multi-view images (Cao et al. 2018b) to a single image (Li et al. 2018; Hassner 2013; Riviere et al. 2020). Different models also restore 3D faces from different priors. Some works determine a statistical face model previously and predict or analyze the coefficients (Tu et al. 2019; Chang et al. 2018; Thies et al. 2016; Ploumpis et al. 2020). These methods take advantage of fast speed, but with the limitation of fixed shape space, producing smooth results.

Several methods predict 3D face vertexes of meshes instead, and model more detailed information of faces (Wei, Liang, and Wei 2019; Feng et al. 2018; Jackson et al. 2017). Nevertheless, these methods call for explicit 3D supervision provided by other models and therefore capture coarse shape information.

Face animation is taken into consideration for detailed reconstruction. For example, DECA (Feng et al. 2021) enables animation and relighting by capturing high-fidelity textures. It predicts expression, pose, and other parameters from one image and generates a UV color map from the person-related low-dimensional representation. Deep3DFace (Deng et al. 2019) employ the 3D MorphableModel (3DMM) (Blanz and Vetter 1999) and train an CNN model to predict the corresponding coefficients. It also models the expression for accurate reconstruction.

Face-based Speech Synthesis

Recent TTS systems are competent to synthesize realistic and natural speech (Wang et al. 2017; Shen et al. 2018; Ren et al. 2019; Li et al. 2019). These years, multi-speaker TTS is developed to generate the voices of different speakers by one system (Arik et al. 2017; Jia et al. 2018; Park et al. 2019; Cooper et al. 2020). These models usually implement the multi-speaker function with a speaker encoder trained by a speaker verification task (Snyder et al. 2018a; Wan et al. 2018). Speaker embedding vectors are extracted from the hidden layer of the speaker encoder and concatenated to the hidden states of the TTS encoder.

Based on multi-speaker TTS, face-based TTS models are proposed by replacing speaker embedding vectors with face embedding vectors. As far as we know, Face2Speech (Goto et al. 2020) is the first model to generate speech from face images. Face2Speech trains a VGG (Simonyan and Zisserman 2014) model to fit face embedding vectors to speaker embedding vectors, and apply the face embedding to Tacotron2 (Shen et al. 2018). Several works are published following this method (Wu et al. 2022; Plüster et al. 2021; Wang et al. 2022), introducing sophisticated modules and detailed experiments to promote the generated voice quality. FaceVC (Lu et al. 2021) introduces a similar approach in voice conversion, but with a three-stage training strategy. A face encoder and a speaker encoder are trained separately, and face embedding space is mapped to speaker embedding space by introducing a Visual-to-Audio Transformation module. Using this approach, FaceVC achieves good audio quality and speaker similarity.

