Interspeech 2024
1-5 September 2024, Kos, Greece

Hear Your Face: Face-based voice conversion with F0 estimation

Jaejun Lee[1], *Yoori Oh*[1], *Injune Hwang*[1], *Kyogu Lee*[1,2,3]

[1]Department of Intelligence and Information, Seoul National University
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University
[3]Artificial Intelligence Institute, Seoul National University
{jjlee0721, yoori0203, dlswns8, kglee}@snu.ac.kr

Abstract

This paper delves into the emerging field of face-based voice conversion, leveraging the unique relationship between an individual's facial features and their vocal characteristics. We present a novel face-based voice conversion framework that particularly utilizes the average fundamental frequency of the target speaker, derived solely from their facial images. Through extensive analysis, our framework demonstrates superior speech generation quality and the ability to align facial features with voice characteristics, including tracking of the target speaker's fundamental frequency.

Index Terms: voice conversion, face/voice association, cross modal generation, speaker embedding

1. Introduction

Voice, the cornerstone of human speech, plays a crucial role in interpersonal communication. Beyond its communicative function, voice is a distinctive feature of an individual, reflecting personal identity. Consequently, individuals who are unable to produce sound not only face a significant barrier to communication but also experience a loss of personal expression.

Conventional speech synthesis techniques, such as Text-to-Speech (TTS) and Voice Conversion (VC), have made significant strides in emulating a target voice while retaining the nonverbal content elements. Yet, these techniques predominantly rely on the availability of the target voice's acoustic data to replicate its unique speech style effectively.

The human face represents another intrinsic aspect of individual identity, containing details such as biological gender, ethnicity, and age. More than just exploring visual information from face, recent studies have increasingly focused on understanding the relationship between facial features and vocal attributes [1, 2]. This field of study may hold the key to a new form of speech synthesis, one that retains the target speaker's identity even in the absence of vocal information.

Recent advancements in face-based speech synthesis have experienced a notable surge, particularly through the integration of conventional TTS [3, 4, 5, 6] and VC [7, 8, 9] techniques. While this growing interest and development show a promising view, the field remains in its formative phases. Specifically, identifying a 'voice that matches the face' presents significant challenges, and metrics for evaluating it also remain a pivotal question.

The fundamental frequency ($F0$), one of the key components in voice conversion process [10, 11], not only serves as a pitch information of speech but also has an aspect of containing information of speakers identification [12]. It has been found that facial features have correlation with voice pitch information, even in cases where gender is controlled [1]. It implies that voice pitch information, indicated by the $F0$, could be derived from the speaker's facial images, and it is not merely from basic gender identification, but also from further biological associative information.

In this study, we propose a novel framework for face-based speech synthesis, focusing particularly on the voice conversion that imprints a face-based target voice's characteristics onto the original source audio. Our framework specifically utilize the $F0$ of target speaker, derived solely from facial images. This approach aims to enhance the face-based voice conversion process, generating speech that is well aligned with the target individual's vocal identity without using any acoustic data of the target speaker.

To contextualize our research, we delineate our contributions as follows:

- We present a framework that sets a new benchmark in performance for face-based voice conversion, demonstrating state-of-the-art results.
- We propose a novel approach for speech synthesis by estimating the $F0$ of the target speaker through their facial images.
- Through extensive analysis and the introduction of a novel evaluation metric, we demonstrate that our framework not only produces high-quality synthetic speech but also suggests that the synthesized voice aligns reasonably well with the corresponding facial image.

The demo is available on the link, `https://jaejunL.github.io/HYFace_Demo/`.

