Progressive Facial Granularity Aggregation with Bilateral Attribute-based Enhancement for Face-to-Speech Synthesis

Yejin Jeon[1], **Youngjae Kim**[1], **Jihyun Lee**[1], **Hyounghun Kim**[1,2], **Gary Geunbae Lee**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[2]Department of Computer Science and Engineering, POSTECH, Republic of Korea
{jeonyj0612, yj122198, jihyunlee, h.kim, gblee}@postech.ac.kr

Abstract

For individuals who have experienced traumatic events such as strokes, speech may no longer be a viable means of communication. While text-to-speech (TTS) can be used as a communication aid since it generates synthetic speech, it fails to preserve the user's own voice. As such, face-to-voice (FTV) synthesis, which derives corresponding voices from facial images, provides a promising alternative. However, existing methods rely on pre-trained visual encoders, and finetune them to align with speech embeddings, which strips fine-grained information from facial inputs such as gender or ethnicity, despite their known correlation with vocal traits. Moreover, these pipelines are multi-stage, which requires separate training of multiple components, thus leading to training inefficiency. To address these limitations, we utilize fine-grained facial attribute modeling by decomposing facial images into non-overlapping segments and progressively integrating them into a multi-granular representation. This representation is further refined through multi-task learning of speaker attributes such as gender and ethnicity at both the visual and acoustic domains. Moreover, to improve alignment robustness, we adopt a multi-view training strategy by pairing various visual perspectives of a speaker in terms of different angles and lighting conditions, with identical speech recordings. Extensive subjective and objective evaluations confirm that our approach substantially enhances face-voice congruence and synthesis stability.

1 Introduction

The ability to communicate using one's own voice is an intrinsic and fundamental aspect of human identity, self-expression, and social interaction. However, a range of neurological and physiological conditions can severely impair speech production mechanisms. For instance, dysarthria is a speech disorder that results from various etiological factors, including cerebrovascular incidents (e.g., strokes) and degenerative neuromuscular disorders such as multiple sclerosis and Parkinson's disease (Darley et al., 1969; Mulfari et al., 2021). The severity of dysarthria varies, but nevertheless manifests as slurred, unintelligible, and phonetically distorted speech. In more extreme cases such as total glossectomy or orofacial myofunctional disorders, individuals may completely lose the ability to generate speech. These impairments significantly hinder verbal communication, often leading to frustration, social isolation, and a reduced quality of life (Mertl et al., 2018).

Towards this, text-to-speech (TTS) systems (Shen et al., 2018; Ren et al., 2021; Kim et al., 2021) have been utilized as assistive technologies to convert typed text into artificial speech (Mertl et al., 2018; Jeon et al., 2025). However, while they are effective for communication, they fail to preserve the speaker's unique vocal identity. Due to this, personalized multi-speaker TTS methods (Hu et al., 2020; Badlani et al., 2023; Jeon et al., 2024) could be a potential solution (Mertl et al., 2018) as they are able to imitate a target speaker's vocal characteristics for speech generation. Yet, achieving high-fidelity voice cloning typically requires extensive multi-speaker training or few-shot adaptation with a large amount of speech recordings. Thus, while zero-shot multi-speaker synthesis can ideally enable speaker adaptation, it remains infeasible for those without any accessible prior recordings, such as individuals with complete vocal muscle paralysis. This limitation highlights the need for alternative biometric modalities to infer speaker identity and enable personalized voice synthesis that is independent of speech recordings.

Both early and recent psychological research suggests the existence of a strong correlation between facial identity with vocal characteristics. Specifically, studies such as Kamachi et al. (2003); von Kriegstein et al. (2005); Smith et al. (2016)

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2799–2811
November 4-9, 2025 ©2025 Association for Computational Linguistics

have indicated that humans can infer aspects of a person's voice, such as pitch and timbre, based solely on their facial features. As a result, this insight has driven investigations into face-to-voice (FTV) synthesis, where a facial image serves as a reference for generating a corresponding voice, and then producing speech in that inferred voice.

Early methods of FTV synthesis relied on statistical techniques that mapped facial images onto facial landmarks, which were then transformed into eigenvoice representations (Ohsugi et al., 2018). These approaches, however, were limited by the need for manual alignment between facial and vocal features. More recent work has adopted multi-stage training pipelines (Goto et al., 2020; Yang et al., 2023; Kang et al., 2025), wherein a face encoder is first trained to produce embeddings that align with outputs from a separately trained audio-based speech encoder. Afterwards, during inference, the face encoder substitutes the speech encoder. While this strategy improves performance, it introduces training complexity and inefficiency. Moreover, because the facial representations are specifically trained to align with audio features, they fail to directly capture and utilize fine-grained local facial features such as gender or other demographic attributes that are essential in producing identity-consistent speech. Although Yang et al. (2023) attempts to explicitly incorporate such facial attributes, their method requires a combination of demographic metadata, 2D facial features, textures (e.g., skin, muscle), and 3D cranial structures. Not only does this introduce substantial complexity in terms of data acquisition and input modalities, but also necessitates a three-stage learning process to integrate these sources of information, which ultimately results in a complex and resource-intensive training pipeline.

To address the limitations of prior work, specifically the reliance of inference-time-only visual encoders, underutilization of fine-grained facial features, and dependence on multi-stage training, we present a novel end-to-end FTV synthesis framework, which eschews external models and instead focuses on effective facial feature learning. Our method starts with a progressive facial encoder that decomposes and hierarchically aggregates local facial regions to form a rich visual identity embedding. Additionally, to ensure alignment of high-level semantic attributes such as gender and ethnicity, we introduce a bilateral attribute enhancement mechanism, which is applied to both the facial representation and to the synthesized audio output. A multi-view data augmentation strategy is further adopted where each speech sample is paired with a diverse set of facial images from the same speaker, which are captured under varying poses, lighting conditions, and facial movements. Through both subjective and objective evaluations, we demonstrate that our method generates speech with higher speaker fidelity and stronger FTV associations.