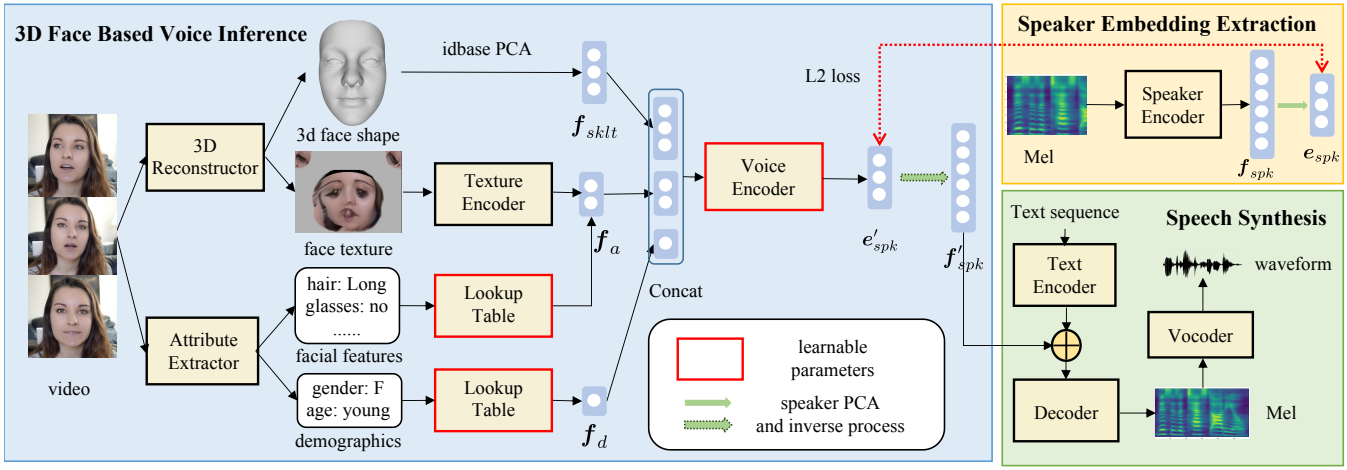

Figure 1: The pipeline of our model. We extract the 3D face shape, face texture, facial features, and demographics from the speaker's face video. These factors are sent to a voice encoder to generate the speaker embedding. The generated speaker embedding and the text are then sent to a multi-speaker speech synthesis module to get the synthetic speech with proper voice characteristics. The voice encoder is trained with L2 loss between the generated speaker embedding coefficients and the ground truth extracted by the speaker embedding extraction module.

However, none of them generate speech from 3D face shapes.

Methodology

We first select four kinds of voice-related speaker characteristics including 3D face as our input according to sound production and attributes related to voice characteristics. Then we implement a 3D face-based voice inference method unifying inputs with different dimensions to predict the voice embedding vector of the speaker. Finally, we introduce a three-stage training strategy to complete the pipeline of speaker embedding extraction, voice inference, and speech synthesis, as shown in Figure 1.

Voice-Related Factors

As introduced in Section 1, we select four voice-related factors to contrive the input of our model to predict voice characteristics. In this section, we give explanations of these features and introduce how we extract them.

- **3D face shape** is the 3D model of a face, composed of vertexes in 3D space. We adopt 3D face shape as the representation of anatomical voice-related skeleton.
- **Face texture** is the texture map of the 3D face shape, mapped as a 2D image. We utilize face texture to represent soft tissues and muscles that affect sound production.
- **Facial features**. We employ facial features as part of the personality and character of a speaker. Our selected facial features include: *eyeglasses*, *hair color*, *beard*, *mustache*, *hat* and *smile*.
- **Demographics** stand for different groups of speakers with different voice patterns, including *gender* and *age*.

We leverage a 3D reconstructor and an attribute extractor to get these features. First, we adopt the 3DMM (Blanz and Vetter 1999) face model to represent each 3D face. The face shape $\mathbf{S}$ is the weighted sum of face skeleton and face expression:

$$\mathbf{S} = \mathbf{B}_{sklt} f_{sklt} + \mathbf{B}_{exp} f_{exp} + \bar{\mathbf{S}} \quad (1)$$

where $\mathbf{B}_{sklt}$ and $\mathbf{B}_{exp}$ are the PCA bases of speaker identity and expression; f_{sklt} and f_{exp} are the coefficients of the PCAs respectively; $\bar{\mathbf{S}} \in \mathbb{R}^{N \cdot 3}$ is the average face shape and N is the number of the vertexes in the face model. We adopt the 2009 Base Face Model (Paysan et al. 2009) for $\bar{\mathbf{S}}$, and use the expression bases $\mathbf{B}_{exp}$ of (Guo et al. 2018). Only a subset of the bases and coefficients are in use, with $f_{sklt} \in \mathbb{R}^{80}$ and $f_{exp} \in \mathbb{R}^{64}$.

The face texture $\mathbf{T}$ is represented by an affine model:

$$\mathbf{T} = \mathbf{B}_t \delta + \bar{\mathbf{T}} \quad (2)$$

where $\mathbf{B}_t$ is the PCA base of face texture and $\delta \in \mathbb{R}^{80}$ is the corresponding coefficient vector. Then we convert $\mathbf{T}$ from a linear space into FLAME (Li et al. 2017) layout to output a UV albedo map $\mathbf{A} \in \mathbb{R}^{d \cdot d \cdot 3}$.

We adopt an attribute extractor with CNN architecture (Hernandez 2021) to infer the facial features L_f and demographics L_d from images.

3D Face Based Voice Inference

As described in the previous section, we obtain the four inputs: 3D face shape $\mathbf{S}$, face texture $\mathbf{T}$, facial features L_f, and demographics L_d. We adopt different methods to convert these inputs with different dimensions into a unified embedding space.

As for the 3D face shape, we select the skeleton coefficient f_{sklt} as representation, disentangled with the interference of expression and the average face. As for the face

texture, we adopt a texture encoder with VGG19 architecture (Simonyan and Zisserman 2014) and send the texture albedo map $\mathbf{A}$ as input to it. We extract the hidden layer output of the texture encoder as the representation of face texture. We utilize lookup tables to embed facial features and demographics to embedding vectors. We concatenate the embedding vector of facial features with the hidden state of the texture encoder to form the appearance feature f_a, representing the appearance information. The embedding vector of demographics forms demographic feature f_d.

