HarmoniVox: Painting Voices to Match the Avatar's Soul

Songtao Zhou
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Xiaoyu Qin*
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Yixuan Zhou
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China

Qixin Wang
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Zeyu Jin
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Zixuan Wang
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Zhiyong Wu
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China

Jia Jia*
Department of Computer Science and Technology,
BNRist, Tsinghua University
Beijing, China

Abstract

Imagine James Bond speaking like Mr. Bean—such a mismatch would create a jarring dissonance and break the viewer's immersion. Current research on virtual avatar animation has focused on modeling 3D geometry, appearance, motion generation, however, neglecting the harmony between speech prosody and the avatar's visual presentation and contextual environment. In this paper, we seek to bridge this gap by firstly identifying and defining the key elements necessary for achieving audiovisual harmony, such as appearance, expression, body posture, backgrounds and colors. Subsequently, we propose a method that jointly models semantic consistency in avatar animation, named HarmoniVox, specifically on *crafting prosodic speech consistent with the avatar's essence from given visual image*. To achieve this, we implement a technical framework with a mutual modal contrastive learning strategy, enhancing multimodal alignment in a coarse-to-fine fashion. To support this method, we establish a experimental dataset HarAvaSpeech comprising 28,929 image-audio pairs, designed to encompass expressive speech prosody and rich avatar visual presentations across a wide range of contexts. Leveraging this dataset, our experiments demonstrate that the proposed method outperforms the baselines in manipulating the nuanced tone and harmonious rhythm of speech with the avatar visual presentations, and reveal generalizability on out-of-domain cases. Demo would be provided in https://harmonivox.github.io/harmonivox/.

CCS Concepts

• **Applied computing** → **Media arts**; • **Information systems** → **Multimedia content creation**.

Keywords

Virtual Avatar Animation, Audiovisual Harmony, Multi-Modal Contrastive Learning

ACM Reference Format:
Songtao Zhou, Xiaoyu Qin, Yixuan Zhou, Qixin Wang, Zeyu Jin, Zixuan Wang, Zhiyong Wu, and Jia Jia. 2025. HarmoniVox: Painting Voices to Match the Avatar's Soul. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3746027.3755736

1 Introduction

In virtual avatar animation, the harmony between speech and visual presentation is as crucial as geometry modeling, appearance generation and motion control [49]. For example, would James Bond speaking in Mr. Bean's tone still appear as *a tough, charismatic spy in secret operation*? Or, would Mulan talking like Snow White's voice still come across as the *determined warrior in battlefields*? The character's true essence, or "soul", such as the role settings, scenario settings and action settings, would be lost in this mismatch. Therefore, it is significantly crucial to maintain the harmony between the avatar voice with the internal attributes (such as *appearance*, *personalities*, etc.) reflected in visual presentations, to better conveying the unified audiovisual experiences, as shown in Fig. 1.

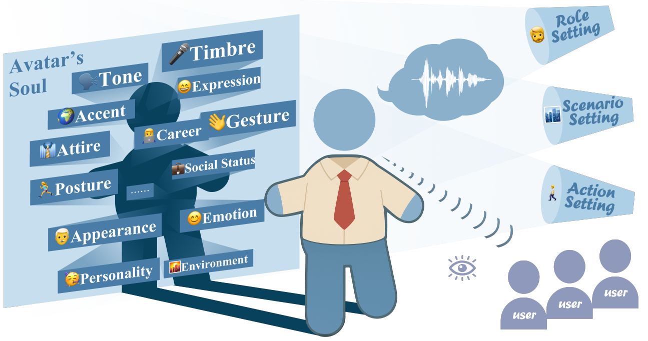


Figure 1: Overview of virtual avatars attributes. In practical applications of virtual avatars, all these attributes should not only maintain consistency with one another but also align with the contextual settings of *role, scenario, action* to achieve a harmonious human-computer interactions.