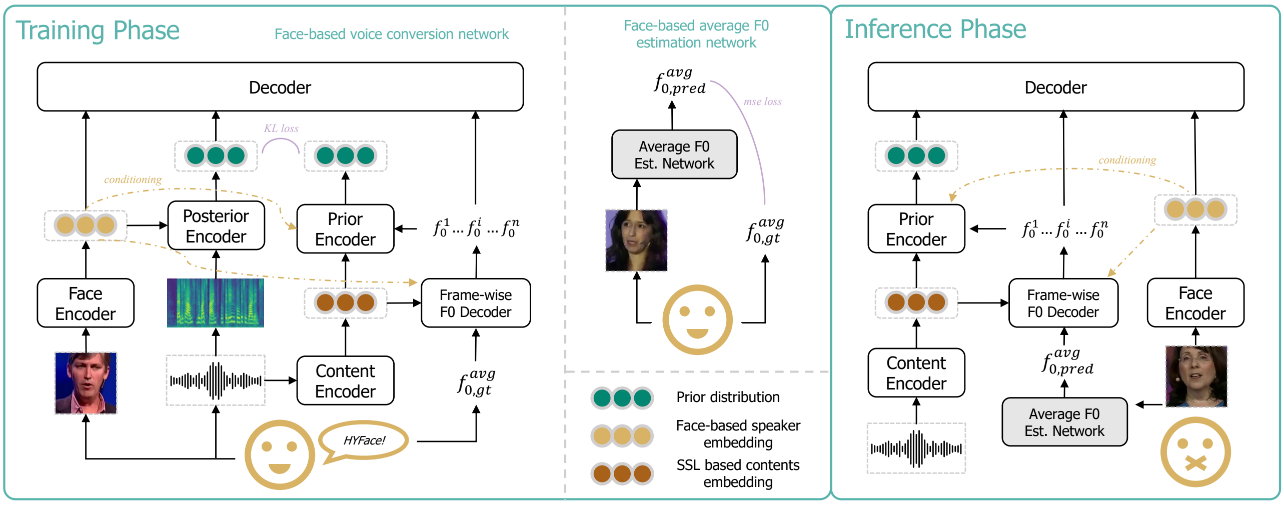
2. Related work

2.1. Voice conversion

Voice conversion, a specialized subset of speech synthesis, is a process that automatically transforms speech from one source speaker into a voice resembling that of a target speaker, all the while maintaining the original linguistic content. The challenge primarily arises in non-parallel voice conversion scenarios, where the lack of directly corresponding parallel data. The disentanglement of linguistic content in speech and its acoustic voice, timbre is a crucial problem. To address this, various methodologies have been explored, including adversarial training [13, 14], vector quantization [15, 16], and information perturbation [10, 11].

Notably, recent advancements have been made with the advent of self-supervised learning (SSL) techniques. Pretrained representations trained on large data corpus exhibit a remarkable capacity for disentangling the contents information of speech [17, 18], thereby significantly enhancing voice conversion processes [19]. Recently, the collaborative voice conver-

10.21437/Interspeech.2024-232



(a) Training phase

(b) Inference phase

Figure 1: *Overview of the proposed method, HYFace, conditional VAE based network that its speaker embedding is learned on face images only. In training phase, a predefiend speaker-wise average F0 ($f_{0,gt}^{avg}$) is used to estimate frame-wise $F0$ values. However, as the $f_{0,gt}^{avg}$ values for unseen target speakers are not available during the inference phase, we independently train an average $F0$ estimation network based solely on facial inputs. This module is then utilized in the inference phase.*

sion project, named Sovits-SVC[1] (SoftVC VITS Singing Voice Conversion), has demonstrated impressive outcomes in both standard voice conversion and singing voice conversion domains. It leverages SSL representations for content representations, and employs a neural-source filter vocoder, specifically designed to track the $F0$ of the source audio, which plays a significant role in its original intention for singing voice conversion.

2.2. Face-voice association

Early studies, especially through human behavioral and neuroimaging approach, demonstrate that humans use both facial and vocal cues for identity recognition [12, 20]. Furthermore, similar studies reveal human ability to match faces with voices of unfamiliar individuals [1, 21, 22]. Particularly, the authors in [1] showed that humans can significantly match faces and voices under the controlled attributes such as gender, race, and age. Specifically, they also revealed that there is a correlation between the target speaker's voice pitch and facial features.

Building on these finding, interest has surged in learning based methods for associations between faces and voices. An application of such methods includes generating face from a given speech [2, 23] or vice versa. Specifically, face-based speech synthesis, the focus of this paper, is categorized based on the type of input: text for TTS [3, 4, 5, 6] and source audio for VC [7, 8, 9, 24]. Recently, Sheng *et al.* [9] showed prominent result in zero-shot face-based voice conversion, employing memory based methods. All these works tried to learn cross-modal speaker representations implicitly, without explicit voice characteristic such as $F0$. Moreover, their evaluation primarily relied on metrics such as the mean opinion score (MOS) or speaker embedding similarities, rather than on assessments directly related to explicit voice characteristics.

