Applied Intelligence (2022) 52:14839–14852
https://doi.org/10.1007/s10489-022-03227-7

Speech synthesis with face embeddings

Xing Wu[1,2] · Sihui Ji[1] · Jianjia Wang[1,2] · Yike Guo[3,4]

Accepted: 9 January 2022 / Published online: 18 March 2022

Abstract

Human beings are capable of imagining a person's voice according to his or her appearance because different people have different voice characteristics. Although researchers have made great progress in single-view speech synthesis, there are few studies on multi-view speech synthesis, especially the speech synthesis using face images. On the basis of implicit relationship between the speaker's face image and his or her voice, we propose a multi-view speech synthesis method called SSFE (Speech Synthesis with Face Embeddings). The proposed SSFE consists of three parts: a voice encoder, a face encoder and an improved multi-speaker text-to-speech (TTS) engine. On the one hand, the proposed voice encoder generates the voice embeddings from the speaker's speech and the proposed face encoder extracts the voice features from the speaker's face as f-voice embeddings. On the other hand, the multi-speaker TTS engine would synthesize the speech with voice embeddings and f-voice embeddings. We have conducted extensive experiments to evaluate the proposed SSFE on the synthesized speech quality and face-voice matching degree, in which the Mean Opinion Score of the SSFE is more than 3.7 and the matching degree is about 1.7. The experimental results prove that the proposed SSFE method outperforms state-of-the-art methods on the synthesized speech in terms of speech quality and face-voice matching degree.

Keywords Multi-view speech synthesis · Face to voice · Visual-audio · Multi-speaker text-to-speech

Introduction Machine learning has been widely studied and applied in medical, entertainment, finance, security, and other fields [1–4]. With the explosive growth of data, there are a large number of various types of non-linear data. Multi-view learning takes multiple modal data sources or multiple feature representations as input, in which the information of multiple perspectives will complement each other, to improve the performance of feature representation. In general, multi-modal data includes audio, image, text, video, and etc. Multiple features refer to features extracted from different modal data or different levels. Multi-view learning makes full use of the complementary information of these data or features and achieves great success in the field of machine learning and artificial intelligence, which has attracted the attention of researchers.

The face image and voice of the same person can be regarded as two data modalities with the same identity attribute. As we all have a common feeling that there is a connection between a person's face and voice. Although the association is not completely one-to-one correspondence, we can imagine a person's face through his voice to some extent and vice versa. From the perspective of face structure, the distribution of facial tissues, the shape, size, and position of skeletal muscles, and the vocal organs' acoustic characteristics determine the vocal-tract of sound production [5]. Neuroscientists also demonstrated a neurological relationship between faces and voices: human voice and face share a common neuro-cognitive pathway structure [6].

Based on the proven relationship between the speaker's face image and voice, we focus on speech synthesis tasks with face embeddings. Specifically, given a person's face image to produce a voice that sounds consistent with the person's identity. It is noteworthy that we are not to clone the speaker's voice completely but to learn the underlying associations between voices and faces to produce a voice that matches the speaker's facial features as much as possible.

✉ Xing Wu
xingwu@shu.edu.cn

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China

[2] Shanghai Institute for Advanced Communication and Data Science, Shanghai, China

[3] Department of Computing, Imperial College London, London, United Kingdom

[4] Hong Kong Baptist University, Kowloon Tong, Hong Kong

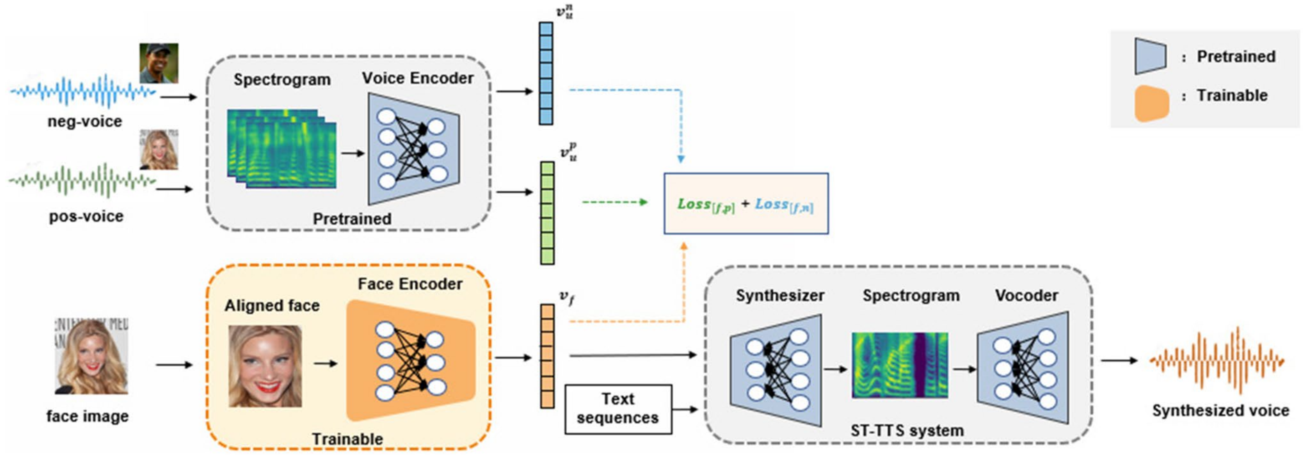


Fig. 1 The training pipeline of the SSFE framework

The key of this work is to learn the common feature representation containing both the speaker's face and voice information and further synthesize the speech through the TTS system. The research can be used for video dubbing, voice simulation of speech-impaired people, and voice reconstruction of people who have passed away.

The speech synthesis with face embeddings is a two-stage task, in which the first stage extracts voice features from speaker's faces and the second stage converts features into speech through Text-to-Speech (TTS). TTS is a technique that produces a speech from given text. The similarity and naturalness of synthetic speech are key indicators for evaluating the TTS system. However, many traditional TTS models only learned an averaged prosodic distribution of the speaker's corpus in the modeling of prosody, so it is difficult to learn the dynamic characteristics of the voice. The speech generated from the end-to-end architecture proposed in [7] could be highly similar to the human voice. While the speech generated by the model sometimes had sudden silent fragments, which cannot break sentences well. The modeling of prosody could improve the fluency and naturalness of pronunciation to some extent. As proposed in [8], style tokens capture the speaker's style features from different dimensions and have been proven to be effective at synthesizing expressive and long-form sentences.

We propose an end-to-end cross-modal speech synthesis framework with face embeddings (SSFE), as shown in Fig. 1. The whole framework is divided into three independent training modules: voice encoder, face encoder, and multi-speaker TTS system. The role of the voice encoder is to guide the learning of the face encoder. During training, the voice encoder takes the spectrogram of the speaker's speech as input and outputs a voice embedding. When training the face encoder, the voice encoder is fixed. The face encoder learns the feature representation containing both the speaker's face and voice information under the guidance of the voice embedding, denoted as f-voice embedding. Take the voice embeddings of the same person obtained from the voice encoder as positive samples; and the voice embeddings of other speakers in the same batch as negative samples. The distance between the f-voice embedding and the positive samples is as close as possible, and the distance between the f-voice embedding and the negative samples is as far as possible to guide the feature extraction process of the face encoder. The TTS system takes text sequences and voice embeddings as input and outputs synthesized speech during training. We extended the end-to-end TTS model by adding the style token layer to further capture style and prosody information. In the inference stage, we input a face image into the face encoder to extract the f-voice embedding. The speech is then synthesized from the given text sequence and the f-voice embedding.

We evaluate the model in terms of speech quality, speech similarity, face-voice matching degree and robustness. The results show that the speech generated from our SSFE framework matches the face well, and the speech quality is comparable to the speech synthesized using the voice embeddings extracted from the voice encoder. The main contributions of this work can be summarized as follows:

- A two-stage training strategy is adopted proposed for the voice encoder to speed up the extraction of speech features.
- A Style Token based Text-To-Speech (ST-TTS) system is proposed to extract the style features of voices and they are fused with the traditional voice features to obtain more expressive voice representation.
- A tailor-designed domain adaption method is proposed, which the proposed SSFE could have a common representation of voice and face feature spaces.

The remaining chapters are organized as follows: In Section 2, we present related work on multi-view learning and speech synthesis. Section 3 is the focus of this paper. We elaborate the pipeline and training settings of the SSFE framework in this part. Section 4 shows the improvement effect of each module through sufficient experiments and gives the specific adjustment processes. In addition, we give the evaluation results of the model from three perspectives of speech quality, speech similarity, and face-voice matching degree, and give the comparison with the current advanced methods. Finally, we summarize the current work and look into the future work in Section 5.