[*] Corresponding authors: xyqin@tsinghua.edu.cn and jjia@tsinghua.edu.cn.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755736

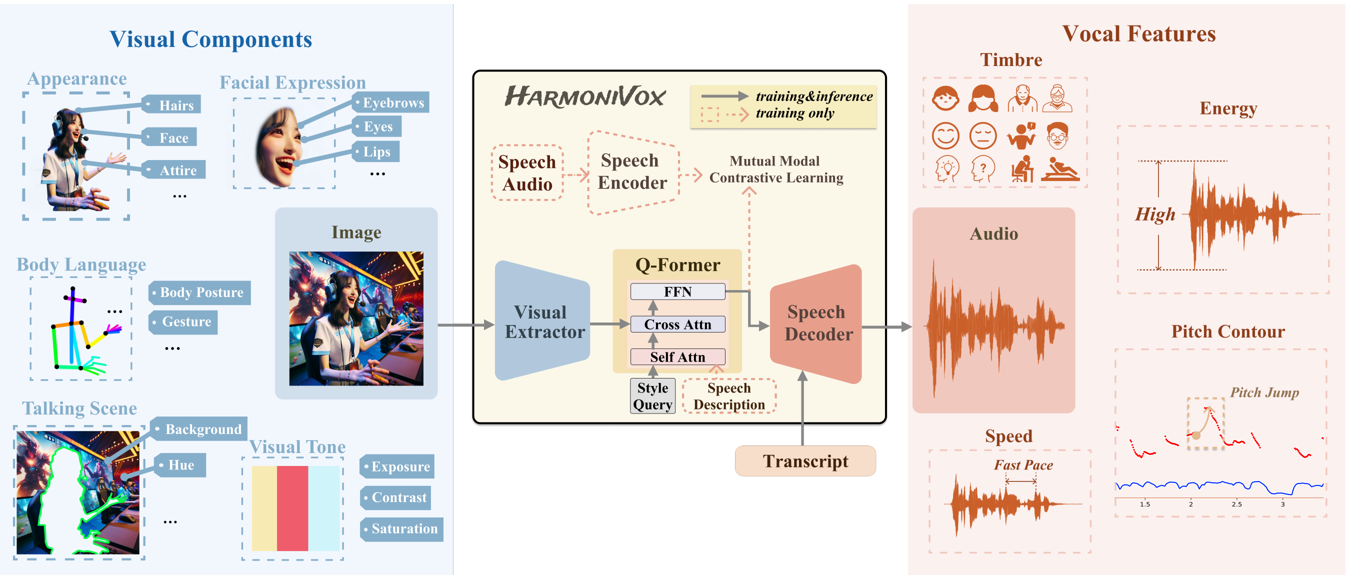


Figure 2: Overview of HarmoniVox, a model positioned in two modalities implicitly extracts the visual contexts on the leftmost, such as *appearance*, *posture*, and *expression*, then aligns them with vocal features shown on the rightmost, such as *pitch* and *energy*, to achieve the audiovisual harmony. The solid gray arrows represent the data flow during the training and inference processes, while the dashed arrows and the modules enclosed in dashed lines indicate components used only for training.

In this paper, we first seek to investigate the specific aspects of harmony between voice and visual presentation that impact human perception. Some existing research has (1) generally proven, from a stimuli perspective, that voice affects the attractiveness of characters [49], and (2) examined the relationship between voice and visual factors (such as facial skeletons) from a human cognition perspective [1, 6, 18, 27, 31, 48, 52]. Inspired by these research, we identify and define the (a) five speech features that crucial for emotional and intentional communication: *timbre*, *rhythm*, *energy*, *emotion* and *topic*, and (b) five major components in visual presentation that related to speech features: *appearance*, *expression*, *posture*, *scene*, and *colors*. To gain a deeper insight into the underlying mechanism, we conduct semi-structured interviews and subjective questionnaires to identify the visual components that must remain consistent for each speech feature. Our interview results and survey findings highlight the importance to ensure the audiovisual harmony of user perception through the combination of explicit consistency and implicit consistency.

Derived from the survey findings, we propose a novel method to jointly modeling the multi-modal consistency for virtual avatar, specialized on synthesizing harmonious speech with the key elements reflected in given visual images. Our approach mainly address issues of the cross-modal alignment from two aspects. *Modality harmonization*: The inherent semantic mismatch between speech and image modalities makes it challenging to achieve multi-modal harmony. Inspired by the success of contrastive learning in image captioning [40, 41] and speech captioning [66], we introduce Q-former and a three-stage training pipeline for speech style-related visual representation learning. *Coarse-to-fine alignment*: Despite the great potential of contrastive learning in image/speech captioning tasks, the modality differences between image and speech are far greater than those between either modality and text. To enhance learning efficiency, we propose a mutual modality contrastive learning (MMCL) strategy, which integrates coarse supervision from text and fine supervision from audio to guide the Q-former in extracting speech-style related visual representations. We implement a technical framework model to address the problem in an end-to-end fashion, as illustrated in Fig. 2.