2 Related Work

2.1 Stylistic Speech Generation

Advancements in TTS technology have greatly enhanced the naturalness of synthesized speech, thereby expanding its applications to areas such as voice imitation in the form of multi-speaker TTS. Methodologies of this domain incorporate a speaker encoder that extracts a speaker representation from a reference audio sample, and then conditions it into the backbone TTS model. Existing methodologies typically follow either a few-shot adaptation strategy, where a pre-trained multi-speaker TTS model is fine-tuned with multiple target speaker samples (Huang et al., 2022; Chen et al., 2019, 2021), or a zero-shot adaptation approach, which generates an embedding from a single target speaker sample to condition the model directly. Regardless of the adaptation objective, research in multi-speaker TTS has either focused on improving speaker encoders (Cooper et al., 2020; Jeon et al., 2024) or refining speaker conditioning techniques by evolving from simple concatenation to more adaptive methods (Min et al., 2021; Choi et al., 2022; Yoon et al., 2023). While FTV synthesis is a form of multi-speaker TTS, it takes a fundamentally different approach. Instead of relying on reference audio at training or inference time, it leverages a facial image to infer a speaker's vocal characteristics. As such, this task is inherently more challenging, as it requires learning to associate visual identity cues such as gender, with corresponding acoustic traits.

2.2 Multi-Modal Speech Synthesis

The innate human ability to associate voices with faces (Ellis, 1989), along with evidence of overlapping neural-cognitive pathways for processing these modalities (Joassin et al., 2011), has fueled extensive research in multi-modal visual-acoustic learning. A prominent direction in this field is voice-to-face generation, where models synthesize

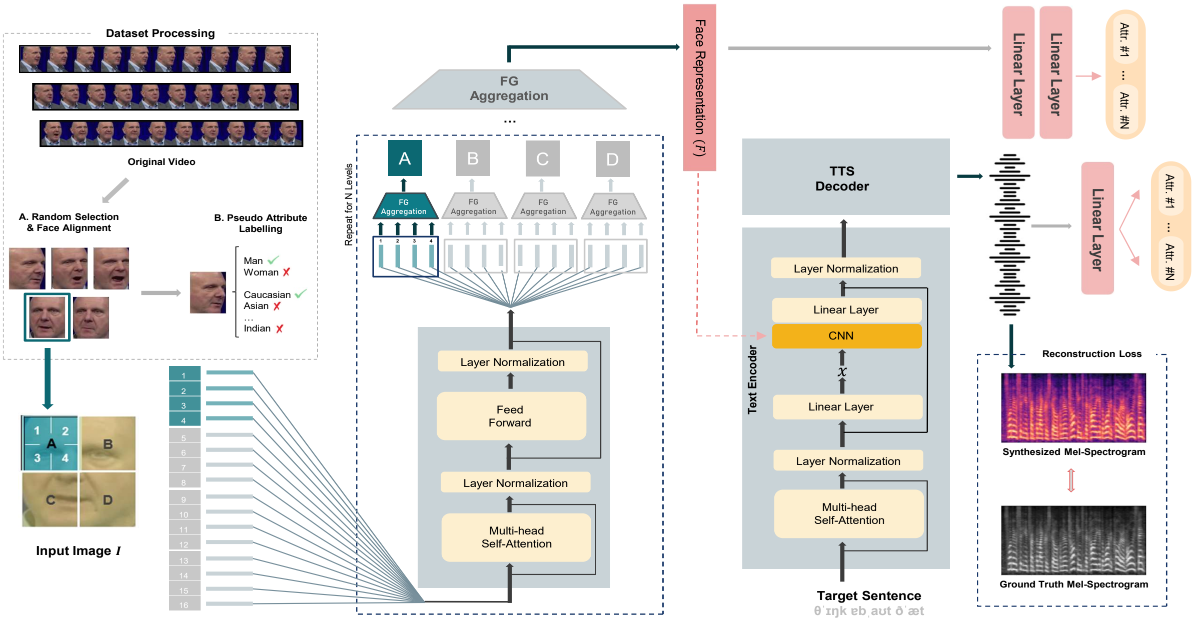


Figure 1: Diagram illustrating the proposed model. On a high level, an input image is segmented into 16 smallest-size patches. Four sets of adjacent patches are then combined to create four larger areas. These four areas are subsequently aggregated into a final representation F, which ultimately integrates information from all 16 original patches. Moreover, attribute enhancement is further conducted on both the facial and acoustic domains. Note that the architecture following the text encoder output is identical to the original VITS (Kim et al., 2021) model.

plausible facial features from speech, effectively constructing a visual identity from audio (Sinha et al., 2020; Liu and Wang, 2023; Xu et al., 2024.). In contrast, face-to-voice generation has been studied in applications like biometric security (Jiang et al., 2024), lip-dubbing (Chung and Zisserman, 2016; Park et al., 2022; Mukhopadhyay et al., 2024; Choi et al., 2023; Lei et al., 2024) and voice synthesis (Yang et al., 2023; Sheng et al., 2023; Kang et al., 2025). Previous approaches in voice synthesis often rely on multi-stage training pipelines by aligning face and audio embeddings, and then only leveraging face encoders during inference (Goto et al., 2020; Kang et al., 2025). To enhance identity preservation across modalities, various loss functions such as tri-item (Wang et al., 2022) and disentanglement losses have also been formulated (Nagrani et al., 2020). In this work, we focus on face-to-voice synthesis while addressing the limitations of previous methods by eliminating multi-stage training. Towards this objective, we extract effective facial embeddings enriched with identity-relevant attributes like gender and ethnicity for higher speaker fidelity.

3 Methodology

3.1 Problem Setup

In a conventional multi-speaker TTS framework, the training dataset is defined as $D = \{(X_n, Y_n)\}_{n=1}^{N}$, where X_n and Y_n represent the input text and corresponding speech waveform, respectively. The objective is to generate an acoustic output $\hat{Y}_n$ that reflects both the linguistic content of X_n and the vocal identity of the target speaker. To enable speaker-specific synthesis, an auxiliary speaker encoder is employed to extract a speaker embedding from the reference audio Y_n.