1 Related work

1.1 Multi-view learning

Multi-view learning In contrast to single-view learning, multi-view learning extracts interrelated and complementary information from multi-view data, which is beneficial to explain the essential characteristics of the research object and improve learning performance. Therefore, multi-view learning has attracted more and more attention. One of the earliest and representative research achievements of multi-view learning is Canonical Correlation Analysis (CCA) [9], which is a statistical method to search linear mapping of two eigenvectors.

In recent years, multi-view learning has achieved success in many fields. To solve the problem of insufficient multi-task shared class label set, a multi-task and multi-view clustering algorithm based on local linear embedding under heterogeneous conditions was proposed in [10]. Wu et al. [11] achieved excellent results in the recognition of multi-modal mixed character CAPTCHA, which is composed of English letters, Arabic numerals, Chinese characters, and mathematical operators. Zhang et al. [12] proposed a multi-view clustering model based on a non-negative matrix for multi-stage factor analysis of Alzheimer's disease. Zhou et al. [13] proposed using different depth information as a supplement to RGB information to obtain more expressive features. The depth information is extracted and fused at different levels, and then fused with RGB features. In [14], they also fused RGB and thermal modal features to realize semantic segmentation of urban scenes and achieved excellent performance. In the field of cross-modal retrieval, an effective cross-modal retrieval method MS2GAN is proposed [15]. This method can jointly extract and utilize both the modality-specific and modality-shared features effectively. In face recognition, Fei Wu et al.[16] improved multi-spectral face recognition by considering both spectrum and class label information. The proposed approach significantly outperforms state-of-the-art multi-spectral face recognition methods. Like our multi-view speech synthesis work, there are some studies on spiking neural networks that are biologically inspired neural networks [17, 18], and these studies provide a bridge between biological learning and machine learning. In addition, the idea of multi-view learning was also integrated with ensemble learning [19].

Cross-modal audio-visual research There has been an increasing amount of literature on audio-visual research. Tamura, S. et al. [20] built a cross-modal audio-visual voice conversion model. Hori, C. [21] used teacher-student networks to research the audio-visual scene-aware dialog. Tae-Hyun et al. [22] proposed a three dependently trained Speech2Face framework to generate a face image from a given speech. There are some studies about generating face from audio using Generative Adversarial Network [23]. Nagrani et al. [24] conducted a biometric matching between faces and voices. Experiments confirmed that there is indeed biological information associated with human faces and voices. Some scholars have committed to audio-visual matching researches for speech recognition [25]. There are some lip-reading tasks from video to speech [26, 27] and some image2speech studies on image description [28]. Ryan Jenkins et al. [29] demonstrated that those with excellent facial memory and matching ability performed better in speech matching and memory. This proves that cross-modal face and voice and cross-task mechanism of memory and perception drive superior performance.

1.2 Speech synthesis

Speech synthesis on single-view learning The earliest method is statistical parametric speech synthesis (SPSS) appeared in the late 1990s. It uses a statistical generation model to learn the relationship between the calculated features on the input text and the output acoustic features. With the development of deep learning, the TTS systems based on DNN have emerged [30] and WaveNet [31] and Tacotron [32] were proposed and breakthroughs were made in the TTS field. WaveRNN was proposed in [33] to improve the slow speed of WaveNet.

Some researchers have begun to conduct the end-to-end training of the TTS model. Sotelo j et al. [34] trained TTS models directly from $< text, audio >$ pairs without hand-made intermediate representations. Luong, H. [35] introduced an "unsupervised speaker adaptation" method using a small amount of speech data for speech synthesis. Tacotron2 [36] used WaveNet as a vocoder to reverse the spectrogram generated by the encoder-decoder structure with an attention mechanism. Morita, T.et al. [37] proposed an end-to-end unsupervised TTS system without text. Zhang et al. [38] proposed an unsupervised pre-training mechanism that could be more effectively applied to low-resource language synthesis.

Speech synthesis on multi-view learning Although research on audio-visual cross-modal tasks has been increasing, there are still few cross-modal speech synthesis research, especially speech synthesis using face images. Beskow and Jonas [39] developed a multi-modal speech synthesis system, which could synthesize audio-visual animation from an arbitrary text by using parameter-controlled face and head models. Goto, S. et al. [40] proposed a three-stage cross-modal TTS framework, which used supervised GE2E loss, a measure based on cosine distance, to guide the face encoder.

Style and prosody model There are several approaches to style and prosody modeling that have been studied, such as the DNN-based speech synthesis model [41] although it requires explicit labels. Cluster-based unsupervised modeling method [42] relies on the features of manual design. After that, the concept of reference embedding was introduced in [43]. Wang et al. Guided attention [44] could accelerate the training speed by adding prior knowledge. Namely, there is a linear relationship between the position of each word and the moment of pronouncing it when we read the text.

2 Method

2.1 Problem formulation

Our SSFE framework consists of three modules, as shown in Fig. 1: 1) the voice encoder, taking spectrograms as input and voice embeddings as output; 2) the face encoder, extracting the f-voice embeddings from face images; and 3) the ST-TTS system to synthesize the speech from $< voice\, embedding, text\, sequence >$ pairs. When training the face encoder, the voice encoder and ST-TTS are pretrained. We will give the definition of the SSFE framework in accordance with these three modules.

During training, the speaker's audio and face are simultaneously input into the voice encoder and face encoder, respectively. The information flows through the models as follows:

The role of the **voice encoder** in the SSFE framework is to extract the discriminative voice features from a given speaker's speech. During training, it takes a sequence of log-Mel spectrogram frames from an arbitrary length ground-truth utterance as input, and maps them to a fixed d-length voice embedding. The model computes the utterance embedding as v_u for each utterance $u_{ji}(1 \leq j \leq N, 1 \leq i \leq M)$, with a total of N speakers and M utterances per speaker:

$$\mathcal{M}_1 : u_{ji} \rightarrow v_u \tag{1}$$

where $\mathcal{M}_1$ represents the map from utterance u_{ji} to embedding v_u. The speaker embedding $v_s \in \mathbb{R}^{d \times L}$ of speech length L is defined as the mean of M utterance embeddings:

$$v_s = \frac{1}{M} \sum_{i=1}^{M} v_{u_i} \tag{2}$$

At the same time, the **face encoder** learns the feature representation from the speaker's face. It takes input as face images $f_{ji}(1 \leq j \leq N, 1 \leq i \leq M)$, which corresponding the j-th speaker's utterances u_{ji} and outputs the f-voice embeddings v_f:

$$\mathcal{M}_2 : f_{ji} \rightarrow v_f \tag{3}$$

where $\mathcal{M}_2$ represents the map from the i-th face of the j-th speaker f_{ji} to face embedding v_f.

The voice embeddings extracted from the same identity are denoted as pos-voice embeddings v_u^p, and the voice embeddings with different identity are denoted as neg-voice embeddings v_u^n. For the domain adaption of face images and voices, the f-voice embedding v_f extracted from the face image is expected to be close to the v_u^p and far away from the voice embeddings v_u^n in a mini-batch as much as possible.

After that, the learned v_f and the specified text sequence will be fed into the ST-TTS system for speech synthesis. The ST-TTS system is composed of two independently trained neural networks: (1) a style token based synthesizer, which takes a $< voice\, embedding, text\, sequence >$ pair as input and synthesizes a log-Mel spectrogram mel as output. (2) a WaveRNN based vocoder, which converts synthesized log-Mel spectrogram mel into synthetic speech sp:

$$\begin{aligned} \mathcal{M}_3 &: < voice\, embedding, text\, sequence > \rightarrow mel \\ \mathcal{M}_4 &: \quad mel \rightarrow sp \end{aligned} \tag{4}$$

where $\mathcal{M}_3$ and $\mathcal{M}_4$ represents the map of two networks respectively.

Specifically, the style token based synthesizer is pretrained with the voice embeddings extracted from the voice encoder. During training, the f-voice embedding v_f derived from the face encoder is replaced by the speaker embedding v_s derived from the voice encoder. The module takes pairs of $< voice\, embedding, text\, sequence >$ as input and outputs the synthesized speech, as shown in Fig. 2.