To support this method, we establish a bilingual image-speech avatar dataset encompassing rich speech expressiveness and diverse visual scenes, with data from over 10,000 speakers. As shown in Tab. 1, existing avatar datasets have limitations in visual contents and speech expressiveness. To address these limitations, we leverage Large Language Models to synthesize paired visual presentation with highly-expressive emotional speech corpora, resulting in a multimodal speech dataset HarAvaSpeech. To our knowledge, HarAvaSpeech is the first bilingual multimodal dataset for avatar animations that focus on the overall harmony between the visual cues and speech styles. Leveraging the HarAvaSpeech dataset, extensive experiments on the HarAvaSpeech demonstrate that our method outperforms the baseline method in harmonious avatar speech synthesis. The ablation studies have verified the effectiveness of MMCL strategy in boosting the cross-modal alignment. Comparative experiments have been conducted between the MEAD-TTS [60] and HarAvaSpeech, evaluated on the out-of-domain test set from Web. The results indicate that the model trained on HarAvaSpeech significantly outperforms MEAD-TTS [60] in gender and emotion accuracy, verifying the effectiveness and superiority of HarAvaSpeech dataset. To further illustrate the generalizability of HarmoniVox, a case study on avatar speaking transcripts with distinct visual presentations is provided in Sec. 6.4.

Summarily, the contributions of this paper are as follows:

- **We propose HarmoniVox**, a novel cross-modal modeling method for virtual avatar animation, concentrating on manipulating speech features with given visual images.
- **We define the key framework of harmony between visual images and speech**, specifying five key visual components related to speech and their consistency with each speech feature.
- **We propose a multi-modal contrastive learning strategy**, boosting the model in harmonizing a more nuanced tone and rhythm of synthesized speech for both intra-domain and out-of-domain inputs.
- **We establish HarAvaSpeech**, an AIGC-based multimodal visual-text-speech dataset that collects the body and scene visual context for the first time.

Table 1: A summary of multimodal avatar datasets. The source can be categorized into YouTube, Studio, and AIGC. Face, Body, and Scene indicate the corresponding visual information is provided or not.

Dataset	Source	Language	Face	Body	Scene	Speaker
VoxCeleb2 [10]	YouTube	EN	✓	✗	✗	6,112
MEAD [60]	Studio	EN	✓	✗	✗	60
MMFace4D [63]	Studio	ZH	✓	✗	✗	465
HDTF [71]	YouTube	EN	✓	✗	✗	> 300
MEAD-TTS[1] [23]	Studio	EN	✓	✗	✗	47
HarAvaSpeech	AIGC	ZH+EN	✓	✓	✓	> 10,000

[1] MEAD-TTS is derived from MEAD dataset.

2 Related Works

The section first reviews the cognitive study on the visual-vocal relationship for human behaviors and then examines the relevant research for virtual avatar animation.

2.1 Audiovisual Consistency in Human Behaviors

The natural consistency between voice and visual identity has always been the hotspot in cognitive science. Research has shown that the voice provides comparable information of identity as facial features do [31]. For instance, the facial skeletal measurements have been found to significantly correlate with F0 and habitual frequency [52]. Additionally Evans et al. [18] discover a significant negative relationship between fundamental frequency and measures of body shape and weight. These findings suggest that people can match the unfamiliar voice with static face images [48]. Horiguchi et al. [27] further demonstrated that distances between facial features and audio features can be utilized to to classify the best-matched face according to audio clips.

Aside from timbre-related skeletal features, facial and body movements also play a significant role in shaping perception. Studies have pointed out that the static posture [12, 51], as well as the direction and intensity of body sway [6], convey different emotion states. For example, individuals with positive emotions tend to lean forward and shown activation, while those with negative emotions exhibit opposite behaviors [3]. Co-speech gestures, as shown by Kelly and Tran [33], provide both emotional and cognitive functions in various communicative contexts. Another form of implicit consistency can be found in aesthetics considerations that preferences for color hues, such as saturation, contrast and brightness, can reflect individual personalities and social contexts [35].

In this paper, we address both implicit and explicit consistency within our defined framework, guided by insights from existing literature and our empirical studies in Sec. 3.