Extending this framework to the FTV synthesis task, we aim to generate speech that aligns with the input text and conveys the vocal characteristics of the individual, which is given as a facial image. Accordingly, the training dataset is reformulated as $D = \{(X_n, Y_n, Z_n)\}_{n=1}^{N}$, where Z_n denotes the facial image that is extracted from a video. A facial representation F is then extracted from Z_n and conditioned into the end-to-end acoustic model of the VITS architecture (Kim et al., 2021). Following the adaptive conditioning approach of Yoon et al. (2023), F is projected through a linear layer to obtain a gain g, a normalized direction vector $\frac{d}{\|d\|}$, and a bias term b. These parameters are used in a 1-D convolutional layer to modulate the phonemic hidden representations from the VITS text encoder x as $(g \times \frac{d}{\|d\|}) \times x + b$. This fusion mechanism integrates facial identity cues into the linguistic representation, which guides the model to synthesize speech that is both textually accurate and identity-consistent. Since representation conditioning strategies are a separate and active area of research (Yoon et al., 2023; Min et al., 2021; Choi et al., 2022), we do not elaborate on this in further detail. Interested readers are referred to Yoon et al. (2023). In the following subsections, we instead focus on the learning of facial representation F. The full architecture is illustrated in Figure 1.

3.2 Progressive Feature Extraction

Although the primary objective of FTV is to convert facial images into speech, previous literature have predominantly emphasized audio signal decomposition over image processing. This is exemplified by the learning of latent acoustic characteristics from reference audio, which is subsequently used as a target for training the face encoder (Goto et al., 2020; Wang et al., 2022; Kang et al., 2025). Such strategies emphasize global representations and often overlook the rich spatial hierarchies embedded within facial images. Towards this, we incorporate a hierarchical non-pretrained visual encoder originally utilized for image generation (Zhang et al., 2022) that is trained directly from scratch, and results in more precise and robust mapping from facial geometry to vocal identity.

Formally, given a facial input image $Z \in \mathbb{R}^{3\times H\times W}$, the model first decomposes Z into non-overlapping patches of resolution $P \times P$. Each patch $z_{i,j} \in \mathbb{R}^{3\times P\times P}$ is flattened and linearly projected into a fixed-dimensional embedding space, which results in a set of patch tokens:

$$Z = \{z_{i,j} \mid i \in [1, H/P],\, j \in [1, W/P]\} \quad (1)$$

Each patch token is then passed through canonical Transformers (Vaswani et al., 2017) that is composed of alternating multi-head self-attention (MHSA), feed-forward networks (FFNs), and Layer Normalization (LN):

$$\hat{z} = \mathrm{TransformerBlock}(z) \quad (2)$$

Moreover, it is important to conduct aggregation between neighboring granularities $\hat{z}$ in order to form a comprehensive visual representation and to enhance locality. As such, the first facial granularity aggregation (FGA) iteration combines the representations of the four smallest neighboring $P \times P$ patches using a 3×3 convolutional layer, followed by layer normalization, and then 3×3 max pooling. This process expands the model's receptive area from a single local facial patch to a set of four.

$$\tilde{z} = \mathrm{MaxPool}(\mathrm{LN}(\mathrm{Conv}_{3\times3}([\hat{z}_1, \cdots, \hat{z}_n]))) \quad (3)$$

This Transformer–aggregation cycle is applied iteratively across hierarchical levels, where each aggregation stage reduces the number of spatial patches by a factor of four, progressively expanding the receptive field. The process continues until a single spatial token remains, thus capturing the global summary of the entire facial image. The final output is a 128-dimensional embedding that serves as the facial identity vector F.

3.3 Bilateral Attribute-Based Enhancement

While the hierarchical visual encoder described in the previous section effectively captures spatially localized geometric features, it does not account for high-level semantic attributes that humans often leverage when inferring vocal traits from facial appearance. Specifically, psychological studies suggest that demographic cues such as gender and ethnicity influence perceived voice characteristics (Doty, 1998; Kim and Davis, 2010). Thus, to complement the recursive visual learning of facial structure, we introduce an auxiliary attribute-based learning objective that explicitly aligns and enhances high-level facial semantics such as gender and ethnicity into the training process.

Towards this, we employ a pseudo-labeling approach using the pretrained DeepFace model[1] (Serengil and Ozpinar, 2024), which automatically annotates each facial image with predicted demographic categories pertaining to gender and race. Attribute prediction is then conducted across both visual and auditory modalities to enhance cross-modal consistency. On the visual side, the facial representation F is passed through two fully connected layers to predict an attribute, which is then

[1] https://github.com/serengil/deepface

Evaluation Metric	Modality	Comparison Set
MOS	Audio ⇔ Audio	1 : 1
ABX	Audio ⇔ Audio	1 : 4
F2V Alignment	Image ⇒ Audio	1 : 4
Age Alignment	Audio	—
Gender Classification	Audio	—

Table 1: Overview of subjective evaluation methodologies. *Modality*: MOS, ABX, and Face-to-Voice (F2V) Alignment evaluations involve comparing the synthesized audio output with either the ground truth audio or a corresponding face image. *Comparison Set*: Indicates the number of synthesized audio samples that are evaluated against each other to determine the one most closely aligned with the ground truth.

compared against the pseudo-labeled ground truth via a standard cross-entropy loss. Facial embedding F is subsequently integrated into the backbone TTS system using adaptive conditioning techniques from Yoon et al. (2023). On the auditory side, the synthesized waveform of variable lengths is first temporally linearly resampled into a fixed-size vector representation. A separate linear layer then predicts the same set of attributes, which enables semantic supervision in the audio domain. This bidirectional supervision scheme promotes alignment of high-level identity cues across modalities. The overall attribute prediction loss is

$$\mathcal{L}_{\text{attr}} = \mathcal{L}_{\text{face}} + \alpha \cdot \mathcal{L}_{\text{audio}} \tag{4}$$

where $\mathcal{L}_{\text{face}}$ and $\mathcal{L}_{\text{voice}}$ denote the cross-entropy losses for the attribute predictions from the facial and voice modalities, respectively, and are combined through weighted summation. These losses are linearly combined with the original reconstruction loss calculated between ground truth and synthetic mel-spectrograms, along with the KL-divergence, adversarial, and duration losses utilized in VITS (Kim et al., 2021) for training.