The speaker embedding v_s is fed into the **style token** layer before being input to the synthesizer. The module measure the similarity between the v_s and K style tokens $T_k(1 \leq k \leq K)$ using the multi-head attention. The output $S_i \in \mathbb{R}^{d \times L}$ is a weighted style-token combination as shown in (5), which then is used to condition the Mel spectrogram synthesis of text sequences.

$$v_{sty} = \sum_{k=1}^{K} w_k \cdot MultiHead(v_s, T_k) \tag{5}$$

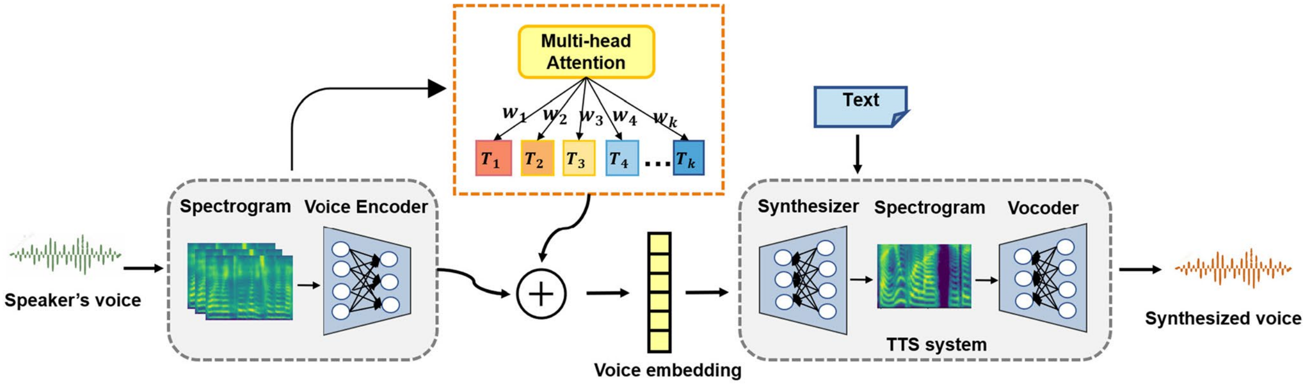


Fig. 2 The training pipeline of the ST-TTS system

The speaker embedding v_s and the style embedding v_{sty} will be concatenated and text sequence will be passed through synthesizer to predict a log-Mel spectrogram *mel*.

$$mel = Synth(text, Concat(v_s, v_{sty})) \tag{6}$$

In the synthesizer, the text sequence is encoded as text embedding t. The attention mechanism acts as a bridge between the encoder and the decoder to learn the alignment information between voice features and text features. Finally, the spectrogram is decoded by the decoder.

The WaveRNN is regarded as our vocoder to complete the signal conversion from frequency domain to time domain. The model replaces the entire 60 convolutions from WaveNet as a single GRU layer and improves the computing speed than WaveNet. During training, it takes the ground truth aligned Mel spectrogram in a batch generated by the synthesizer as the input, and the ground truth audio as the target.

$$sp = Voc(mel) \tag{7}$$

During inference, the synthesized *mel* will be inverted to time-domain waveform, namely synthesized speech *sp*.

2.2 Training settings and model details

2.2.1 Voice encoder

The goal of the voice encoder is to extract the discriminative voice features from a given speaker's speech. The network is composed of 3 LSTM layers, which takes 40-channel log-Mel spectrograms as input and outputs 256-dimensional embeddings as characteristics of different speakers. Different from [7], we increase the hidden size from 256 to 768 and replace ReLU activation with Tanh before the L2-normalization. The model is optimized by a GE2E loss. The similarity comparison against the correct speaker's own embedding is removed to reduce the bias. The similarity matrix $\mathcal{S}_{ji,k}$ is defined as:

$$\mathcal{S}_{ji,k} = \begin{cases} w \cdot cos(v_u, v_s^{(-i)}) + b & if \quad j = k \\ w \cdot cos(v_u, v_s) + b & if \quad j \neq k \end{cases} \tag{8}$$

where w and b are learnable weight and bias parameters and $j = k$ represents utterances of speaker j matches the k-th speaker embedding (1<=k<=N). For the j-th speaker, the exclusive speaker embeddings $v_s^{(-i)}$ are defined as:

$$v_s^{(-i)} = \frac{1}{M-1} \sum_{m=1,m\neq i}^{M} v_{u_m} \tag{9}$$

Then, each embedding v_u is optimized to get as close to the speaker embedding $v_s^{(-i)}$ as possible. The voice embeddings extracted from voices are used to guide the learning of f-voice embeddings, to realize the domain adaption of voice and face features.

2.2.2 Face encoder

The face encoder learns the feature representation containing both the speaker's face and voice information under the guidance of the voice embedding. Specifically, to reduce the gap between face and voice, the pre-trained voice encoder described in Section 3.2.1 is used to extract voice embeddings. Take this feature as the learning goal of the face encoder, the feature representation containing both face and voice features are learned by minimizing the loss between the face embeddings extracted by the face encoder and the voice embeddings extracted by the voice encoder. At the same time, to make the extracted voice features reflect the differences of voice features from different facial structures, it is necessary to make the distance between the learned voice features and the voice features of other speakers extracted by the voice encoder as large as possible.

Inception-ResNet-v1 is regarded as the architecture of our face encoder. To make the network learn the features of faces more accurately, we use MTCNN[1], a pre-trained face detection

[1] https://github.com/davidsandberg/facenet/tree/master/src/align

model, to extract the face parts from raw images and resize them to 160 × 160 before the image is input into the network. The face encoder takes aligned face images of speakers as input and outputs v_f with the size of 256 that equals v_u extracted from the voice encoder. The dropout is set to 0.8 on the fully connected layer. To further speed up the calculation, we randomly select 40 images and 40 voices for each speaker in a mini-batch rather than all face images and audios.

As mentioned in Section 3.1, the f-voice embedding v_f is expected to be closer to all voice embeddings v_u^p of the same identity and far away from the voice embeddings v_u^n extracted from other speakers in a mini-batch. To speed up convergence, for each generated v_f, we select hard $< v_u^p, v_u^n >$ pairs that satisfy:

$$\| v_f - v_u^n \|_2^2 - \| v_f - v_u^p \|_2^2 > \alpha \tag{10}$$

where α is a hyper-parameter restricts the v_u^p and v_u^n in a fixed margin. This constraint means that when the distance between v_f and v_u^p exceeds the distance between v_f and v_u^n by α, the contribution of face image feature v_f to training is small and can be discarded. Training only those v_u^p and v_u^n that are most difficult to distinguish from v_f can accelerate the convergence of the model. Then, the L_2 distance measure is applied between the feature embeddings. The model is optimized by minimize the loss function defined as:

$$\mathcal{L} = \sum_{j}^{N} [\| v_f - v_u^p \|_2^2 - \| v_f - v_u^n \|_2^2 + \alpha]_+ \tag{11}$$

where N is the number of speaker in a mini-batch. We select $< v_f, v_u^p, v_u^n >$ triplets from each mini-batch with size of 90, and speakers per batch is set to 45. The v_f learned by face encoder contains both the face and voice information, and then could be input to the ST-TTS system to synthesize the speech with specified text sequences.

2.2.3 ST-TTS system

The ST-TTS system consists of a synthesizer and a vocoder for speech synthesis. During training, the input of the synthesizer are pairs of $<$ *voice embedding*, *text sequence* $>$. For arbitrary length text sequences, each character is encoded as 512-dimension.

The synthesizer uses Tactron2 as the basic architecture and the style token module is added to enhance the ability of extracting voice features. Specifically, the voice features extracted from the voice encoder and the style features extracted from the style token module are fused through concatenation to obtain voice representation containing richer information. As shown in Fig. 2, the module consists of N random initialized style token embeddings $T = T_1, T_2, T_3, ..., T_N$ to capture rich style dimensions in speaker embeddings. We apply a multi-head attention with h heads after random initialized style tokens to learn the style embedding, a weighted combination of tokens. In our experiments, we set $N = 10, h = 4$ and each token embedding is 512/h, so that the style embedding is 512-D after the concatenation of 4 heads attention output. The style token module is jointly trained with the synthesizer, and no any other loss based on this is introduced.

The 256-dimensional voice embedding is concatenated with text embedding at each timestep. The dimension of attention space is set to 128 and batch size is 64. The Adam optimizer is used with $\beta1 = 0.9$, $\beta2 = 0.999$, $\epsilon = 1 \times 10^{-6}$. For the synthesizer, the initial learning rate is 1×10^{-3} and the final learning rate is 1×10^{-5} with a decay rate equaling 0.5. When training the vocoder, we use raw audio and ground truth aligned Mel spectrograms generated from the synthesizer as input to learn time-domain waveforms.