2.2 Audiovisual Consistency for Virtual Avatar

The preference for harmony is rooted in human cognitive mechanisms: research from psychology has shown people tends to rely on cross-modal correspondences to perceive real-life objects [47, 54]. Therefore, mismatches between multi-modalities can interrupt their perception of the avatar [9]. This notion is not new—during the era of World War I, Kandinsky et al. [32] have emphasized that the sound, color, movements should work in harmony to serve the creator's intent [56]. Recent investigations into the interaction between modalities reveal how different aspects contribute to avatar perception. Ondřej et al. [49] explored how voice, face motion, body motion and appearance affect the distinctiveness and attractiveness of characters. Ennis et al. [17] suggested that combining facial and body motion can enhance the user's perception of an avatar's emotional state. Ferstl et al. [19] argued that maximizing the realism of speech and motion is preferable even when it leads to a mismatch with the appearance realism.

However, existing research is mostly limited to highlighting the necessity of multi-modal harmony. The specific contribution of each pair of visual components and speech features to the overall harmony remains an open question.

2.3 Virtual Avatar Animation

Recent research on the virtual avatar animation primarily focus on the technological advancements in modeling of 3D geometry [5, 16, 42, 43], facial motion [13, 22, 24, 29, 45, 57–59, 63, 64, 70], body motion [62, 68, 69] and appearance [65]. To tackle the audiovisual harmony between speech and visual presentation, research has primarily focused on speech-driven talking face and face-based speech synthesis. Studies have explored on rendered-based generation with 3D geometry priors [57, 63, 64, 70] and video-based generation with diffusion prior [13, 29, 45, 58]. Some research has explored inferring timbres from human face [21, 23, 38, 44, 67]. Face2Speech [21], Face-VC [44] and Face-TTS [38] adopts face embedding as a substitute for speaker embedding in speech synthesis. MM-TTS [23] adopts the emotion and gender of the speaker inferred from facial close-ups as implicit prompts for text-to-speech under a multi-modal prompt framework. These aforementioned works use 2D image-based methods, while 3D facial skeletal structures are reconstructed anatomically by Yang et al. [67] for better timbre control.

Their investigation of the harmony of speech and visual images is relatively insufficient. In this paper, we innovatively include the comprehensive visual image (not just the face) in the consideration of audiovisual harmony.

3 Specifying Harmonious Audiovisual Consistency

In this section, we seek to answer the following question: *what specific aspects of visual and audio consistency need to be considered in virtual avatar?* Based on the existing literature, we first define the key visual components and speech features. Subsequently, we

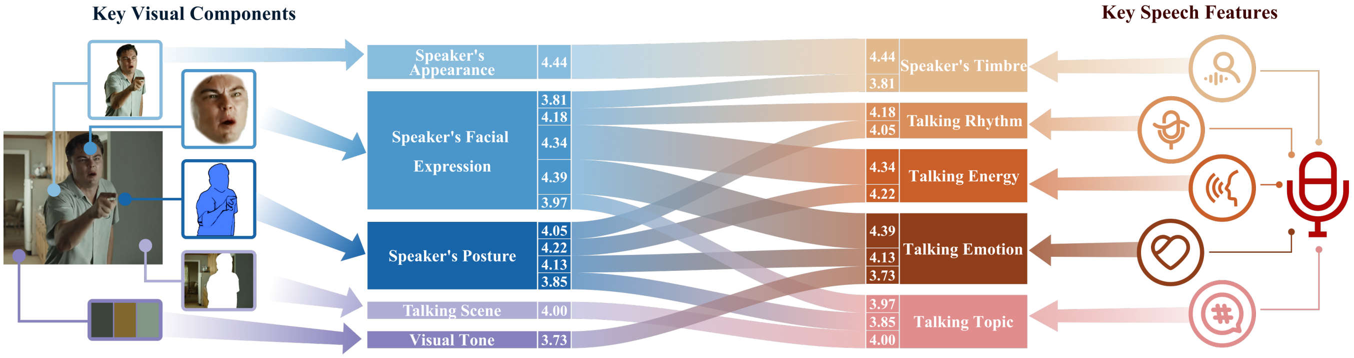


Figure 3: The framework illustrates the underlying relationship between visual components and vocal characteristics in Sankey diagram. The left side shows the key visual components of an image, and the right side displays the key speech style attributes. Every flow refers to a correlation between two ends, with the width representing the strength of correlation.

recruited twelve individuals to conduct a twenty-minute interview on their opinions about the visual-vocal relationship, and then come up with a follow-up user survey for further investigation.