Moreover, to enhance robust learning, multi-view data augmentation of the training data is further conducted. This is because previous methodologies have only relied on datasets featuring static relationships (Goto et al., 2020; Lee et al., 2023; Plüster et al., 2021; Kang et al., 2025), wherein each speaker is represented by a single face image paired with corresponding ground truth audio for FTV model training. Recognizing the need for vocal consistency despite variations in head orientation and facial expression, we employ a pretrained s3fd (Zhang et al., 2017) face alignment model to automatically extract face-centered frames from each video in the LRS (Afouras et al., 2018) training dataset. From these frames, five images are randomly selected and paired with the identical corresponding audio to train the previously detailed model. An overview of this process is provided in the upper left side of Figure 1.[2]

4 Experimental Settings

We leverage the trainval subset of the LRS open-source dataset (Afouras et al., 2018) for training, and conduct dataset multi-view augmentation and pseudo-labelling[3] as detailed in Sections 3.2 and 3.3, respectively. Out of a total of 159,511 utterances, the training-validation is split using a ratio of 9:1. Audios are resampled to 16,000 Hz, and we use a filter and window length of 1024, and hop size of 256 for mel-spectrogram processing. For fair comparisons, all experiments adhere to a batch size of 20, and 350,000 training iterations are conducted for approximately two days using a single A6000 GPU. The total number of parameters is approximately 43.6 million.

4.1 Evaluation Protocol

For validation, we juxtapose our method with three baseline systems, and conduct subjective and objective assessments. The first baseline follows the two-stage training method proposed by Plüster et al. (2021), which jointly trains a Global Style Token (Wang et al., 2018) speaker encoder with the TTS backbone. Concurrently, a pretrained visual encoder is finetuned to align with the learned speaker embeddings[4]. The second model is a diffusion-based framework (Lee et al. (2023), FaceTTS) that incorporates an auxiliary speaker alignment loss that encourages the visual and acoustic embeddings of the same identity to be close in a shared latent space. Lastly, FVTTS (Lee et al., 2024) combines both global and local facial representations; a pretrained facial recognition network FaceNet (Schroff et al., 2015) is used to to extract global embeddings, while an additional face encoder is utilized to extract local features from the same input image.

To comprehensively assess the perceptual quality and speaker fidelity of the generated speech, we conduct a range of subjective evaluations via

[2] Further processing details are provided in Appendix A.

[3] We do not use age attributes due to significant bias, as the data is heavily concentrated in the 30–40 age range.

[4] Although the original implementation is based on Tacotron, we re-implement the same methodology using the VITS architecture, and refer to this model as Pluster*.

Model	Subjective Metrics			Objective Metrics			
	MOS (↑)	ABX (↑)	F2V (↑)	Seen SECS (↑)	Unseen SECS (↑)	CER (↓)	UTMOS (↑)
Pluster*	3.25±0.08	20.83%	16.55%	67.04	64.81	0.2381	3.0470
FaceTTS	2.98±0.09	28.45%	25.71%	60.11	56.54	**0.1070**	2.1268
FVTTS	3.20±0.08	19.64%	27.38%	62.50	59.49	0.2743	2.2620
Ours	**3.51±0.09**	**31.07%**	**30.36%**	**79.96**	**71.39**	0.1302	**3.2218**

Table 2: Objective and subjective metric evaluation results. MOS scores are calculated with 95% confidence. F2V denotes the face-to-voice alignment metric.

Model	Younger	Identical	Older
Pluster*	0.225	0.538	0.237
FaceTTS	0.156	0.520	0.324
FVTTS	0.255	0.508	0.237
Ours	0.164	0.649	0.187

Table 3: Age alignment accuracy expressed as a percentage. The inferred age of the voice in the synthesized speech is classified as younger, identical, or older relative to the ground truth audio.

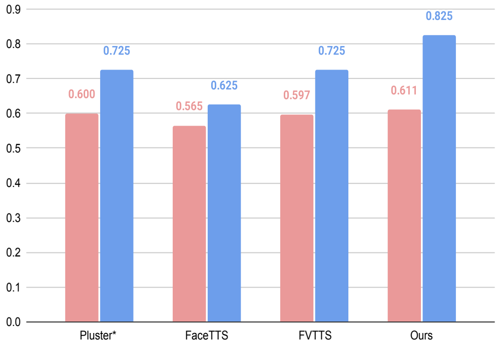


Figure 2: Gender classification agreement (pink) and accuracy (blue) scores in percentage. *Agreement* is quantified by the number of annotator votes assigning the synthesized voice to a specific gender (male or female). *Accuracy* is determined based on the majority vote outcome relative to the ground truth gender.

Amazon Mechanical Turk[5] with twenty-one participants. The first subjective evaluation task involved participants assessing vocal similarity between synthesized and ground truth audio using a 5-point Likert scale (Mean Opinion Scores (MOS)). In the second ABX task, participants selected the audio sample from a set of four (each generated by a different model) that best matched the target speaker's ground truth audio.

While traditional evaluations in this domain typically rely on just MOS and ABX tests, we expand our protocol (Table 1) to include face-to-voice (F2V) and age alignment, and gender classification tasks to better analyze speaker identity preservation. Specifically, the third task of F2V alignment required participants to choose one audio sample out the four audios generated from each of the three baselines and the proposed model, that best matched a given target speaker image. For the age classification task, participants categorized each synthesized audio as having a younger, older, or similar vocal age compared to the ground truth audio. Finally, in the gender classification task, each synthesized audio sample were indicated as either having a male or female voice.

In addition, to support the validity of the subjective evaluation scores, we incorporate a set of automatic objective metrics. Speaker similarity is quantified via cosine similarity between the ground truth target speaker and the synthesized audio (SECS), using the Resemblyzer[6] package. To substantiate the naturalness of the synthesized speech, we employ UTMOS[7] (Saeki et al., 2022), which is a pre-trained model that is designed to predict MOS scores for synthesized speech. Lastly, given that FTV synthesis is fundamentally a speech generation task, maintaining clear and accurate pronunciation is essential. To objectively evaluate pronunciation quality, we transcribe the synthesized speech using a pretrained Wav2Vec-base-960h (Baevski et al., 2020) model, and then compute the Character Error Rate (CER) with the jiwer[8] library.