To better illustrate the training procedure of the whole SSFE framework, the pseudo code is given in Algorithm 1:

Algorithm 1 The training pipeline of SSFE framework.

```
Input: reference speaker's utterance: $utt$,reference speaker's face image: $face$, text sequence: $text$ and model parameter: $\Theta_1$
Output: synthesized speech: $speech$.
Initialize model parameter $\Theta_1$.
for $u, f$ in epoch do:
    Preprocess $utt$ to spectrogram: $u \leftarrow utt$.
    Preprocess $face$ to aligned-face: $f \leftarrow face$.
end for
for mini-batch in epoch do:
    for $u, f$ in mini-batch do:
        Compute the voice embedding using voice encoder: $v_u \leftarrow u$.
        Select hard $(v_u^p,\ v_u^n)$ pairs that satisfy (10).
        Learn the f-voice embedding using face encoder: $v_f \leftarrow f$, under the guidence of
        (11).
        Pass through the style-token layer: $v_{sty} \leftarrow v_f$.
        Synthesize the log-Mel spectrogram: $mel \leftarrow (text, Concat(v_f, v_{sty}))$.
        Converts synthesized $mel$ to time-domain waveform: $speech \leftarrow mel$.
    end for
end for
```

Table 1 Datasets in different modules of the SSFE framework

dataset	type	voice encoder	face encoder	ST-TTS
LibriTTS	audio	√(other)		√(clean)
Common-Voice	audio	√		
VoxCeleb1	audio,image	√(audio)		
VoxCeleb2	audio,image	√(audio)	√(audio)	
VGGFace2	image		√	
GRID	audio,video		√	

3 Experiments

3.1 Datasets

The three modules of our SSFE framework are trained by different datasets, as shown in Table 1. According to the performance reported in [7], LibriTTS-other [45], VoxCeleb1 [24] and VoxCeleb2 [46] are used for training the voice encoder and LibriTTS-clean is used for ST-TTS.

The LibriTTS dataset, rather than the LibriSpeech dataset, is a large-scale corpus comprising approximately 585 hours of speech and consists of the union of the “clean” and “other” sets, in which the “clean” set contains cleaner speech than the “other” set. It is added in the encoder training process first. This is because the LibriTTS dataset is a purer version of the LibriSpeech dataset, with a sampling rate of 24kHz, and the sentence contains punctuation marks not found in the Librispeech dataset, which helps to learn more natural prosodic information. When the loss of model training is stable below 0.01, CommonVoice [47], VoxCeleb1 [24] and VoxCeleb2 [46] datasets were added. CommonVoice dataset is an English dataset with 60K speakers. A large amount of training data can greatly improve the feature extraction capability of the encoder model. Both VoxCeleb1 and VoxCeleb2 are derived from celebrity videos on YouTube. VoxCeleb1 has 1251 speakers for about 150k utterances, while VoxCeleb2 has 6112 speakers for over 1128k utterances. All datasets are sampled at 16kHz. Noisy audio datasets VoxCeleb1 and VoxCeleb2 were added in order to further improve the robustness of the model.

In the training process of the synthesizer and vocoder, we both use the train-clean-100 and train-clean-360 parts of the LibriTTS dataset. The reason for using “clean” parts is because the quality of the dataset affects the quality of the synthesized audio.

For the training of the face encoder, we carry out experiments on different datasets. On the one hand, we compare the proposed method with existing cross-modal speech synthesis methods in terms of speech quality and gender recognition accuracy. Consistent with the setup of the comparison methods proposed in [26, 27], we train and test the model on the Grid dataset. Grid dataset is an audio and video pair dataset composed of 33 speakers, with 1000 utterances per speaker. Each sentence is composed of command + color + preposition + letter + digit + advertisement. Two male speakers s1, s2, and two female speakers s4, s29 were selected to perform a test task of speaker dependence. Each speaker is divided into training, validation, and test sets in a ratio of $90\% : 5\% : 5\%$.

On the other hand, to compare the matching degree between speech and face synthesized by face feature, we prepare $< image, audio >$ pairs, in which the images are derived from the VGGFace2 dataset [48] and the audios from VoxCeleb2 dataset. VGGFace2 is also extracted from YouTube videos of celebrities and contains all speakers of VoxCeleb2. To match identities with the VoxCeleb2, we select the intersection of the VGGFace2 dataset and VoxCeleb2 dataset to train the face encoder. For all 6,112 speakers in the intersection, 5,994 speakers are used for training and 118 speakers for evaluation. This setting is consistent with [40].

3.2 Results

3.2.1 Voice encoder

The purpose of the first part of our experiment is to obtain a voice encoder model to not only extract sound features accurately, but also to converge faster. In [7], an internal dataset is used to train the voice encoder. Because we cannot obtain the internal dataset, we add the CommonVoice dataset to train the voice encoder.

In addition, we propose a two-stage training strategy. In the beginning, we used the “other” part of LibriTTS for training. When the loss of the model stabilized below 0.01 (about 200k-250k steps), we add CommonVoice, VoxCeleb1, and VoxCeleb2 datasets. When the model is trained to about 800k steps, the ReLU activation function of the last layer of lstm is changed to the Tanh activation function, and the loss gradually decreases. When it reaches 1 million steps, the model gradually tends to converge.

As shown in Fig. 3, the UMAP dimension reduction results of the speaker embeddings training at 1k, 5k, 20k, 50k, 1M and 1.2M steps are respectively shown in subfigures. Different colors indicate different speakers, and 10 utterances are randomly selected for each speaker. We can see that as the training step increases, the different utterances of the same speaker are more and more closely clustered together, and the distance between clusters of different speakers is getting larger and larger. It indicates that our voice encoder can recognize the voices of different speakers. Namely, it has learned the characteristics of different speakers’ voices.