[1]https://github.com/svc-develop-team/so-vits-svc

3. Methods

In this section, we present our proposed method, HYFace (short for 'Hear Your Face'), a novel approach to face-based voice conversion, it begins in Section 3.1. Then, Section 3.2 provides detailed architectures of our proposed model. Figures 1(a) and 1(b) illustrate the procedures of the training phase and the inference phase of our method, respectively.

3.1. HYFace

Our HYFace network is a voice conversion (VC) framework fundamentally inspired by Sovits-SVC, utilizing a conditional variational autoencoder architecture. It incorporates pretrained SSL representations as content input for the prior encoder. However, distinct from traditional VC frameworks, HYFace uses the facial image of the target speaker to modify the style of the source audio, instead of using the target speaker's voice. In this system, the speaker embedding, which is learned from the facial images, conditions the prior encoder, posterior encoder, decoder and the frame-wise $F0$ decoder (FF).

Additionally, to enhance the model's capacity to incorporate target voice characteristics, frame-wise $F0$ values, f_0^i ($i = 1, ..., n$. n is the number of frames) conditions both the prior encoder and the decoder. A speaker-wise average $F0$ ($f_{0,gt}^{avg}$) is adjusted to f_0^i within the FF, in conjunction with the content embedding (c) and face-based speaker embedding (s). The loss $\mathcal{L}_{ff}$ for training FF is as follows:

$$\mathcal{L}_{ff} = \frac{1}{n}\sum_{i=1}^{n}(f_{0,gt}^{i} - FF(f_{0,gt}^{avg}, c, s))^2, \tag{1}$$

where $f_{0,gt}^{i}$ refers to the ground-truth frame-wise $F0$ values. Note that $f_{0,gt}^{avg}$ value represents the average of $f_{0,gt}^{i}$ values across all audio frames for each speaker in the training dataset. Importantly, due to the unavailability of $f_{0,gt}^{avg}$ information for unseen target speakers during the inference phase, we independently train a face-based average $F0$ estimation network (AF) solely on face image (v) of target speakers, which constitutes one of the key components of our proposed method. The

Table 1: *Evaluation result. The definitions of all metrics are provided in Section 4.3.*

	Homogeneity↑		Diversity↓		Consistency(obj)↑		Consistency(sub)↑		Naturalness↑		ABX test(%)↑	
	HMG	HTG	HMG	HTG	HMG	HTG	HMG	HTG	HMG	HTG	HMG	HTG
GT	0.7456	-	0.5418	-	-	-	3.9048	-	4.0469	-	-	-
FVMVC[9]	0.6391	0.6401	**0.5942**	**0.5976**	0.5105	0.5086	3.5705	3.5009	3.4096	3.2470	0.395	0.420
HYFace	**0.6770**	**0.6793**	0.6072	0.6103	**0.5696**	**0.5632**	**3.8916**	**3.8189**	**3.8313**	**3.7651**	**0.605**	**0.580**

loss $\mathcal{L}_{af}$ for training AF is as follows:

$$\mathcal{L}_{af} = (f_{0,gt}^{avg} - AF(v))^2. \quad (2)$$

Then, AF is utilized during the inference phase, enhancing our face-based voice conversion network to produce speech that better aligns with the voice characteristics of the target speaker. To clarify our HYFace training procedure, we describe our other loss functions, which include reconstruction loss, KL (Kullback-Leibler) divergence loss, adversarial loss, and feature matching loss in Supplementary A.

3.2. Model architecture

This section details the architecture of the modules employed in our model. We note that all modules were trained from scratch, except for the contents encoder, which is based on pretrained models.

Posterior Encoder: It is consists of WaveNet-based residual blocks. To integrate face-based speaker embeddings, we employed global conditioning similar to [25].

Prior Encoder: It has a transformer-based architecture similar to [26], atop which is stacked a normalizing flow layer comprised of residual coupling blocks [27].

Face Encoder: Vision Transformer [28] architectures with projection layer.

Contents Encoder: We used ContentVec [18], pretrained SSL represenations, especially hugging face version[2]

Decoder: It basically has architecture similar to the generator of HiFi-GAN [29] so as to our discriminator network (D), but for careful conditioning of $F0$ information, we used a neural source filter method [30] based conditioning similar to Sovits-SVC.