5 Results and Analysis

5.1 Feature Extraction and Speaker Fidelity

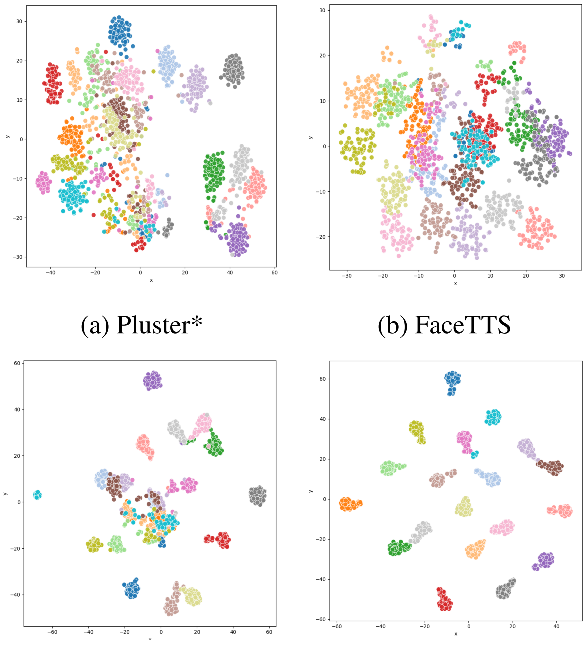
The results in Table 2 clearly demonstrate the superior performance of our proposed model across both subjective and objective evaluation metrics. In

[5] Participants were compensated according to the hourly wage in the authors' country of residence.

[6] https://github.com/resemble-ai/Resemblyzer

[7] https://github.com/sarulab-speech/UTMOS22

[8] https://pypi.org/project/jiwer/


(a) Pluster*

(b) FaceTTS

(c) FVTTS

(d) Ours

Figure 3: t-SNE visualizations of speech embeddings generated using Resemblyzer for seen speakers. Each color denotes a distinct speaker identity. Compact and well-separated clusters indicate higher speaker consistency and better identity preservation. Visualizations of unseen speakers (Figure 6) are included in Appendix C.

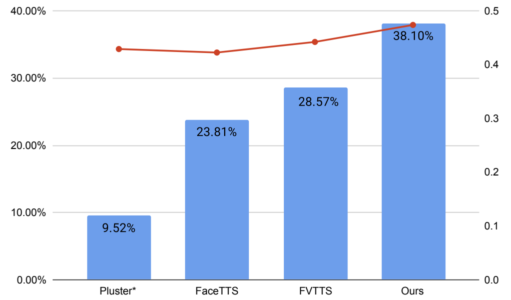


Figure 4: ABX (blue) and SECS (red) scores for the out-of-domain GRID dataset.

terms of MOS, our model achieves a score of 3.51 with a notable 0.26 difference than the closest performing baseline (Plüster et al. (2021), 3.25), which indicates significantly higher speaker fidelity in 1:1 comparisons with ground truth reference recordings. This is further reinforced by ABX preference scores where our model achieves 31.07%, which indicates stronger perceptual similarity to the reference voices in 1:4 comparisons. Most critically, in the F2V alignment task, which evaluates the appropriateness of the inferred voice given only facial inputs, our model records the highest alignment score at 30.36%.

To better understand this perceptual alignment, we further examine age consistency via subjective preference scores across three categories: Younger, Identical, and Older (Table 3). Our model yields the highest preference for the Identical category, achieving 0.649, which is +0.111 higher than the best-performing baseline (Plüster et al. (2021), 0.538). In contrast, baseline models exhibit a greater tendency to generate voices perceived as either too young or too old relative to the target, which suggests unstable synthesis and reduced identity coherence. These results underscore our model's ability to generate voices that are more faithfully aligned with the target speaker's perceived age. Additionally, in terms of gender appropriateness (Figure 2), our model exhibits a 10% improvement in classification accuracy and 2.4-point gain in agreement scores compared to the best-performing baseline.

These subjective findings are strongly corroborated by objective evaluations in Table 2. In SECS, which measures speaker embedding similarity, our model achieves the highest scores for both seen (79.96) and unseen speakers (71.39), with respective margins of 12.92 and 6.58 over the closest baseline model. Although Lee et al. (2023) achieves a lower CER of 0.1070, the difference with our model is minimal at 0.0232. Furthermore, our model performs significantly better than Plüster et al. (2021) and Lee et al. (2024), with CER differences of 0.1079 and 0.1441, respectively. In addition, in terms of naturalness, as measured with UTMOS, our model again leads with 3.2218, which indicates that our synthesized speech is perceived as more natural and human-like. These results are further supported by t-SNE (van der Maaten and Hinton, 2008) visualizations of speaker embeddings across different speakers, where tighter clusters of the same speaker demonstrate speaker consistency (Figure 3).

In order to evaluate generalization to out-of-domain settings, we conducted additional tests using the GRID dataset (Cooke et al., 2006). As shown in Figure 4, our model maintains the highest ABX preference score, with a +10% margin over the second-best model. Additionally, we achieve the top SECS score at 47.38, representing a +3.16 improvement over the next best-performing system, further validating robustness.

	Attribute	Seen SECS (↑)	Unseen SECS (↑)	Overall
Image	Race	77.34	69.37	73.36
	Gender	74.59	68.72	71.66
	Race + Gender	**77.76**	**70.27**	**74.63**
Audio	Race	75.44	67.57	71.51
	Gender	75.28	68.48	71.88
	Race + Gender	**75.84**	**69.52**	**72.48**
Image + Audio	Race	76.14	68.34	72.24
	Gender	75.02	69.83	72.43
	Race + Gender	**79.96**	**71.39**	**75.68**

Table 4: SECS-based speaker fidelity scores for both seen and unseen speakers under varying configurations of attribute-based supervision. Enhancement applied exclusively to the facial representation space is denoted as `Image`, while supervision constrained to the acoustic modality is indicated by `Audio`. The `Image+Audio` setting corresponds to joint attribute enhancement across both visual and auditory domains.

	Seen SECS (↑)	Unseen SECS (↑)	Overall
Ours	79.96	71.39	75.68
- Aud. Enhancement	77.76	70.27	74.63
- Vis. Enhancement	74.73	68.35	71.54
- FGA	66.25	64.31	65.28
- AUG	64.43	62.04	63.23

Table 5: Ablation study results showing the effect of removing each individual component.

5.2 Attribute-based Enhancement

The contribution of attribute-informed supervision, which is implemented via auxiliary prediction objectives for race and/or gender, is quantitatively substantiated by the SECS scores reported in Table 4. When applied exclusively to the facial representation space, unimodal conditioning on either race or gender yields modest improvements in speaker identity preservation, with overall SECS scores of 73.36 and 71.66, respectively. Their joint integration, however, results in an elevated score of 74.63, which is indicative of the complementary nature of multi-attribute cues in enhancing speaker fidelity. In contrast, restricting attribute-based enhancement to the acoustic modality results in a relatively small improvement, which suggests that facial embeddings provide a more robust supervisory signal for identity conditioning. Nonetheless, consistent with the visual-only configuration, combining race and gender supervision within the audio domain still outperforms single-attribute variants for both seen and unseen speaker subsets.