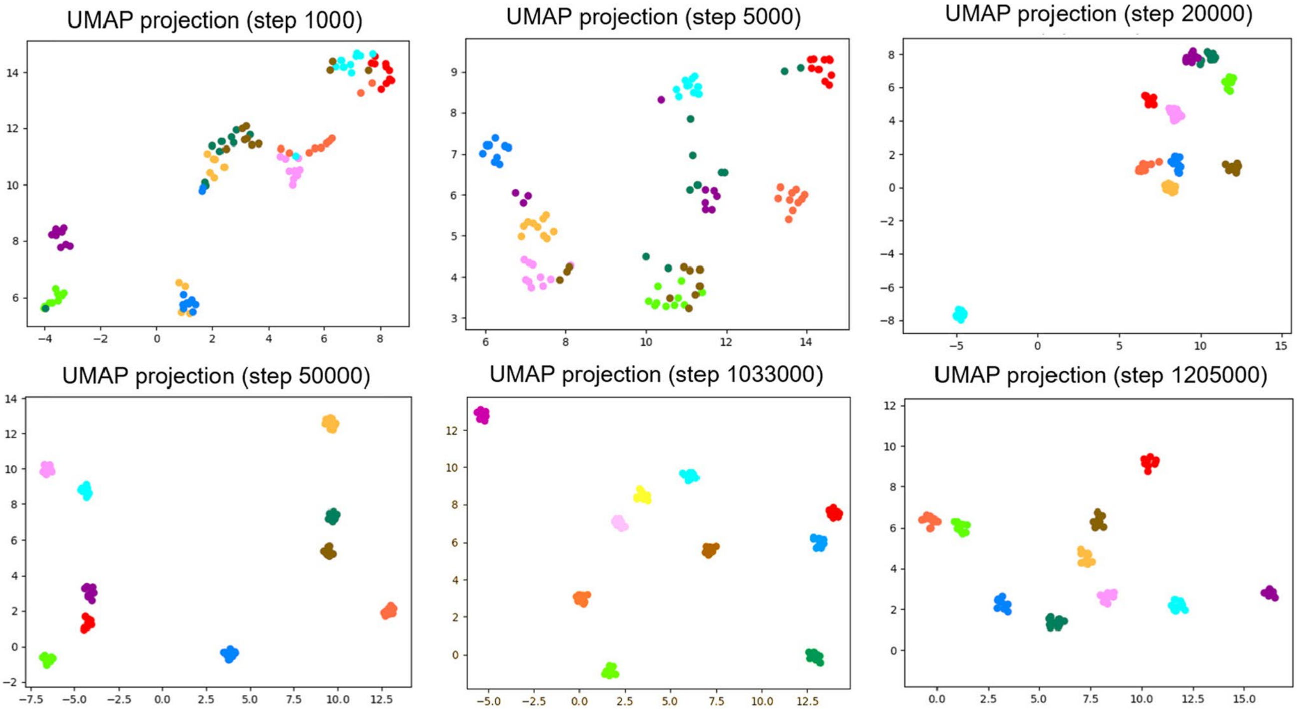


Fig. 3 UMAP visualizations of the voice encoder at different training steps

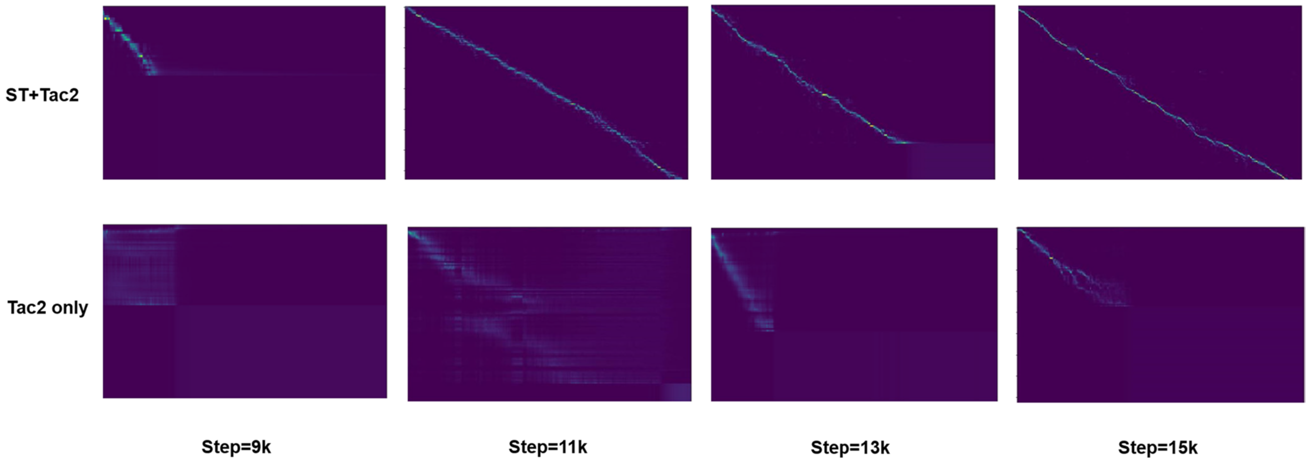


Fig. 4 Visualization of learning effect of attention alignment information in the synthesizer

3.2.2 ST-TTS system

In this section, we will measure the performance of the style token based synthesizer. We use "Tac2" to indicate the Tacotron2 based synthesizer, and "ST" to indicate the style tokens. Therefore, "Tac2 only" means that the voice features learned from the voice encoder are used directly to synthesize the spectrogram; and "ST+Tac2" means that the voice features that input to the synthesizer contain the style information extracted from the style token module. In Tacotron2, attention is used as a bridging mechanism between the encoder and the decoder, reflecting the synthesizer's ability to align pairs $< \textit{voice features}, \textit{text features} >$. The better the alignment information is learned, the more diagonal the curve is.

As shown in Fig. 4, the first row represents the ability to learn attention alignment information in the "ST+Tac2" mode. The model can learn alignment information from about step 9k, and can learn a clear curve at steps 11k, 13k, and 15k. The learning curve is diagonal, indicating that attention has learned the alignment information correctly. In the “Tac2 only” mode, the speed of attention learning is significantly slower than in the “ST+Tac2” mode. It can be seen that the model only learns a little alignment information at step 15k, and the learned curve is not clear or even divergent. Therefore, it can be proved that “ST+Tac2” can learn the alignment information faster and more effectively than the “Tac2 only” mode. Finally, the Mel spectrogram predicted by the synthesizer is shown in Fig. 5.

3.2.3 Face encoder

Comparison of algorithm complexity For Face2Speech, VGG-19 is adopted as the backbone of the face encoder. In the SSFE framework, we consider Inception-ResNet-v1 or Inception-ResNet-v2 as the face encoder. The Floating Point Operations (FLOPs) are often used to measure the time complexity of an algorithm and the number of model parameters are used to measure the space complexity. Taking these three networks as the backbone of the face encoder respectively, the calculation results of FLOPs and the number of parameter are shown in Table 2. As we can see from Table 2, both the FLOPs and the number of parameters of Inception-Resnet networks are less than VGG-19. Compared with Inception-Resnet-v2, Inception-ResNet-v1 has lower time and space complexity. Based on above, we consider using Inception- Resnet-v1 or Inception- Resnet-v2 as the backbone of our face encoder.

Table 2 Comparisons of algorithm complexity on different models

Framework	Input	FLOPs	Params
Inception-ResNet-v1	160 × 160 × 3	240M	23.0M
Inception-ResNet-v2	160 × 160 × 3	564M	54.7M
VGG-19	160 × 160 × 3	20654M	143.7M

Comparison of clustering effects of the f-voice embeddings We compare the clustering effects of the f-voice embeddings when Inception-ResNet-v1 and Inception-ResNet-v2 are used to conduct experiments under the same experimental conditions. It’ s reported that Inception-ResNet-v1 has a smaller model structure, and the classification accuracy is lower than Inception-ResNet-v2. To determine the effects of the two models in our experiments, we use them to train the face encoder respectively. The UMAP dimension reduction visualization results of the feature embeddings extracted from the two models are shown in the first and the second column of Fig. 6, where the UMAP distance matrix is the Euclidean distance.

As can be seen from Fig. 6(a), the voice features extracted from the same speaker by Inception-ResNet-v1, represented with the same color, are almost gathered together, and the voice embeddings of different genders are significantly divided into male and female clusters. For Inception-ResNet-v2, as showed in Fig. 6(b), most of the extracted voice features can be correctly divided into two clusters by gender. However, there are individual speakers that are misclassified, such as n004029. Compared with the Inception-ResNet-v1 model, the overlap of voice feature locations is more significant among different speakers of the same gender. As shown

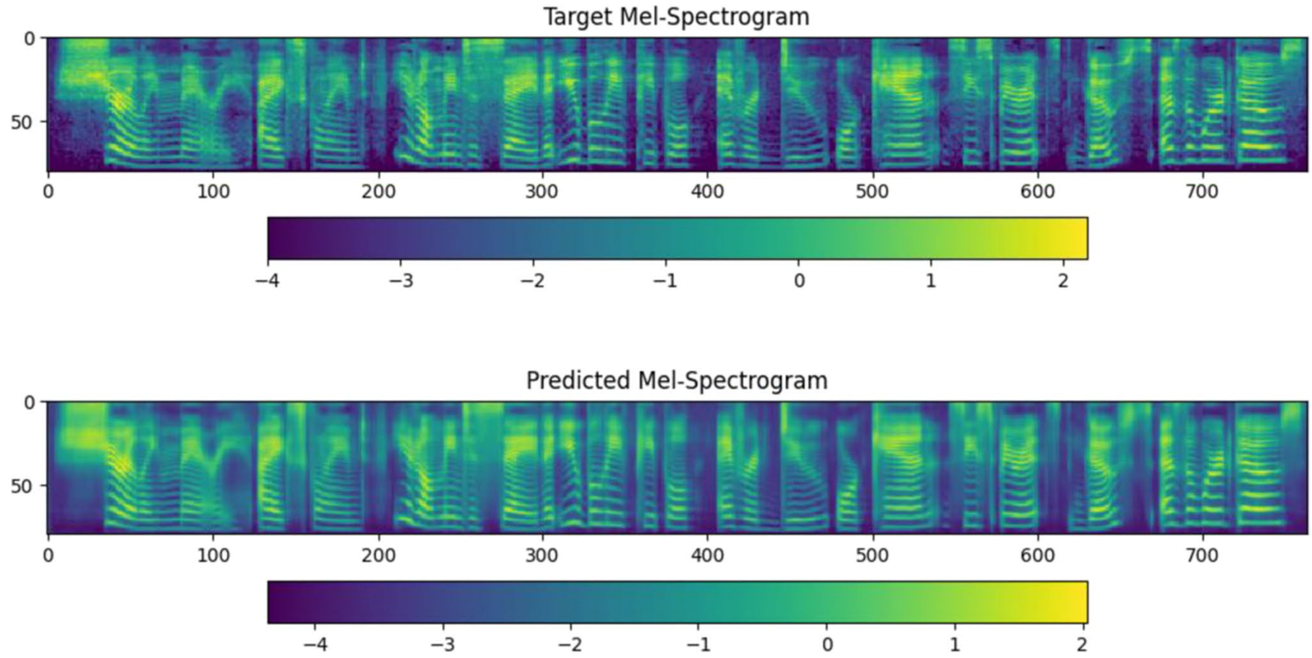
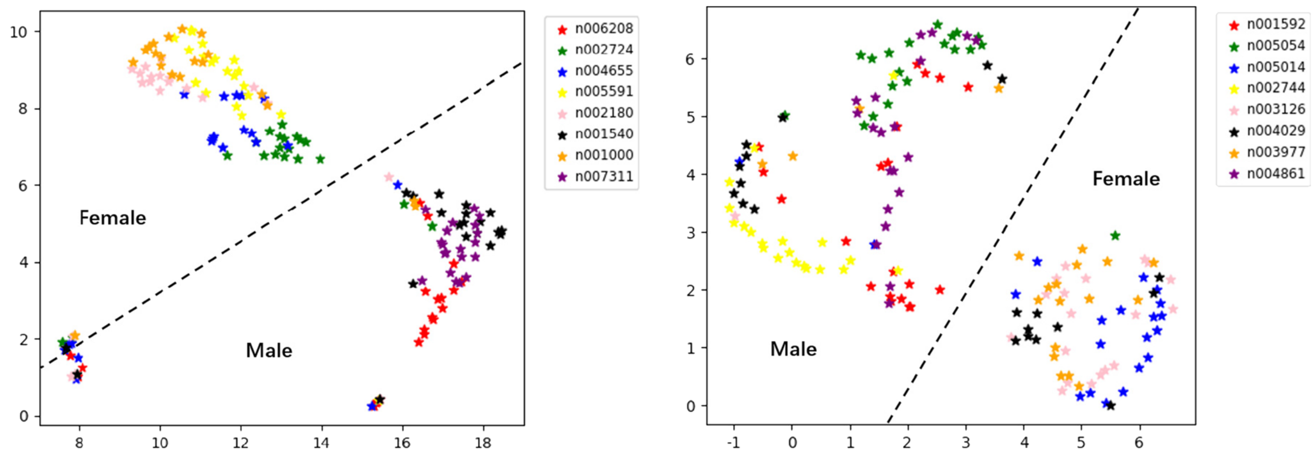