Frame-wise F0 Decoder (FF): It is based on architecture with self-attention layers and feed forward layers conditioned with both content embedding and face-based speaker embedding. Fast Context-base Pitch Estimator[3] (FCPE) is used to extract frame-wise $F0$ value and speaker-wise average $F0$ value.

Average F0 estimation network (AF): Similar to face encoder, vision transformer based architectures with projection layer.

4. Experiments

4.1. Dataset

We used LRS3 [31], the dataset consists of 5,502 videos from TED and TEDx which has more than 430 hours long. Each video is cropped on speaker's face and it has a resolution of 224×224 with 25 frames per seconds images and 16kHz single channel audio. We used predefined *pretrain*, *trainval* and *test* set for training, validation and evaluation, respectively. Expecting our proposed model to more carefully associate detailed face features with speaker's voice characteristics, we used only frontal images from the dataset. Especially, we employed OpenCV haarcascades[4] for image selection, resulting in about 20% of image data being filtered out.

For the evaluation, we picked 50 male and 50 female speakers on *test* set, ranked by the amount of data available. On average, there are about 5.8 speech audio files and 280 facial images available per speaker. We hypothesized that converting the voice to a target speaker of a different gender from the source is more challenging than converting to the same gender. Therefore, we constructed two types of evaluation sets: Homogeneous Gender (HMG) set and Heterogeneous Gender (HTG) set. The HMG set pertains to face-based voice conversion scenarios in which the target speaker's gender is the same as the source speaker's (either male to male (M2M) or female to female (F2F)). In contrast, the HTG set applies to scenarios where the target speaker's gender is different from that of the source speaker (from male to female (M2F) or female to male (F2M)). Thus, technically we have four evaluation sets: M2M and F2F for HMG and M2F and F2M for HTG.

4.2. Comparison systems

Ground truth (GT): The original speech audio, which serves as the upperbound.

FVMVC: Face-based memory-based zero-shot Face Voice Conversion model [9] which recently demonstrated state-of-the-art performance on LRS3 dataset.

HYFace: Our proposed method, detailed in Section 3.

4.3. Metrics

Following Sheng *et al.* [9] and other conventional VC studies, for objective evaluation, we assess the homogeneity, diversity, and objective consistency. For subjective evaluation, we examine subjective consistency, naturalness, and ABX tests. Furthermore, we propose a new evaluation metric: pitch deviation. For all objective evaluations, we randomly selected 10 source speakers for each of the 50 target speakers and repeated this process for 10 trials. This resulted in a total of 5,000 conversion pairs for each of the four evaluation sets. To measure cosine similarity, we utilized speaker embeddings generated by *Resemblyzer*[5]. For subjective evaluations, we use Mean Opinion Scores (MOS) collected via Amzon Mechanical Turk (MTurk). We described detailed procedure of MTurk on Supplementary B. The explanation of each metric is as follows.

Homogeneity: It measures cosine similarity of speaker embeddings in synthesized audio generated from different facial images of the same speaker. Similarity value is expected to be high regardless of different view of face image on same speaker. We randomly select 10 face images from each target speaker.

Diversity: It measures cosine similarity of speaker embeddings in synthesized audio generated from different speakers. In contrast from homogeneity, here the model aims to capture distinct

[2]https://huggingface.co/lengyue233/content-vec-best
[3]https://github.com/CNChTu/FCPE
[4]https://github.com/opencv/opencv/tree/master/data/haarcascades
[5]https://github.com/resemble-ai/Resemblyzer

speaker information for different target speakers.
Consistency(obj): It compares the speaker embedding similarity of the synthesized audio with that of the ground-truth audio from the same speaker. To ensure a robust comparison, we also assess this metric with ground-truth audio from a random speaker, referred to as 'Consistency(rnd)'.
Consistency(sub): This metric measures consistency for subjective evaluation using a 5-point MOS scale (completely inconsistent to completely consistent). It assesses whether the synthesized audio aligns with the corresponding facial images.
Naturalness: It assesses the sound quality of the synthesized audio using 5-point MOS scale (completely unnatural to completely natural).
ABX test: This evaluates the subjective preference between two models. Participants are shown a face image and asked to decide which of two synthesized audio samples, one from HYFace and the other from FVMVC, more closely matches the face in the image.
Pitch deviations: It is our newly proposed metric. Since the $F0$ is one of the key component of voice, we assess the deviation between the $F0$ of the synthesized audio and the average $F0$ (represented as $f_{0,gt}^{avg}$ in Section 3) of the ground-truth target speaker. Note that the standard deviations (stdv) of the $F0$ for all audio samples are 29.18 Hz for male speakers and 37.50 Hz for female speakers. These values serve as a baseline, reflecting the deviation is based solely on gender class. If the model captures the associations between facial features and voice characteristics within a gender-controlled set, then it should demonstrate a deviation lower than these baseline values.