The most pronounced gains emerge when attribute supervision is concurrently applied across both modalities. The `Race+Gender` configuration in this dual-domain setting yields a peak overall SECS score of 75.68, which is a +1.05 improvement over the best-performing unimodal setup. This enhancement generalizes across data partitions, with seen speaker performance increasing from 77.76 to 79.96, and unseen from 70.27 to 71.39. These findings demonstrate the efficacy of multi-domain, multi-attribute conditioning in reinforcing speaker identity consistency in the generated speech.[9]

5.3 Compositional Analysis

We conduct ablation studies to verify the components of the proposed model in Table 5. Eliminating audio-based attribute enhancement results in a noticeable drop of 1.05 points overall, while the further removal of visual enhancement leads to a larger decrease of 3.09 points, which highlights the greater impact of visual attribute alignment on preserving speaker identity. Moreover, we implement a separate experiment that employs just one vanilla transformer and takes the input speaker image as is, without any patch segmentation or feature aggregation (FGA). This results in a 6.26-point drop, which reaffirms the importance of conducting effective initial feature extraction of the input image. Finally, removing the data augmentation (AUG) strategy results in the lowest overall SECS score (63.23), highlighting the necessity of dataset augmentation for achieving robust generalization and speaker consistency. Taken together, these results demonstrate the synergistic effect of each architectural component in optimizing speaker similarity.

6 Conclusion

In this paper, we have presented a novel end-to-end framework for FTV synthesis that emphasizes effective facial representation learning and cross-modal attribute alignment. In contrast to prior methods that depend on multi-stage pipelines or external pretrained models, our approach progressively aggregates local facial features to construct a robust identity embedding and then enforces semantic consistency through bilateral supervision of demographic attributes such as gender and ethnicity. Additionally, by incorporating a multi-view face-audio alignment strategy, we improve the model's ability to maintain vocal consistency across diverse visual inputs. Comprehensive evaluations across five assessment tasks and multiple objective met-

[9] Speaker fidelity scores according to each specific category can be found in Appendix B.

rics confirm that our method significantly improves speaker fidelity and identity preservation, thus offering a viable solution for personalized speech synthesis even in the absence of voice recordings.

Limitations

While our framework advances the generation of identity-consistent speech from facial images, it currently focuses on replicating overall vocal timbre. As such, it does not incorporate facial expressions, which are closely tied to emotional prosody. In future research, we plan to integrate facial expression information—potentially through emotion tagging—to enable expressive and emotionally congruent speech synthesis. Furthermore, although our model includes demographic attributes such as gender and ethnicity via bilateral supervision, age-related features were excluded due to the highly skewed age distribution in the available dataset. Addressing this data imbalance and exploring age-aligned voice synthesis remains an important direction for future work. More broadly, a systematic investigation into the full range of facial attributes that influence perceived vocal characteristics will further enrich personalized and realistic voice synthesis.

Ethical Considerations

The primary objective of this work is to develop a framework capable of synthesizing realistic, identity-consistent speech from facial images. This technology holds promising applications, particularly in assistive communication for individuals with speech impairments or those who have lost vocal capabilities due to neurological or physiological conditions. By enabling the generation of personalized voices, our framework may contribute to restoring a sense of vocal identity and enhancing self-expression for such individuals. Nevertheless, FTV synthesis inherently involves the replication of personal identity cues, and thus raises potential ethical concerns. In particular, the potential for misuse, including unauthorized voice cloning, biometric spoofing, and identity impersonation, poses risks to personal privacy, consent, and digital security. These risks could lead to harmful applications such as fraud, misinformation, or defamation. We acknowledge these risks and emphasize the importance of responsible deployment and strict consent-based usage of FTV systems. Future work should incorporate robust safeguards such as imperceptible watermarking of synthetic speech, which may help to detect and deter misuse of generated voices.

Furthermore, our model incorporates commonly defined demographic attributes such as gender and race during training for improved speaker identity modeling, which was informed by prior studies that indicate their perceptual relevance in voice characteristics. While these features were derived from publicly available datasets and used solely for the purpose of enhancing face-vocal alignment, we recognize the potential sensitivity of such demographic variables. We remain committed to conducting ethical and inclusive research and advocate for ongoing interdisciplinary dialogue between machine learning researchers, ethicists, and affected communities to ensure the responsible development of FTV technologies.

Acknowledgements

This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2024-00437866)(47.5%), Smart Health-Care for Police Officers Program(www.kipot.or.kr) through the Korea Institutes of Police Technology(KIPoT) funded by the Korean National Police Agency(KNPA, Korea)(No. RS-2022-PT000186)(47.5%), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH))(5%).

References

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

Rohan Badlani, Rafael Valle, Kevin J. Shih, João Felipe Santos, Siddharth Gururani, and Bryan Catanzaro. 2023. Multilingual multiaccented multispeaker tts with radtts. In *ICASSP*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12449–12460.

Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Sheng Zhao, and Tie-Yan Liu. 2021. Adaspeech: Adaptive text to speech for custom voice. In *International Conference on Learning Representations (ICLR)*.

Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C. Cobo, Andrew Trask, Ben Laurie, Caglar Gulcehre, Aäron van den Oord, Oriol Vinyals, and Nando de Freitas. 2019. Sample efficient adaptive text-to-speech. In *International Conference on Learning Representations (ICLR)*.

Byoung Jin Choi, Myeonghun Jeong, Joun Yeop Lee, and Nam Soo Kim. 2022. SNAC: Speaker-Normalized Affine Coupling Layer in Flow-Based Architecture for Zero-Shot Multi-Speaker Text-to-Speech. *IEEE Signal Processing Letters*, 29:2502–2506.

Jeongsoo Choi, Joanna Hong, and Yong Man Ro. 2023. Diffv2s: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *he Journal of the Acoustical Society of America*, 120:2421–2424.

Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. 2020. Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings. In *ICASSP*.

F.L. Darley, A. E. Aronson, and J. R. Brown. 1969. Differential diagnostic patterns of dysarthria. In *Journal of Speech and Hearing Research*.