Fig. 5 The target Mel-Spectrogram and predicted Mel-Spectrogram predicted by style token layer based synthesizer



(a) Inception-ResNet-v1 (b) Inception-ResNet-v2

Fig. 6 UMAP visualization of the f-voice embeddings extracted from the face encoder, which trained with (a) Inception-ResNet-v1 and (b) Inception-ResNet-v2, respectively

Table 3 MOS and Gender recognition evaluation results of the proposed and reference models

Framework	MOS	Gender recognition
Video2Speech [26]	1.35	43.2%
Lip2AudSpec [27]	1.63	85.1%
SSFE(ours)	3.94	87.5%

Table 4 Matching scores and preference scores with 95% confidence intervals

Framework	Matching Score		Preference Score	
	SF	SV	SF	SV
Face2Speech [40]	2.01 ± 0.07	1.91 ± 0.06	0.548 ± 0.049	0.452 ± 0.049
SSFE(ours)	1.70 ± 0.114	1.61 ± 0.129	0.577 ± 0.090	0.433 ± 0.090

that different voice characteristics from the same speaker have larger intraclass distances and more dispersed distributions. In addition, considering that the algorithm reported in Table 2, we choose Inception-ResNet-v1 as the backbone of our face encoder.

3.2.4 Evaluations

3.2.4.1 Speech quality evaluation. Similar to the settings used in cross-modal speech synthesis methods [26, 27], we perform speech quality evaluation on the GRID dataset. It is worth mentioning that although both the research direction of these two studies is cross-modal speech synthesis, the specific goal is not to synthesize speech corresponding to the speaker's face but to perform lip-reading tasks based on videos. Therefore, we can not compare the face-voice matching degree with them, but compare the quality of speech generated under the condition of cross-modal speech synthesis. To be consistent with these studies, we evaluated the speech on four speakers s1, s2, S4 and s29.

Mean opinion score(MOS) is an important measure of the naturalness, fluency and clarity of synthesized speech. MOS is rated ranging from 1 to 5 and higher values indicate better audio quality. In our experiment, each speech of the four speakers s1, s2, S4 and s29 is scored by 20 volunteers. The average scores of MOS are shown in the first column in Table 3. Compared with the existing cross modal speech synthesis methods, our method can synthesize higher quality speech from images in video. The average of MOS is 3.94, which means that most speech sounds fluent, pause naturally, and can basically achieve the quality comparable to natural speech.

Another indicator is the accuracy of gender recognition. The values include male, female and hard to judge. The synthesized speech is considered to be gendered correctly only if the speaker's gender can be correctly identified by the volunteers through the synthesized speech. The accuracy of gender recognition is shown in column 2 of Table 3. Compared with the existing methods, our method does achieve the highest recognition accuracy.

Table 5 MOS and Gender recognition evaluation results on VoxCeleb and VGGFace datasets

Framework	MOS	Gender recognition
SSFE(ours)	3.76	82.9%

Matching evaluation Face2Speech is the only audio-visual cross-modal speech synthesis framework that aims to synthesize speech matching the speaker's face image to the best of our knowledge. Although the evaluation of speech is subjective, we still compare our results with theirs, as shown in Table 4. We have prepared two systems to evaluate: 1) SF, the speech generated with f-voice embedding extracted from the face encoder, and 2) SV, the speech generated with f-voice embedding extracted from the voice encoder.

We test the matching degree between the speaker's face image and synthesized speech of all 118 speakers in the test set, that is, the degree to which synthesized voice sounds more consistent with the face's identity. We divide 118 voice pairs (SF, SV) from all subjects into 20 groups of 6 pairs. Each group of voices is scored by three volunteers, and a total of 60 persons participated in the scoring. The scoring metric, proposed in [22], is a four-point scale: 1) Match well, 2) Match moderately, 3) Match slightly, and 4) Not match.

As we can see from Table 4, the matching score of SF in the SSFE framework is higher than SV and lower than the reported value of Face2Speech. The matching score indicates that the f-voice embedding extracted from the face with the SSFE framework can achieve an alternative effect for the speaker's voice characteristics.

We also conducted an AB test on the naturalness of the speech generated from the two systems. The AB test for SF and SV were carried out under the same evaluation environment as the matching evaluation. From the Table 4, it can be seen that the score of SF is even higher than that of SV under the 95% confidence interval. It implies that the voice embeddings learned from face images can further synthesize clear and natural-sounding speech.

In addition, we report MOS and gender recognition evaluation results on VoxCeleb and VGGFace datasets in Table 5. It can be seen that the MOS has achieved an average value of 3.78 and the accuracy of gender identification reached 82.9%, indicating that the synthesized speech sounds natural and has good gender classification accuracy.

The performance of our SSFE framework is better than other existing methods in terms of the quality of synthetic speech and the face-voice matching degree . This proves that our multi-view feature learning method is effective. Through the fusion of different levels of voice features, the synthesized speech can better capture the multi-dimensional information of the speaker's timbre, intonation, and pause. In addition, the domain adaption of voice and face features is reasonable and effective to obtain a new representation containing both two kinds of information. Experiments show that our synthetic speeches does match the speaker's face images well.