To infer the voice from these input, we choose speaker embedding f_{spk} as the representation of the speaker's voice characteristics and the output of our model. Instead of predicting f_{spk} directly, we first apply a PCA to f_{spk} to retain the principal components e_{spk}, reducing the feature dimension for predicting simplicity.

$$f_{spk} = \mathbf{B}_{spk} e_{spk} \tag{3}$$

where $\mathbf{B}_{spk}$ is the PCA base applied to the speaker embedding vector f_{spk}. All extracted features are concatenated together and sent to a voice encoder to predict the principal components e'_{spk}.

$$e'_{spk} = F_v([f_{sklt}; f_a; f_d]) \tag{4}$$

where F_v is the voice encoder. Then we restore the speaker embedding f'_{spk} from e'_{spk} by the inverse process of PCA applied to the speaker embedding.

$$f'_{spk} = \mathbf{B}_{spk} e'_{spk} \tag{5}$$

During the inference stage, the restored speaker embedding f'_{spk} is sent to a multi-speaker TTS system to generate the utterance with corresponding voice characteristics.

Three Stage Training Strategy

We employ a three-stage training strategy to complete the pipeline of voice extraction, inference, and synthesis. Firstly, we employ the pretrained DNN structure XVectors (Snyder et al. 2018b) as our speaker encoder to extract the speaker embedding. It takes 24-dimensional filterbanks feature as input, and we extract 512-dimensional embedding vector f_{spk} from *segment7* as the speaker embedding vector.

Secondly, for the speech synthesis module, we adopt the Conformer-FastSpeech2 (Guo et al. 2020) as our TTS system. A text sequence t is converted to a phoneme sequence and encoded into a hidden sequence h.

$$h = Encoder(t) \tag{6}$$

The speaker embedding vector is replicated and added to the hidden sequence.

$$h' = h + f_{spk} \tag{7}$$

Finally, a decoder convert h' into Mel-spectrogram, and a vocoder restore the waveform afterward.

$$Mel' = Decoder(h') \tag{8}$$

L2 loss between the predicted and ground truth Mel-spectrograms is used to train the TTS system.

Thirdly, for the training of the voice inference module, we adopt L2 loss between e'_{spk} and e_{spk} to learn the parameters of the voice encoder and lookup tables:

$$\mathcal{L} = ||e_{spk} - e'_{spk}||_2^2 \tag{9}$$

Experiment

Dataset

Dataset Construction. The first part of the dataset comes from VoxCeleb2 (Chung, Nagrani, and Zisserman 2018) and VGGFace2 (Cao et al. 2018a). VoxCeleb2 is a video dataset including more than 6,000 celebrities, and VGGFace2 contains face images of these persons. Additionally, we utilize ChaLearn LAP (Ponce-López et al. 2016) video dataset, containing videos of more than 30,000 clips.

Dataset Partition. Our final face-speech pair dataset consists of 5,995 speakers from VoxCeleb2-VGGFace2 and 2,624 speakers from ChaLearn LAP, after filtering the speakers failing in extracting speaker embedding. We split the first 1,200 and 525 speakers off VoxCeleb2 and ChaLearn LAP for validation, remaining 4,795 and 2,009 speakers for training respectively.

Implementation Details

We adopt an open-source pretrained speaker encoder XXvectors from Kaldi[1]. We set PCA_s to retain 95% variation with $e_{spk} \in \mathbb{R}^{59}$. We employ a pretrained Attribute Extractor[2]. Each attribute is embedded to 3 dimension. The texture encoder is a pretrained VGG-19 model (Simonyan and Zisserman 2014). Our Conformer-Fastspeech2 is a pretrained version[3]. We implement our voice encoder with an MLP with 3 Linear layers. Each layer is followed by a ReLU activation and Dropout Layer except the last one.

We train our model on an NVIDIA Geforce 2080 Ti for 50 epochs, with a batch size of 64. We adopt the Adam optimizer with a learning rate of 0.002.

Evaluation Method

Comparison Systems. We compare the performance of our model with some alternative systems to synthesize speech.

- **synth-speech**, i.e. ground truth. We directly apply the speaker embedding provided by the speaker encoder to the multi-speaker TTS.
- **synth-face**. The speaker embedding of synth-face is predicted by a pretrained and finetuned VGG-19.

Metrics. We utilize the following metrics to evaluate models.