N. D. Doty. 1998. The influence of nationality on the accuracy of face and voice recognition. In *The American Journal of Psychology*.

Andrew W. Ellis. 1989. Neuro-cognitive processing of faces and voices. In *Handbook of Research on Face Processing*.

Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori. 2020. Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image. In *Interspeech*.

Ting-Yao Hu, Ashish Shrivastava, Oncel Tuzel, and Chandra Dhir. 2020. Unsupervised style and content separation by minimizing mutual information for speech synthesis. In *ICASSP*.

Sung-Feng Huang, Chyi-Jiunn Lin, Da-Rong Liu, Yi-Chen Chen, and Hung yi Lee. 2022. Meta-TTS: Meta-Learning for Few-Shot Speaker Adaptive Text-to-Speech. In *EEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yejin Jeon, Solee Im, Youngjae Kim, and Gary Geunbae Lee. 2025. Facilitating Personalized TTS for Dysarthric Speakers Using Knowledge Anchoring and Curriculum Learning . In *Interspeech 2025*, pages 2108–2112.

Yejin Jeon, Yunsu Kim, and Gary Geunbae Lee. 2024. Enhancing Zero-Shot Multi-Speaker TTS with Negated Speaker Representations. In *The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*.

Nan Jiang, Bangjie Sun, Terence Sim, and Jun Han. 2024. Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image. In *33rd USENIX Security Symposium (USENIX Security 24)*.

Frédéric Joassin, Mauro Pesenti, Pierre Maurage, Emilie Verreckt, Raymond Bruyer, and Salvatore Campanella. 2011. Cross-modal interactions between human faces and voices involved in person recognition. In *Cortex*.

Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. 2003. Putting the Face to the Voice: Matching Identity across Modality. *Current Biology*, 13(19):1709–1714.

Minki Kang, Wooseok Han, and Eunho Yang. 2025. Face-StyleSpeech: Improved Face-to-Voice latent mapping for Natural Zero-shot Speech Synthesis from a Face Image. In *ICASSP*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *ICML*.

Jeesun Kim and Chris Davis. 2010. Knowing what to look for: Voice affects face race judgements. In *Visual Cognition*.

Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. 2023. Imaginary Voice: Face-styled Diffusion Model for Text-to-Speech. In *ICASSP*.

Minyoung Lee, Eunil Park, and Sungeun Hong. 2024. FVTTS: Face Based Voice Synthesis for Text-to-Speech. In *Interspeech*.

Songju Lei, Xize Cheng, Mengjiao Lyu, Jianqiao Hu, Jintao Tan, Runlin Liu, Lingyu Xiong, Tao Jin, Xiandong Li, and Zhou Zhao. 2024. Uni-dubbing: Zero-shot speech synthesis from visual articulation. In *ACL*.

Shiguang Liu and Huixin Wang. 2023. Talking Face Generation via Facial Anatomy. In *ACM Transactions on Multimedia Computing, Communications and Applications*.

Jiří Mertl, Eva Žáčková, and Barbora Řepová. 2018. Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis. In *Disabil Rehabil Assist Technol*.

Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. 2021. Meta-stylespeech : Multi-speaker adaptive text-to-speech generation. In *International Conference on Learning Representations (ICLR)*.

Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, and Abhinav Shrivastava. 2024. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5292–5302.

Davide Mulfari, Gabriele Meoni, Marco Marini, and Luca Fanucci. 2021. Machine learning assistive application for users with speech disorders. In *Applied Soft Computing*.

Arsha Nagrani, Joon Son Chung, Samuel Albanie, and Andrew Zisserman. 2020. Disentangled Speech Embeddings using Cross-modal Self-supervision. In *ICASSP*.

Yasuhito Ohsugi, Daisuke Saito, and Nobuaki Minematsu. 2018. A Comparative Study of Statistical Conversion of Face to Voice Based on Their Subjective Impressions. In *Proc. Interspeech 2018*, pages 1001–1005.

Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. 2022. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*.

Björn Plüster, Cornelius Weber, Leyuan Qu, and Stefan Wermter. 2021. Hearing Faces: Target Speaker Text-to-Speech Synthesis from a Face. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 757–764.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *ICLR*.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Interspeech*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Sefik Serengil and Alper Ozpinar. 2024. A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, 17(2):95–107.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In *Proceedings of ICASSP*.

Zheng-Yan Sheng, Yang Ai, Yan-Nian Chen, and Zhen-Hua Ling. 2023. Face-Driven Zero-Shot Voice Conversion with Memory-based Face-Voice Alignment. In *MM '23: Proceedings of the 31st ACM International Conference on Multimedia*.

Sanjana Sinha, Sandika Biswas, and Brojeshwar Bhowmick. 2020. Identity-Preserving Realistic Talking Face Generation. In *International Joint Conference on Neural Networks (IJCNN)*.

Harriet M. J. Smith, Andrew K. Dunn, Thom Baguley, and Paula C. Stacey. 2016. Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology*, 14(1):1474704916630317.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. In *Journal of Machine Learning Research*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Katharina von Kriegstein, Andreas Kleinschmidt, Philipp Sterzer, and Anne-Lise Giraud. 2005. Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, 17(3):367–76.

Jianrong Wang, Zixuan Wang, Xiaosheng Hu, Xuewei Li, Qiang Fang, and Li Liu. 2022. Residual-guided personalized speech synthesis based on face image. In *ICASSP*.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35 th International Conference on Machine Learning*.

Chao Xu, Yang Liu, Jiazheng Xing, Weida Wang, Mingze Sun, Jun Dan, Tianxin Huang, Siyuan Li, Zhi-Qi Cheng, Ying Tai, and Baigui Sun. 2024. Facechain-imagineid: Freely crafting high-fidelity diverse talkingfaces from disentangled audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhihan Yang, Zhiyong Wu, Ying Shan, and Jia Jia. 2023. What Does Your Face Sound Like? 3D Face Shape towards Voice. In *AAAI*.

Hyungchan Yoon, Changhwan Kim, Seyun Um, Hyun-Wook Yoon, and Hong-Goo Kang. 2023. Sc-cnn: Effective speaker conditioning method for zero-shot multi-speaker text-to-speech systems. *IEEE Signal Processing Letters*, 30:593–597.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. 2017. S^{3}fd: Single shot scale-invariant face detector. In *ICCV*.

Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan O. Arik, and Tomas Pfister. 2022. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In *AAAI*.

Algorithm 1: Face Extraction and Speaker-Wise Pseudo-Label Assignment

```
Input  : Video $\mathcal{V}$
Output : Face images $\mathcal{I} = \{I_1, \ldots, I_5\}$,
         Attribute label $\mathcal{A}$
Step 1: Frame Sampling
$\mathcal{F} \leftarrow$ s3fd($\mathcal{V}$)
$\mathcal{F} \leftarrow$ shuffle($\mathcal{F}$)
$\mathcal{I} \leftarrow \mathcal{F}[:5]$
Step 2: Pseudo-Label Generation
$\mathcal{A} \leftarrow$ NULL
foreach $I_k \in \mathcal{I}$ do
    if $\mathcal{A}$ = NULL then
        $(\mathrm{attr}_1, \mathrm{attr}_2) \leftarrow$ DeepFace($I_k$) ;
        $\mathcal{A} \leftarrow [\mathrm{attr}_1, \mathrm{attr}_2]$ ;
    end
end
return $\mathcal{I}$, $\mathcal{A}$
```

A Dataset Processing

The video from a speaker can be viewed as a sequence of individual frames, each representing a still image. From these images, the face is extracted using the s3fd model. Five frames are randomly selected for further processing. After selecting the five images, one image from each speaker is annotated using the DeepFace pretrained model, as described in Algorithm 1. The remaining four images inherit the annotated attributes of the first image. This process is done to ensure attribute annotation consistency across the images of the same speaker. Since the LRS dataset is used solely for research purposes, this usage is consistent with the Creative Commons Attribution 4.0 International License.

B Sub-attribute Category Evaluations

To incorporate high-level attributes across both audio and visual modalities, we have introduced a bilateral enhancement strategy[10]. To assess the robustness of our method across demographic subgroups, we analyze speaker similarity (SECS) scores by gender and race[11] using pseudo-attribute annotations derived from a pretrained `DeepFace` model on the LRS dataset. As summarized in Table 6, the majority of the face images are classified as male (74.66%) and predominantly Caucasian (72.23%). Despite this imbalance, our model maintains consistent speaker similarity across subgroups, as shown in Table 7. Specifically, the SECS scores for male (75.97) and female (75.03) speakers are closely aligned, which suggests that the model does not disproportionately favor the over-represented gender.

Attribute	Category	Percentage
Gender	Male	74.66%
	Female	25.34%
Race	Caucasian	72.23%
	Asian	12.76%
	Middle Eastern	5.64%
	African American	4.72%
	Latino Hispanic	4.33%
	Indian	0.32%

Table 6: Dataset statistics according to gender and race attributes.

Attribute	Category	Spk Sim
Gender	Male	0.7597
	Female	0.7503
Race	Caucasian	0.7593
	Asian	0.7637
	Middle Eastern	0.6925
	African American	0.7165
	Latino Hispanic	0.7867
	Indian	0.7631

Table 7: Speaker similarity scores across gender and race.

A similar trend is observed across race categories; while Caucasian speakers dominate the dataset, the similarity scores for underrepresented groups such as Asian (76.37), Indian (76.31), and Latino Hispanic (78.67) are comparable or even superior. Notably, the performance for Middle Eastern (69.25) and African American (71.65) speakers, though slightly lower, still remains within an

[10] The weight assigned to classification in the acoustic domain is empirically set to 0.3.

[11] The pretrained DeepFace model originally labels race attributes using terms such as "White" and "Black." For improved clarity and alignment with academic conventions, we adopt the terms "Caucasian" and "African American" in this paper.

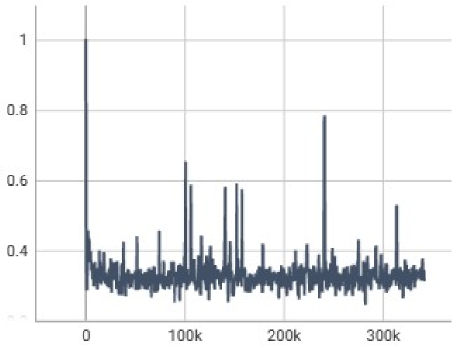
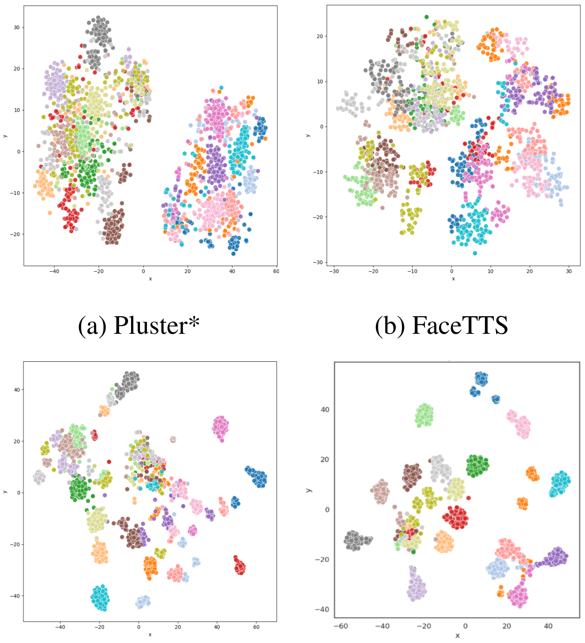


Figure 5: Visualization of collective modality attribute loss.



(a) Pluster*

(b) FaceTTS

(c) FVTTS

(d) Ours

Figure 6: t-SNE visualizations of speech embeddings generated using Resemblyzer for unseen speakers.

acceptable range, which indicates generalizability beyond the majority class. These findings are further supported by the progression of the attribute classification loss (Figure 5); its continuous decline and eventual plateau indicate stable convergence and the effectiveness of incorporating attribute supervision in improving face-to-voice alignment.

C Subjective Evaluations

To comprehensively assess the perceptual quality and speaker fidelity of the generated speech, each model was evaluated using 100 audio samples spanning 40 speakers, comprising of both seen and unseen identities. The evaluation process required approximately two hours per annotator. Among the twenty-one annotators, the gender distribution was seventeen male and four female. Prior to evaluation, all participants were informed that their demographic information and responses would be used solely for research purposes. To further substantiate our subjective assessments, we additionally provide objective visualizations of the unseen speakers in Figure 6.