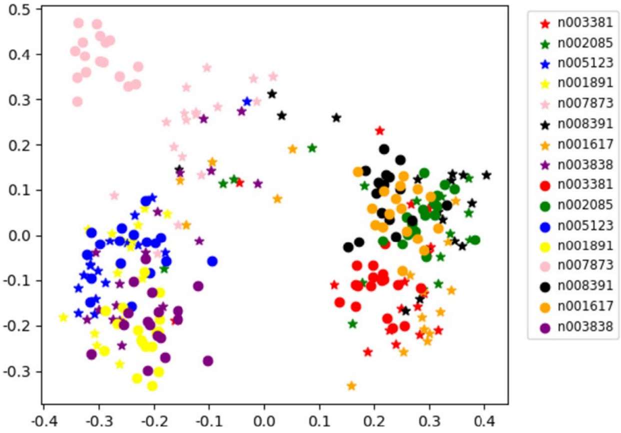
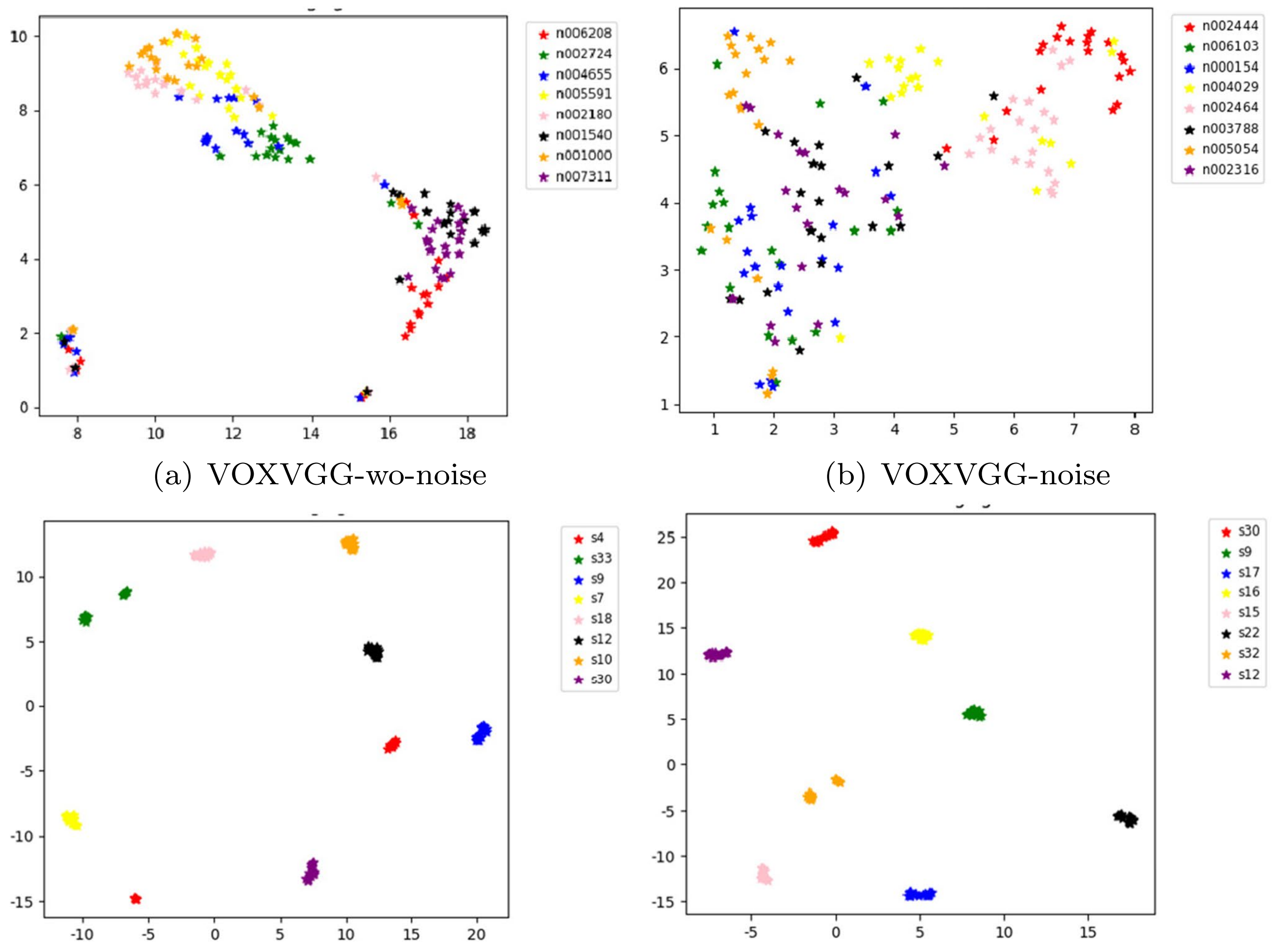


Fig. 7 PCA visualization of the embeddings extracted from SF and SV

Similarity evaluation Although our goal is not to reproduce the speaker's own voice, we still measure the similarity between the embeddings generated from SF and SV. We visualize them by PCA, as shown in Fig. 7. We use the points with same color to represent the same person, where asterisks are embeddings extracted from SF, and dots are embeddings from SV. As we can see, although the points from SF are not completely distributed in the same position as the points from SV, they are still distributed in relatively close locations. Moreover, people of the same gender are distributed on the same side. This shows that our SSFE framework can use the speaker's face to synthesize the voice that sounds similar to the voice synthesized by the speaker's speech and consistent with their gender.

Robustness evaluation To measure the robustness of proposed method, we train the face encoder on two groups of datasets and test the robustness respectively. The first group is the VoxCeleb voice set and VGGFace face image set, and the second group is the GRID dataset in the form of video-audio pair. Two groups of experiments with and without noise were carried out on each dataset, in which the noise is Gaussian noise with a standard deviation equal to 1. For the video of the GRID dataset, the face image is split every 10 frames, the other preprocessing is consistent with VGGFace, and the preprocessing of audios is consistent with the VoxCeleb dataset.

We mark the models trained with VoxCeleb and VGGFace datasets without and with noise as VOXVGG-wo-noise and VOXVGG-noise, respectively. The UMAP visualizations are shown in the first and second rows of Fig. 8. When the VoxCeleb and VGGFace datasets are used, as



(a) VOXVGG-wo-noise (b) VOXVGG-noise

(c) GRID-wo-noise (d) GRID-noise

Fig. 8 UMAP visualization of the f-voice embeddings extracted from the face encoder, which trained with (a)VOXVGG-wo-noise, (b)VOXVGG-noise, (c) GRID-wo-noise and (d)GRID-noise respectively

shown in Fig. 8(a) and (b). We mark the experiments carried on the GRID dataset as GRID-wo-noise and GRID-noise, which are shown in Fig. 8(c) and (d). In general, the embeddings of the same speaker can be gathered together, and the embeddings of different speakers can be distinguished to a certain extent (note: the coordinate distance in Fig. 8(b) is smaller than that in Fig. 8(a)). As we can see, the experiments with the GRID dataset achieve better performance than VoxCeleb, whether it contains noise or not. This is because the GRID dataset contains fewer speakers and have purer face image background. Although intuitively, the performance of the test with noise will be discounted. However, from the experimental results, our model still has a similar performance in the case of noise. This is more obvious on the GRID dataset.

Based on the above experimental results, we believe that the SSFE framework is effective even in a complex environment with noise.

4 Conclusion and future work

From the perspective of multi-view learning, we propose a speech synthesis framework called SSFE. The framework is trained to learn the common feature representation of faces and voices, so that speech synthesis can be carried out under the condition that only the face image of the speaker is provided. The experimental results demonstrate that the speech synthesized by the SSFE framework can match the speaker's face features well and has a high degree of naturalness and a certain degree of similarity with the speaker's speech. The proposed SSFE demonstrate learns a representation that can contain both face and voice information, which makes the connection between different modalities. It can be regarded as a large-scale biophysically meaningful neural network with multi-compartment neurons. In cerebellar morphological theory, it provides the connection between different sensory organs, which may have a certain enlightening effect.

In addition, our modal fusion method is reasonable in theory and practical results, we believe that our method may also be effective in feature fusion in other cross modal tasks. In the future, the application of our model in other cross-modal fusion methods and digital morphological computing is a direction worthy of further exploration.

Acknowledgements This work was supported by the National Key R&D Program of China (Grant No. 2019YFE0190500), the Natural Science Foundation of Shanghai, China (Grant No. 20ZR1420400), the State Key Program of National Natural Science Foundation of China (Grant No. 61936001).