4.4. Results

The evaluation results for the metrics discussed in previous section can be found in Table 1. As mentioned in Section 4.1, we have created four evaluations sets: HMG (M2M and F2F), HTG (M2F and F2M). However, due to space limitations, we present the averaged scores for both HMG and HTG. For detailed results from all four evaluation sets, please refer to Supplementary C. Note that GT refers to the ground-truth audio, in which pairing between source and target of different genders (heterogeneous gender pairing) is not possible. Therefore, only HMG scores are applicable. We have not reported Consistency(obj) scores for GT as it is conceptually identical to Homogeneity.

4.4.1. Objective results

In Homogeneity, our proposed model HYFace presents high scores in both HMG and HTG ($p < 0.01$ in paired t-tests) than FVMVC. For Diversity, the FVMVC shows better scores both in HMG and HTG ($p < 0.01$ in paired t-tests). For Consistency(obj), HYFace scored higher, indicating closer similarity with the ground-truth audio. For fair comparing, we measured the Consistency(rnd), which measures similarity between synthesized audio and the ground-truth audio of 'different' speakers. For HYFace, the Consistency(rnd) scores are 0.5577 for HMG and 0.5524 for HTG. The Consistency(obj) scores of HYFace are statistically higher than Consistency(rnd) ($p < 0.01$ in paired t-tests) both in HMG and HTG, suggesting HYFace has meaningful correlation with voice characteristics of ground-truth speaker. However, in case of FVMVC, the Consistency(rnd) scores are 0.5074 for HMG and 0.5064 for HTG, showing no significant difference from its Consistency(obj) scores ($p > 0.02$ for HMG and $p > 0.1$ for HTG in paired t-tests). It means that there is no correlation with speaker embedding of synthesized audio from FVMVC and that

Table 2: *Comparison of pitch deviation (in Hz) between synthesized audio and ground-truth average $F0$*

	HMG↓		HTG↓	
	M2M	F2F	M2F	F2M
FVMVC[9]	29.55	34.15	34.75	28.50
HYFace	**24.01**	**29.58**	**29.15**	**24.31**

of ground-truth audio. The objective results suggest that although HYFace may show slightly poorer Diversity compared to the benchmark, its speaker embeddings significantly align with those of the ground-truth speaker, a feature not observed in the benchmark model. Additionally, our model demonstrates higher Homogeneity scores.

4.4.2. Subjective results

In all subjective evaluations, including Consistency(sub), Naturalness, and ABX test, our proposed HYFace model outperformed the benchmark for both HMG and HTG ($p < 0.01$ in paired t-tests). Remarkably, in the Consistency(sub) metric, which assesses how well the synthesized audio matches the corresponding ground-truth facial image, HYFace achieved scores comparable to those of the ground-truth audio. Furthermore, HYFace's Naturalness score nearly approached that of the ground-truth audio. In terms of performance differences between HMG and HTG sets, only the Naturalness scores in the FVMVC model showed a significant decrease in the HTG set compared to HMG ($p < 0.01$ in paired t-tests).

4.4.3. Pitch deviations

Table 2 presents the $F0$ deviation of two models, our proposed HYFace and FVMVC, the benchmark. Across evaluation sets of all gender pairings of source and target speakers (M2M, F2F, M2F, and F2M), the proposed HYFace exhibited superior $F0$ estimation performance compared to the benchmark ($p < 0.01$ in paired t-tests). Furthermore, HYFace consistently demonstrated significantly lower deviations compared to the stdv of the ground truth. In the male cases, the deviations are 24.01 for M2M and 24.31 for F2M, both below the GT male stdv of 29.18. In female cases, the deviations are 29.58 for F2F and 29.15 for M2F, each below the GT female stdv of 37.50. It suggests that our model can nearly estimate the pitch of the target speaker based solely on facial images, even under the controlled gender set. To our knowledge, this marks the first instance of evaluating explicit voice characteristics, pitch, associating with facial features within the face-based voice conversion domain, and also yielding meaningful results.