3.1 Harmony Framework

Among the attributes illustrated in Fig. 1, we select the following ***five key visual components***, as shown in left of Fig. 3: ***(1) Appearance*** includes physical features, body shape, and attire, which serve 0as the external manifestation of the role. It encompasses attire that indicates age and gender, as well as clothing that may suggest the character's career and scenario. ***(2) Expression*** refers to facial expressions, which serve as the external manifestation of the action. ***(3) Posture*** includes body posture and hand gestures, which represent the external manifestation of the action. ***(4) Scene*** includes background elements, which serve as the external manifestation of the scenario. ***(5) Color*** refers to the lighting, brightness, and color tones in the background, serving as the external manifestation of scenarios. As for the speech features, we identify ***five speech features*** inspired by existing text-to-speech research [25, 28, 30, 39], as shown in right of Fig. 3: ***(1) Timbre*** refers to the unique quality or tone of a voice that helps identify one speaker from another. ***(2) Rhythm*** pertains to the speed and cadence of speech. ***(3) Energy*** refers to the volume or loudness. ***(4) Emotion*** is a higher-level feature, capturing the affective state conveyed through voice. ***(5) Topic*** relates to the content of the speech itself.

3.2 Empirical Study

For the interview, we invited 12 individuals aged 22 to 26 with foundational knowledge in areas such as human factors, psychology (art/design psychology), design, art, aesthetics, or experience in audiovisual artistic creation and design work. During the semi-structured interview, we requested participants to identify which visual components are expected to align with each given speech feature and to provide explanations. We organized the results by computing the selection rate for each pair of visual and vocal elements. Detailed results are provided in Appendix Fig. 4. We define pairs with more than 50% approval as *explicit consistency*, e.g., timbre and appearance, indicating that the consistency is widely recognized and explicit. As for pairs with less than 50% but more than 25% approval, e.g., emotion and color , we define them as *implicit consistency*, as suggested by the participants.

To verify the interview findings, we conducted a quantitative survey on clips from multiple web platforms[1]. Twenty pairs of samples are collected and processed into questionnaires, with the sample source ranging across movies, TV series, talk shows, lectures, and interviews. We select the most expressive keyframe image that captures the character essence and corresponding audio clips and ask another group of participants to rate them on a scale of one to five, from *not consistent* to *highly consistent*. For each visual-vocal attribute pair, we define the average rates over samples and participants as the relevance value of the pair and construct the relevance framework in Fig. 3, where only the highest 50% relevance is preserved. The survey interface and the original score matrix are provided in App. A (supplementary materials). Combined with the interview findings, our results verify the existence of explicit consistency, such as *posture* - *emotion* pairs and *posture*, and implicit consistency, such as *color* - *emotion*. Although the *scene* is mostly related to the *topic*, we will consider this during the dataset construction process but will not address it in the methodology section. The generation of speech content will be left for future work. In the following sections, we will utilize these findings to further enhance our research.

4 Painting Voice to Match Avatar's Soul

This section elaborates on the method for multi-modal semantic consistency modeling, HarmoniVox, with a particular focus on how we address modality harmonization and coarse-to-fine alignment.

4.1 Problem Formulation

We first formulate the avatar animation problem in the context of a given visual image and content transcripts. Let I be a visual image of avatar, containing the five visual components: appearance I_a, facial expression I_f, body posture I_b, scene component I_s and color I_c. Let $T = \{c_1, c_2, \ldots, c_n\}$ be the transcripts, a sequence encompassing multi-lingual characters. Our target is to implicitly infer the inner states of avatar and cast them into talking style S based on the visual image I. We then synthesize speech audio A with style S and

[1] http://youtube.com; http://bilibili.com; http://youku.com.

content T, as shown in Eq. (1).

$$A = f(S, T) = f(S(I_a \cup I_f \cup I_b \cup I_s \cup I_c), T), \quad (1)$$
$$A = g(S, T) = g(S(I_a \cap I_f), T), \quad (2)$$

While previous methods [21, 23, 67] focus on facial appearance $I_f \cap I_a$, i.e., Eq. (2), this paper take a comprehensive approach by considering the visual representations $I_a \cup I_f \cup I_b \cup I_s \cup I_c$. Methods like this can be integrated with speech-driven animation techniques using the visual image, resulting in the final avatar videos.