- **MSE**. We adopt *MSE* error to compare the ground truth embedding from the ground truth audio and the embedding reconstructed.
- **Speaker Similarity**. We calculate the cosine score of speaker embedding extracted from ground truth audio, i.e. the ground truth embedding, and our generated ones.
- **MOS** (mean opinion score). We use the mean opinion score (MOS) to evaluate the degree of satisfaction of users in terms of audio quality, voice characteristic diversity, and face-voice matching degree respectively. For

[1] http://kaldi-asr.org/models/m3

[2] https://github.com/buenohernandez/Face-detection-feature-extraction

[3] https://github.com/espnet/espnet

input	VoxCeleb2		Chalearn LAP	
	MSE (↓)	Sim (↑)	MSE (↓)	Sim (↑)
synth-face	1.86	0.7558	1.82	0.7752
ours	**1.50**	**0.8128**	**1.52**	**0.8169**

Table 1: MSE and Sim(speaker similarity) of speaker embedding vectors on the validation dataset.

audio quality and voice diversity, the range of MOS is 1-5 with 1 point interval, the higher the better. For the face-voice matching degree, the range of MOS is 1-4 without decimal scores by following the setting of (Chen et al. 2017), the lower the better. Totally 24 evaluators participated in the subjective tests.

Comparison with Baselines

To compare our method with other systems, we train these models on the ChaLearn LAP dataset. We list the MSE and speaker similarities of speaker embedding vectors given by different models on validation datasets in Table 1. Our model outperforms the synth-face system on both datasets in these two indexes. This result indicates that our model produces more accurate embedding vectors than synth-face, with a significance p-value less than 0.05.

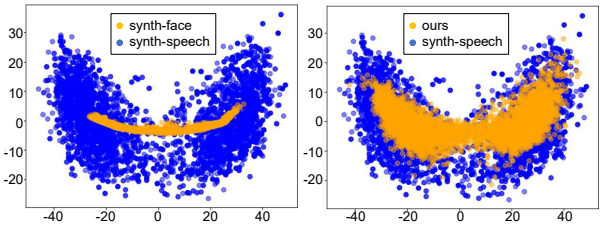

Figure 2: The tSNE visualization of speaker embedding vectors of synth-speech, synth-face and ours on ChaLearn LAP validation dataset.

Obeying the same setting, we select validation speakers and visualize the predictions of each model by tSNE. From Figure 2, synth-speech has a wild distribution, while synth-face fails to fit the ground truth, and produces vague predictions between the two gender clusters. Our method successfully fits the ground truth embedding vectors with distinguishable clusters. But for some marginal cases on the borderlines of clusters, it is also difficult for our model to give precise predictions.

User Study

To examine the generation quality, we carried out the subjective test to assess the *Audio Quality*, *Voice Diversity* and *Matching Score* of the speech synthesized by the generated speaker embedding vectors following the description of MOS.

Audio Quality. We adapt this index to evaluate the noise level of the generated speech. Participants are asked to score given audio from 1 to 5, i.e., from very bad to very good.

Embedding	MOS		
	Audio Quality (↑)	Voice Diversity (↑)	Matching Score (↓)
synth-speech	3.42 ± 0.09	$\mathbf{3.72 \pm 0.20}$	$\underline{1.49 \pm 0.09}$
synth-face	$\underline{3.45 \pm 0.07}$	3.05 ± 0.18	1.64 ± 0.09
ours	$\mathbf{3.50 \pm 0.06}$	$\underline{3.21 \pm 0.18}$	$\mathbf{1.39 \pm 0.07}$

Table 2: The results of subjective tests. Scores are presented with 95% confidence intervals.

As shown in Table 2, all MOS scores are above 3, indicating generated utterances are acceptable. Our model achieves a higher level than synth-face, even better than synth-speech because our model tends to produce a smooth and averaged embedding, causing fewer jitters and noise than synth-speech in generated audio. Whereas synth-face produces several over-smooth embedding vectors, for example, locating between the male distribution and the female distribution, resulting in difficulty in TTS model synthesizing.

Voice Diversity. We carry out this subjective test to examine to what extent our model learns the real distribution of human voice characteristics. Evaluators are asked to judge whether a group of generated utterances have a voice diversity comparable to human ones. In detail, good audio samples should cover as many voice characteristics as possible, dull or clear, soft or powerful. Samples are rated from 1 to 5, from homogeneous or unreal to various. Results are in Table 2, it is clear that voices from real humans are of the widest distribution. Our model still surpasses synth-face obviously, showing the same results in Figure 2. However, there also exists a large margin between our generated voices and real ones, demonstrating the limitation that face-based models are only able to produce averaged and approximate voice types and hardly give perfect predictions.