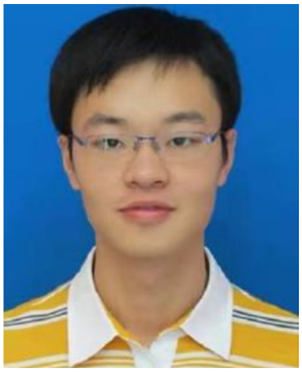
References

1. Pei M, Wu X, Guo Y, Fujita H (2017) Small bowel motility assessment based on fully convolutional networks and long short-term memory. Knowledge-Based Systems 121:163–172
2. Wu X, Chen H, Wang J, Troiano L, Loia V, Fujita H (2020) Adaptive stock trading strategies with deep reinforcement learning methods. Information Sciences 538:142–158
3. Wu X, Du Z, Guo Y, Fujita H (2019) Hierarchical attention based long short-term memory for chinese lyric generation. Applied Intelligence 49(1):44–52
4. Fujita H, Gaeta A, Loia V, Orciuoli F (2019) Improving awareness in early stages of security analysis: A zone partition method based on grc. Applied intelligence 49(3):1063–1077
5. Teager H, Teager S (1990) Evidence for nonlinear sound production mechanisms in the vocal tract. In: Speech production and speech modelling. Springer, pp 241–261
6. Belin P, Fecteau S, Bedard C (2004) Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences 8(3):129–135
7. Jia Y, Zhang Y, Weiss R, Wang Q, Shen J, Ren F, Nguyen P, Pang R, Lopez Moreno I, Wu Y et al (2018) Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Advances in Neural Information Processing Systems 31:4480–4490
8. Wang Y, Stanton D, Zhang Y, Ryan RS, Battenberg E, Shor J, Xiao Y, Jia Y, Ren F, Saurous RA (2018) Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: International conference on machine learning. PMLR, pp 5180–5189
9. Kettenring JR (1971) Canonical analysis of several sets of variables. Biometrika 58(3):433–451
10. Zhang Y, Yang Y, Li T, Fujita H (2019) A multitask multiview clustering algorithm in heterogeneous situations based on lle and le. Knowledge-Based Systems 163:776–786
11. Wu X, Dai S, Guo Y, Fujita H (2019) A machine learning attack against variable-length chinese character captchas. Applied Intelligence 49(4):1548–1565
12. Zhang X, Yang Y, Li T, Zhang Y, Wang H, Fujita H (2021) Cmc: A consensus multi-view clustering model for predicting alzheimer's disease progression. Computer Methods and Programs in Biomedicine 199:105895
13. Zhou W, Guo Q, Lei J, Yu L, Hwang JN (2021a) Irfr-net: Interactive recursive feature-reshaping network for detecting salient objects in rgb-d images. IEEE Transactions on Neural Networks and Learning Systems:1–13. https://doi.org/10.1109/TNNLS.2021.3105484
14. Zhou W, Liu J, Lei J, Yu L, Hwang JN (2021)) Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. IEEE Transactions on Image Processing 30:7790–7802. https://doi.org/10.1109/TIP.2021.3109518
15. Wu F, Jing XY, Wu Z, Ji Y, Dong X, Luo X, Huang Q, Wang R (2020) Modality-specific and shared generative adversarial network for cross-modal retrieval. Pattern Recognition 104:107335. https://doi.org/10.1016/j.patcog.2020.107335
16. Wu F, Jing XY, Feng Y, Ji YM, Wang R (2021) Spectrum-aware discriminative deep feature learning for multi-spectral face recognition. Pattern Recognition 111:107632. https://doi.org/10.1016/j.patcog.2020.107632
17. Yang S, Wang J, Deng B, Azghadi MR, Linares-Barranco B (2021a) Neuromorphic context-dependent learning framework with fault-tolerant spike routing. IEEE Trans Neural Netw Learn Syst:1–15. https://doi.org/10.1109/TNNLS.2021.3084250
18. Yang S, Gao T, Wang J, Deng B, Lansdell B, Linares-Barranco B (2021) Efficient spike-driven learning with dendritic event-based processing. Frontiers in Neuroscience 15:97. https://doi.org/10.3389/fnins.2021.601109
19. Yan X, Ye Y, Qiu X, Yu H (2020) Synergetic information bottleneck for joint multi-view and ensemble clustering. Information Fusion 56:15–27
20. Tamura S, Horio K, Endo H, Hayamizu S, Toda T (2018) Audio-visual voice conversion using deep canonical correlation analysis for deep bottleneck features. In: Proc. Interspeech 2018, pp 2469–2473. https://doi.org/10.21437/Interspeech.2018-2286
21. Hori C, Cherian A, Marks TK, Hori T (2019) Joint student-teacher learning for audio-visual scene-aware dialog. In: Proc. interspeech 2019, pp 1886–1890. https://doi.org/10.21437/Interspeech.2019-3143
22. Oh TH, Dekel T, Kim C, Mosseri I, Freeman WT, Rubinstein M, Matusik W (2019) Speech2face: Learning the face behind a voice. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7539–7548
23. Duarte A, Roldan F, Tubau M, Escur J, Pascual S, Salvador A, Mohedano E, McGuinness K, Torres J, Giro-i Nieto X (2019) Wav2pix: Speech-conditioned face generation using generative adversarial networks. In: ICASSP, pp 8633–8637
24. Nagrani A, Chung JS, Zisserman A (2017) Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH
25. Chung JS, Senior A, Vinyals O, Zisserman A (2017) Lip reading sentences in the wild. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 3444–3453
26. Vougioukas K, Ma P, Petridis S, Pantic M (2019) Video-driven speech reconstruction using generative adversarial networks. In: INTERSPEECH
27. Akbari H, Arora H, Cao L, Mesgarani N (2018) Lip2audspec: Speech reconstruction from silent lip movements video. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2516–2520
28. Effendi J, Sakti S, Nakamura S (2021) End-to-end image-to-speech generation for untranscribed unknown languages. IEEE Access 9:55144–55154. https://doi.org/10.1109/ACCESS.2021.3071541
29. Jenkins R, Tsermentseli S, Monks CP, Robertson DJ, Stevenage SV, Symons AE, Davis JP (2021) Are super - face - recognisers also super - voice - recognisers? evidence from cross - modal identification tasks. Applied Cognitive Psychology
30. Wan L, Wang Q, Papir A, Moreno IL (2018) Generalized end-to-end loss for speaker verification. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4879–4883
31. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: A generative model for raw audio. In: 9th ISCA speech synthesis workshop, pp 125–125

32. Wang Y, Skerry-Ryan R, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S et al (2017) Tacotron: Towards end-to-end speech synthesis. Proc interspeech 2017:4006–4010
33. Kalchbrenner N, Elsen E, Simonyan K, Noury S, Casagrande N, Lockhart E, Stimberg F, Oord A, Dieleman S, Kavukcuoglu K (2018) Efficient neural audio synthesis. In: International conference on machine learning. PMLR, pp 2410–2419
34. Sotelo J, Mehri S, Kumar K, Santos JF, Kastner K, Courville A, Bengio Y (2017) Char2wav: End-to-end speech synthesis. In: ICLR (Workshop Track)
35. Luong HT, Yamagishi J (2018) Multimodal speech synthesis architecture for unsupervised speaker adaptation. In: Proc. Interspeech 2018, pp 2494–2498. https://doi.org/10.21437/Interspeech.2018-1791
36. Shen J, Pang R, Weiss RJ, 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). IEEE, pp 4779–4783
37. Morita T, Koda H (2020) Exploring TTS without T using biologically/psychologically motivated neural network modules (ZeroSpeech 2020). In: Proc. interspeech 2020, pp 4856–4860. https://doi.org/10.21437/Interspeech.2020-3127
38. Zhang H, Lin Y (2020) Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages. In: Proc. interspeech 2020, pp 3161–3165. https://doi.org/10.21437/Interspeech.2020-1403
39. Beskow J (2003) Talking heads-models and applications for multimodal speech synthesis. PhD thesis, Institutionen för talöverföring och musikakustik
40. Goto S, Onishi K, Saito Y, Tachibana K, 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
41. Luong HT, Takaki S, Henter GE, Yamagishi J (2017) Adapting and controlling dnn-based speech synthesis using input codes. In: 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4905–4909
42. Jauk I (2018) Unsupervised learning for expressive speech synthesis. In: IberSPEECH 2018
43. Skerry-Ryan R, Battenberg E, Xiao Y, Wang Y, Stanton D, Shor J, Weiss R, Clark R, Saurous RA (2018) Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In: International conference on machine learning. PMLR, pp 4693–4702
44. Tachibana H, Uenoyama K, Aihara S (2018) Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4784–4788
45. Zen H, Dang V, Clark R, Zhang Y, Weiss RJ, Jia Y, Chen Z, Wu Y (2019) Libritts: A corpus derived from librispeech for text-to-speech. Proc interspeech 2019:1526–1530
46. Chung JS, Nagrani A, Zisserman A (2018) Voxceleb2: Deep speaker recognition. In: INTERSPEECH
47. Ardila R, Branson M, Davis K, Kohler M, Meyer J, Henretty M, Morais R, Saunders L, Tyers F, Weber G (2020) Common voice: A massively-multilingual speech corpus. In: Proceedings of The 12th language resources and evaluation conference, pp 4218–4222
48. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp 67–74

Xing Wu received his Ph.D. degree from the Department of Computer Science and Technology, Shanghai Jiaotong University in 2010. He is currently a professor with the School of Computer Engineering and Science, Shanghai University, China. He is the Vice Dean of the Department of Computer Science and Technology, Shanghai University. His research interests include machine learning and data mining. He is the author of many research studies published in national and international journals, conference proceedings and book chapters.

Sihui Ji is a postgraduate student at the School of Computer Engineering and Science, Shanghai University. Her major research interests are multi-modal speech synthesis.



Jianjia Wang received the M.S. degrees in electronic engineering from HKUST in 2013, and the Ph.D. degree in computer science from the University of York, U.K, in 2018. He is currently an assistant professor (lecturer) at Shanghai University. He is also an adjunct professor at the Council on International Educational Exchange with Rutgers, the State University of New Jersey. His research interests include complex networks, statistical and structural pattern recognition, data science, and so on.

Yike Guo is the vice president of Hong Kong Baptist University. He is a Fellow of the Royal Academy of Engineering (FREng), Member of Academia Europaea (MAE), Fellow of British Computer Society and a Trustee of The Royal Institution of Great Britain. Professor Guo has published over 200 articles, papers and reports.