4.2 Modality Harmonization

To tackle the modality harmonization, as illustrated in Fig. 2, we propose the HarmoniVox consisting of three modules: visual extractor, Q-former, and speech decoder. For the visual contexts capture, we adopt CLIP image encoder [53] as the contexts extractor. Pretrained with over 0.4B pairs of samples, the CLIP model shows an outstanding performance and generalizability in semantic extraction and zero-shot classification [53]. For the speech style fusion, we adopt VITS model [34, 37] as the speech backbone. Pretrained with large-scale speech corpora, the VITS model demonstrates high-fidelity speech reconstruction ability. To efficiently harness these advantages, we introduce a Q-former with a contrastive learning paradigm, which have demonstrated their effectiveness in cross-modal alignment [40, 41, 66]. However, to our best knowledge, we are the first to apply them to tackle the visual-vocal semantic alignment issues. Taking the above into consideration, we design a three-stage learning framework as follows (additional illustrations are provided in App. B):

Stage I: Unsupervised Speech Style Learning. In this stage, only pure speech data and VITS model are involved. To endow the VITS model with powerful speech modeling capability, the speech backbone including the speech encoder and speech decoder is first unsupervisedly trained on large-scale emotional speech data. For simplicity, we annotate all the training loss in this stage, including the reconstruction loss and other items [37], as $\mathcal{L}_{vits}$.

Stage II: Speech Style-Related Visual Representation Learning. In this stage, the paired data (image, speech) is utilized. The Q-former is guided to learn the visual representation associated with speech features, from both the speech encoder and CLIP image encoder. We introduce the MMCL strategy, employing the coarse speech description to enhance the guide the Q-former to learn the speech style-related visual information.

Stage III: Vision-conditioned Speech Style Control. In this stage, both the Q-former and speech decoder are trainable. The Q-former is guided to learn the visual representation associated with speech features, from the speech decoder. We directly connect the Q-former with the speech decoder and employ $\mathcal{L}_{vits}$ as the training objects on image-audio pairs.

4.3 Mutual Modal Contrastive Learning

The MMCL strategy is mainly applied in Stage. II, in which the speech encoder and the image encoder are both frozen. There are two sets of guidance in MMCL: Mutual Modal Contrastive Loss and Mutual Modal Matching Loss. The former focuses on aligning representations within a shared semantic space, while the latter primarily relies on a matching classifier to guide the multimodal fusion [40, 41]. As shown in Fig. 4, there are three modal-specific embeddings: image embedding E_i, audio embedding E_a and text embedding E_t. The E_i and E_t are both the output of Q-former[2], while the E_a is the output of speech encoder.

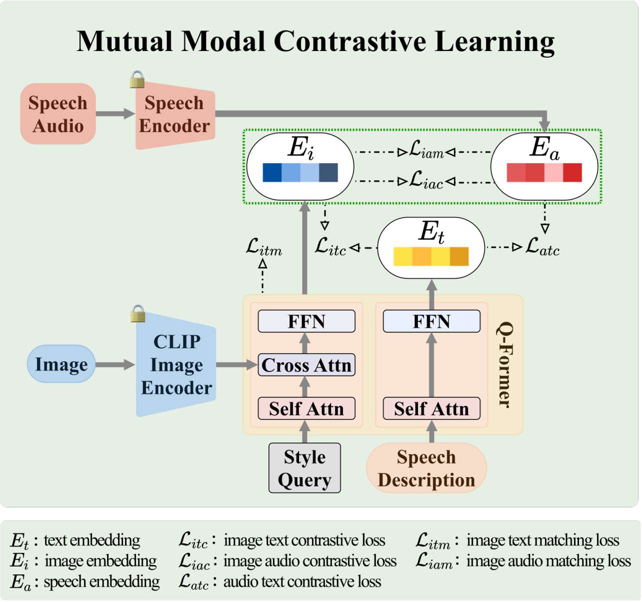


Figure 4: Illustration of Mutual-Modal Contrastive Learning strategy in Stage.II. The Q-former learns to extract speech-related visual presentation from the supervision of ground-truth speech audio and speech descriptions. The legend at the bottom displays the losses and embeddings involved.