Matching Score. The matching score is set from 1 to 4 following (Chen et al. 2017): from matching well to not match, measuring whether generated voices match the faces. Note that a lower match score is better. Subjects are provided with face images and corresponding generated utterances. The results shown in Table 2 indicate that our generated samples match speaker faces moderately, slightly better than ones of synth-speech. This is possibly caused by smoothed predictions in a few cases, resulting in a voice without jitter and shimmer, slightly more convincing than synth-speech ones.

Visualization

We adopt tSNE visualization to observe the distribution of generated speaker embedding vectors. In Figure 3, although without tight clusters, predictions scatter around the ground truth, with the same color meaning speaker embedding vectors from the same speaker. Besides the two clusters of male and female, there exist other distribution features. For example, voices located at the top tend to be deeper, while voices at the bottom are probably softer and brighter.

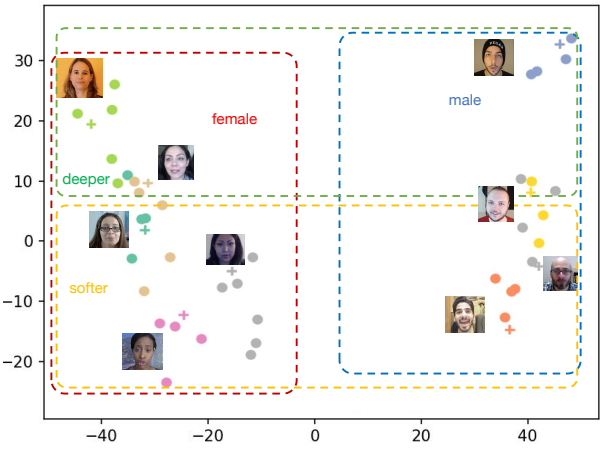

Figure 3: Speaker embedding tSNE visualization of ground truth and generated ones. '+' represents ground truth embedding vectors. '·' represents our generated embedding vectors. Speaker's face is placed nearby. Box with colored dotted lines represents distribution of certain voice type.

Component	MSE (↓)	Speaker Similarity (↑)
ours	1.52	0.8169
w/o face shape	2.12	0.7356
w/o texture	1.54	0.8165
w/o attributes	1.53	0.8190
w/o speaker PCA	1.54	0.8162

Table 3: MSE and speaker similarity of speaker embedding vectors on ChaLearn LAP validation dataset.

Ablation Study

Although we emphasize the importance of 3D face shape, more factors are taken into consideration for completeness, such as face texture. We implement an ablation study to test the performance of our model when each component is absent to examine the contributions of each component to voice prediction. The results on ChaLearn LAP validation dataset are shown as Table 3.

From Table 3, face shape input is certainly the most significant factor to predict voices, confirming our claim. Face texture and attributes (facial attributes and demographics) do not disturb the results too much. It is interesting to compare the performance of 'w/o face shape' with synth-face in Table 1. The former is inferior to the latter, indicating the speaker information within face textures is less than original face images. Despite the cost of information loss in face texture, the 3D face shape obtains more information in the 3D face reconstruction process. Speaker embedding PCA is also not very necessary in terms of performance but helps to explain how face shape influence the principle components of voice characteristics, elaborated later in the following section.

Correlation Analysis

To demonstrate the correlation between 3D face and voice, we visualize the correlation coefficients between 3D face shape embedding f_{sklt} and reduced speaker embedding e_{spk}, as shown in Figure 4. For simplicity, we only demonstrate the first 20 dimensions of face embedding vectors and the first 30 dimensions of speaker embedding vectors after PCA. Empirically, the 1^{st} component of the 3D face shape vector and reduced speaker embedding vector is gender. The gender of the face not only influences the gender of voice but also influences other components in speaker embedding by high correlation coefficients. The phenomenon indicates that gender is the most important factor in the voice. Other face components also influence the voice components, but with a weak correlation.

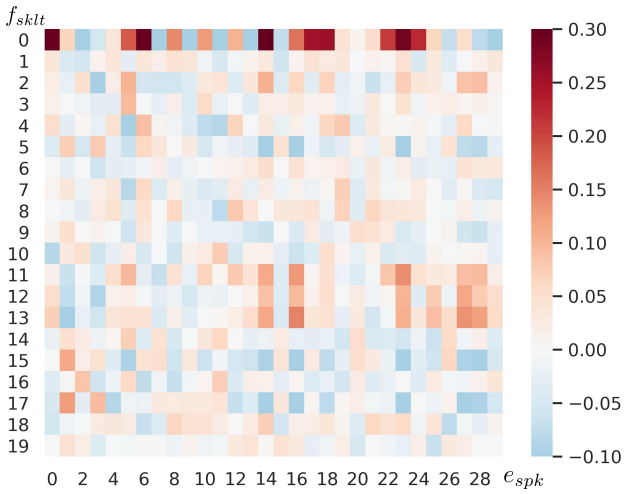

Figure 4: The correlation coefficients between 3D face shape embedding vectors and reduced speaker embedding vectors.