5. Conclusion

In this work, we present a novel framework for face-based voice conversion, particularly utilizing fundamental frequency estimation module, which operates solely on facial images. Through comprehensive objective and subjective evaluations, our model has achieved state-of-the-art performance. Moreover, in our newly proposed metric, which explicitly assesses the association between facial features and voice characteristics, our method has yielded meaningful results. We hope that this research will serve as stepping stones towards providing individuals without a voice with one that fits their identity.

6. Acknowledgements

We sincerely give thanks to *Zheng-Yan, Sheng* who is the author of [9] for dedication to the academy area and kind communication regarding model reproduction. This work was supported by Institute of Information communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) [No. RS-2022-II220641, XVoice: Multi-Modal Voice Meta Learning].

7. References

[1] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*. Springer, 2019, pp. 276–292.

[2] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7539–7548.

[3] S. Goto, K. Onishi, Y. Saito, K. Tachibana, and K. Mori, "Face2speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image." in *INTERSPEECH*, 2020, pp. 1321–1325.

[4] B. Plüster, C. Weber, L. Qu, and S. Wermter, "Hearing faces: Target speaker text-to-speech synthesis from a face," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 757–764.

[5] J. Lee, J. S. Chung, and S.-W. Chung, "Imaginary voice: Face-styled diffusion model for text-to-speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] P. Wen, Q. Xu, Y. Jiang, Z. Yang, Y. He, and Q. Huang, "Seeking the shape of sound: An adaptive framework for learning voice-face association," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 347–16 356.

[7] H.-H. Lu, S.-E. Weng, Y.-F. Yen, H.-H. Shuai, and W.-H. Cheng, "Face-based voice conversion: Learning the voice behind a face," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 496–505.

[8] N. Takahashi, M. K. Singh, and Y. Mitsufuji, "Cross-modal face- and voice-style transfer," *arXiv preprint arXiv:2302.13838*, 2023.

[9] Z.-Y. Sheng, Y. Ai, Y.-N. Chen, and Z.-H. Ling, "Face-driven zero-shot voice conversion with memory-based face-voice alignment," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8443–8452.

[10] H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 251–16 265, 2021.

[11] H.-S. Choi, J. Yang, J. Lee, and H. Kim, "Nansy++: Unified voice synthesis with neural analysis and synthesis," in *The Eleventh International Conference on Learning Representations*, 2022.

[12] M. Xu, F. Homae, R.-i. Hashimoto, and H. Hagiwara, "Acoustic cues for the recognition of self-voice and other-voice," *Frontiers in psychology*, vol. 4, p. 735, 2013.

[13] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[14] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[15] A. T. Liu, P.-c. Hsu, and H.-y. Lee, "Unsupervised end-to-end learning of discrete linguistic units for voice conversion," *arXiv preprint arXiv:1905.11563*, 2019.

[16] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," *arXiv preprint arXiv:2106.10132*, 2021.

[17] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[18] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, "Contentvec: An improved self-supervised speech representation by disentangling speakers," in *International Conference on Machine Learning*. PMLR, 2022, pp. 18 003–18 017.

[19] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6562–6566.

[20] S. Campanella and P. Belin, "Integrating face and voice in person perception," *Trends in cognitive sciences*, vol. 11, no. 12, pp. 535–543, 2007.

[21] L. W. Mavica and E. Barenholtz, "Matching voice and face identity from static images." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 2, p. 307, 2013.

[22] H. M. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey, "Matching novel face and voice identity using static and dynamic facial images," *Attention, Perception, & Psychophysics*, vol. 78, pp. 868–879, 2016.

[23] H.-S. Choi, C. Park, and K. Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," in *International Conference on Learning Representations*, 2019.

[24] J. Wang, Z. Wang, X. Hu, X. Li, Q. Fang, and L. Liu, "Residual-guided personalized speech synthesis based on face image," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4743–4747.

[25] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[26] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[27] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[30] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5916–5920.

[31] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.