Case Study

We carry out two case studies to intuitively demonstrate the relationship between face shape coefficients and the generated voice. We introduce two indexes to measure the differences between voices: $\boldsymbol{F}_0$ **contour** is the base frequency of an utterance in the time domain; **Spectral Centroid** of the mel-spectrogram represents the center of the harmonic distribution.

Principal Component Disturbance. Since the face shape feature we extract is the eigenvalue after PCA, we can modify different principal components to observe the difference in voices. We select one speaker and edit the first principal component of his face shape eigenvalues. As the results shown in Figure 5, the original 3D face and the predicted voice is in the middle, and the modified face shapes and the generated speech spectrograms are on each side. From left to right, the face changes from female to male, and the voice changes accordingly. The pitch gradually decreases, and the centroid shifts from bottom right to top left, representing the

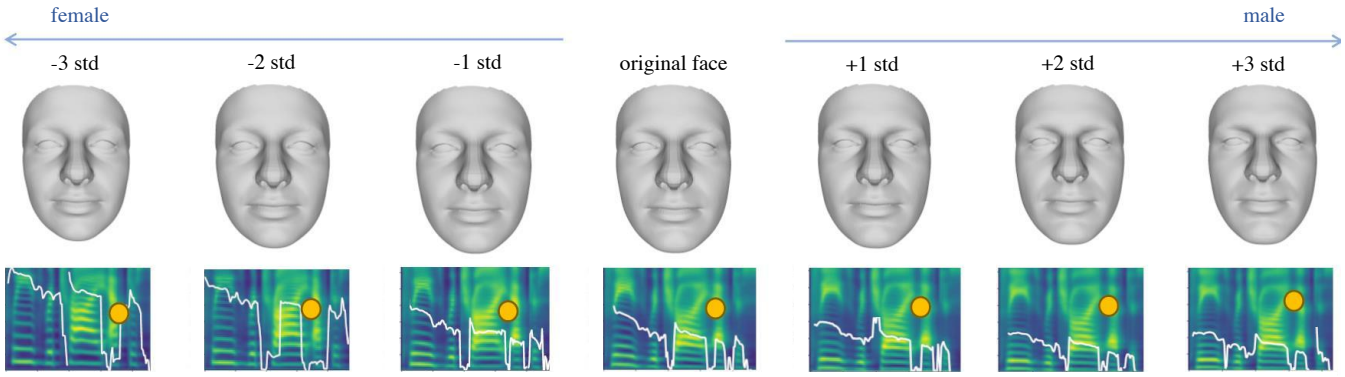

Figure 5: The mel-spectrograms of voices generated by the 1^{st} principal component edited face shapes. The white curves denote F_0 contours. The orange dots denote centroids of the mels. The input text is "*Please retry after several hours.*"

shift of the tone and resonance. Modification of the 2^{st} triggers changes in the resonance, with the case shown in the supplement material.

Interpolation. First, we train our model with only 3D face shapes as input, eliminating the interference of other factors. Then we select two speakers, A and B, and average their face shape eigenvalues with weight λ (0-1). We set λ as $0, 0.3, 0.5, 0.7, 1$ and visualize the averaged face shape and the corresponding generated audio. Figure 6 shows the results of females and males. We do not exhibit the interpolation results of two genders, though which generate acceptable utterances.

For the voices of females in Figure 6, from left to right, the cheekbones appear more prominent, the chin gets stronger, and the age also grows elder. Voice type transformation is corresponding to face shape changes. From left to right, the voice type changes from brighter to duller, the range of F_0 gradually gets lower, and resonance also shifts, with a loss of high-frequency resonance. Similar phenomena also appear in the voices of males. With the face getting longer and more masculine, the voice gets deeper and the vowel patterns in spectrograms also change.

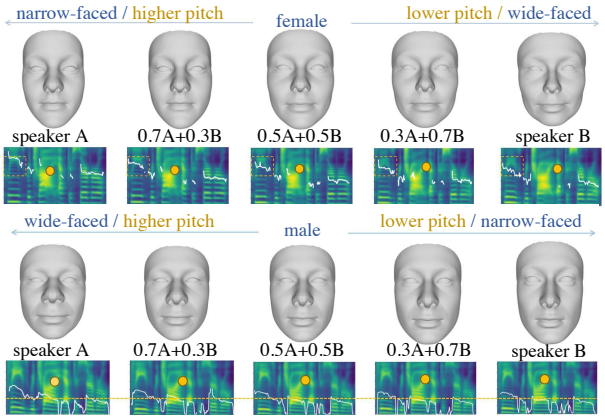

Figure 6: The mel-spectrograms of voice generated by mixed face shapes. The first row is female, and the second row is male. The white curves denote F_0 contours. The orange dots denote centroids of the mels. The input text is "*Please retry after several hours.*"

Ethic Consideration and Error Analysis

Approximation. In this paper, we introduce 3D face shapes to generate speaker embedding. Face shapes are related to the vocal tract structures, but there does not necessarily exist a causal link between voices and faces. This predictive approach can only provide an approximate and compatible solution of voice characteristics, predicting average voices rather than exact voices of individuals. The physical meaning of 3D face and voice factors still remains to be explored.

Data Bias. The datasets we use are from English native speakers on Youtube. Most speakers are from a specific age group. There exists a difference between the datasets and the real world. The populations of datasets can not cover all the communities.

Annotation Error. The attribute labels and the face shapes we utilize are mostly annotated by the pretrained model we employed. Thus there exists errors between the real faces and the ones we reconstructed. We can only try our best to give the explanation of face-voice correlation and make predictions with the given restrictive conditions.

Conclusion

In this paper, we propose the problem of generating speech by face shapes. We introduce a voice inference and speech synthesis framework using 3D face shapes. Compared with 2D face images, 3D face shapes are editable and independent of expression, pose, and other irrelevant factors. Experiments and subjective tests demonstrate our method can generate utterances matching faces well. Voice editing is practicable through face interpolation and 3D face coefficient modification. Further case studies explore how the voice changes with the face. For example, aging causes high-frequency resonance loss. Our method shows a promising prospect in virtual human and dubbing applications.

Acknowledgments

This work is supported by National Key Research and Development Plan (2021QY1500), National Natural Science Foundation of China (NSFC) (62076144), the joint research fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK40415), Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004) and AMiner.Shenzhen SciBrain fund.

References

Arik, S.; Diamos, G.; Gibiansky, A.; Miller, J.; Peng, K.; Ping, W.; Raiman, J.; and Zhou, Y. 2017. Deep voice 2: Multi-speaker neural text-to-speech. *arXiv preprint arXiv:1705.08947*.

Blanz, V.; and Vetter, T. 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.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018a. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Cao, X.; Chen, Z.; Chen, A.; Chen, X.; Li, S.; and Yu, J. 2018b. Sparse photometric 3D face reconstruction guided by morphable models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4635–4644.

Chang, F.-J.; Tran, A. T.; Hassner, T.; Masi, I.; Nevatia, R.; and Medioni, G. 2018. Expnet: Landmark-free, deep, 3d facial expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 122–129. IEEE.

Chen, L.; Srivastava, S.; Duan, Z.; and Xu, C. 2017. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 349–357.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Cooper, E.; Lai, C.-I.; Yasuda, Y.; Fang, F.; Wang, X.; Chen, N.; and Yamagishi, J. 2020. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6184–6188. IEEE.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4): 1–13.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, 534–551.

Goto, S.; Onishi, K.; Saito, Y.; Tachibana, K.; and Mori, K. 2020. Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image. *Proc. Interspeech 2020*, 1321–1325.

Guo, P.; Boyer, F.; Chang, X.; Hayashi, T.; Higuchi, Y.; Inaguma, H.; Kamo, N.; Li, C.; Garcia-Romero, D.; Shi, J.; et al. 2020. Recent Developments on ESPnet Toolkit Boosted by Conformer. *arXiv preprint arXiv:2010.13956*.

Guo, Y.; Cai, J.; Jiang, B.; Zheng, J.; et al. 2018. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6): 1294–1307.

Hassner, T. 2013. Viewing real-world faces in 3D. In *Proceedings of the IEEE International Conference on Computer Vision*, 3607–3614.

Hernandez, F. 2021. Face-detection-feature-extraction. https://github.com/buenohernandez/Face-detection-feature-extraction. Accessed: 2023-04-18.

Horiguchi, S.; Kanda, N.; and Nagamatsu, K. 2018. Face-voice matching using cross-modal embeddings. In *Proceedings of the 26th ACM international conference on Multimedia*, 1011–1019.

Jackson, A. S.; Bulat, A.; Argyriou, V.; and Tzimiropoulos, G. 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the IEEE international conference on computer vision*, 1031–1039.

Jia, Y.; Zhang, Y.; Weiss, R. J.; Wang, Q.; Shen, J.; Ren, F.; Chen, Z.; Nguyen, P.; Pang, R.; Moreno, I. L.; et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv preprint arXiv:1806.04558*.

Kamachi, M.; Hill, H.; Lander, K.; and Vatikiotis-Bateson, E. 2003. Putting the face to the voice': Matching identity across modality. *Current Biology*, 13(19): 1709–1714.

Kriegstein, K. v.; Kleinschmidt, A.; Sterzer, P.; and Giraud, A.-L. 2005. Interaction of face and voice areas during speaker recognition. *Journal of cognitive neuroscience*, 17(3): 367–376.

Lachs, L.; and Pisoni, D. B. 2004. Crossmodal source identification in speech perception. *Ecological Psychology*, 16(3): 159–187.

Li, N.; Liu, S.; Liu, Y.; Zhao, S.; and Liu, M. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6706–6713.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.

Li, Y.; Ma, L.; Fan, H.; and Mitchell, K. 2018. Feature-preserving detailed 3d face reconstruction from a single image. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, 1–9.

Lu, H.-H.; Weng, S.-E.; Yen, Y.-F.; Shuai, H.-H.; and Cheng, W.-H. 2021. Face-based Voice Conversion: Learning the Voice behind a Face. In *Proceedings of the 29th ACM International Conference on Multimedia*, 496–505.

Maurer, D.; and Landis, T. 1990. Role of bone conduction in the self-perception of speech. *Folia phoniatrica*, 42(5): 226–229.

Mavica, L. W.; and Barenholtz, E. 2013. Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance*, 39(2): 307.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Oh, T.-H.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; and Matusik, W. 2019. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7539–7548.

Park, J.; Zhao, K.; Peng, K.; and Ping, W. 2019. Multi-speaker end-to-end speech synthesis. *arXiv preprint arXiv:1907.04462*.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, 296–301. Ieee.

Ploumpis, S.; Ververas, E.; O'Sullivan, E.; Moschoglou, S.; Wang, H.; Pears, N.; Smith, W. A.; Gecer, B.; and Zafeiriou, S. 2020. Towards a complete 3D morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 4142–4160.

Plüster, B.; Weber, C.; Qu, L.; and Wermter, S. 2021. Hearing Faces: Target Speaker Text-to-Speech Synthesis from a Face. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 757–764. IEEE.

Ponce-López, V.; Chen, B.; Oliu, M.; Corneanu, C.; Clapés, A.; Guyon, I.; Baró, X.; Escalante, H. J.; and Escalera, S. 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European conference on computer vision*, 400–418. Springer.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*.

Riviere, J.; Gotardo, P. F. U.; Bradley, D.; Ghosh, A.; and Beeler, T. 2020. Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans. Graph.*, 39(4): 81.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779–4783. IEEE.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, H. M.; Dunn, A. K.; Baguley, T.; and Stacey, P. C. 2016. Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology*, 14(1): 1474704916630317.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018a. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018b. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333. IEEE.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2387–2395.

Tu, X.; Zhao, J.; Jiang, Z.; Luo, Y.; Xie, M.; Zhao, Y.; He, L.; Ma, Z.; and Feng, J. 2019. Joint 3d face reconstruction and dense face alignment from a single image with 2d-assisted self-supervised learning. *arXiv preprint arXiv:1903.09359*, 1(2).

Vorperian, H. K.; Kent, R. D.; Gentry, L. R.; and Yandell, B. S. 1999. Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: preliminary results. *International Journal of Pediatric Otorhinolaryngology*, 49(3): 197–206.

Wan, L.; Wang, Q.; Papir, A.; and Moreno, I. L. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4879–4883. IEEE.

Wang, J.; Wang, Z.; Hu, X.; Li, X.; Fang, Q.; and Liu, L. 2022. Residual-Guided Personalized Speech Synthesis based on Face Image. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4743–4747. IEEE.

Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Wei, H.; Liang, S.; and Wei, Y. 2019. 3d dense face alignment via graph convolution networks. *arXiv preprint arXiv:1904.05562*.

Wu, X.; Ji, S.; Wang, J.; and Guo, Y. 2022. Speech synthesis with face embeddings. *Applied Intelligence*, 1–14.