Intuitively, our goal is to align the image embedding E_i and the speech embedding E_a from the frozen speech encoder, as shown in Fig. 4, We consider the image and audio paired in the dataset as the matched pair and labeled them as positive, with the random pairs considered as the unmatched pair and labeled as negative. Since E_i and E_a are both normalized, we employ the cosine similarity of them as the logits for contrastive learning and donate the image-audio loss as $\mathcal{L}_{iac}$. To retrieve a finer-grained alignment of image and audio, we utilize an extra set of linear layers to fuse the E_i and E_a, with an output linear layer to obtain the 2-class (positive/negative samples) logits for the image-audio matching loss $\mathcal{L}_{iam}$.

Vanilla contrastive learning strategy by BLIP 2 [40] and SECap [66] only deals with contrastive loss and matching loss across two modalities. However, as the empirical findings suggest, both explicit and implicit consistency should be considered. We leverage the textual vocal descriptions to assist in training the image-audio alignment, serving coarse speech styles to enhance the explicit consistency such as appearance-timbre, expression-emotion. Inspired by [40], the contrastive loss $\mathcal{L}_{itc}$ are derived from the uni-model output of Q-former, E_i and E_t, since the style queries are forbidden to attend to textual descriptions in this case. The matching loss $\mathcal{L}_{itm}$ are derived from the bi-directional output of Q-former, in which case the style queries and textual descriptions are concatenated together and allowed to attend to each other [40]. As for

[2] The Q-former implementation of deriving image and text embedding is similar to BLIP 2 in https://github.com/salesforce/LAVIS/tree/main/projects/blip2.

text-audio contrastive learning, we adopt only the contrastive loss $\mathcal{L}_{atc}$ for simplicity.

The total training loss for MMCL is

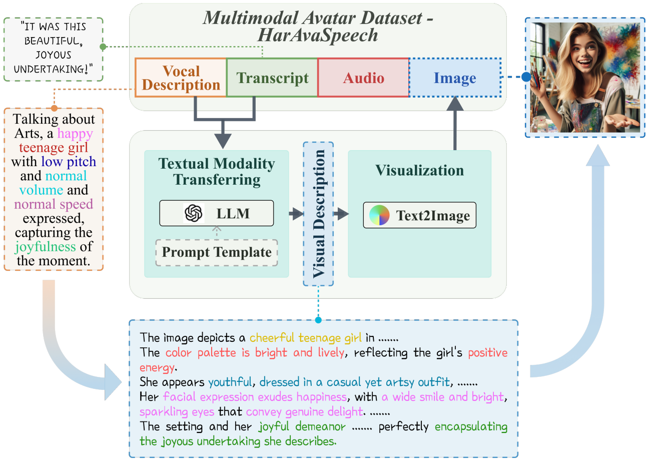
$$\mathcal{L} = \lambda_{ia}(\mathcal{L}_{iam} + \mathcal{L}_{iac}) + \lambda_{it}(\mathcal{L}_{itm} + \mathcal{L}_{itc}) + \lambda_{at}\mathcal{L}_{atc}. \quad (3)$$

Considering the coarse-to-fine supervision of textual description and speech audio, we set an annealing function for λ_{ia} as we found a fixed value for λ_{ia} would downgrade the performance. Starting from zero, the weight of $\mathcal{L}_{ia}$ increases gradually and reaches 1.

5 Augmenting Multimodal Dataset

This section presents our dataset construction pipeline instructed by the defined key audiovisual framework of avatar harmony (Sec. 3).

As shown in Tab. 1, typical multimodal datasets for avatar animations such as MEAD [60] and MMFace4D [63] are collected in recording studios, offering *limited variations on scenario and role settings*. Meanwhile, face-speech datasets collected from YouTube, such as VoxCeleb2 [10] and HDTF [71] provide only facial information, *missing the multi-consistency associated with the body movements and surrounding backgrounds*. To overcome the scarcity of high-quality multimodal data, we turn to explore novel data sources and construction methods.


"IT WAS THIS BEAUTIFUL, JOYOUS UNDERTAKING!"
Multimodal Avatar Dataset - HarAvaSpeech
Vocal Description
Transcript
Audio
Image
Talking about Arts, a happy teenage girl with low pitch and normal volume and normal speed expressed, capturing the joyfulness of the moment.
Textual Modality Transferring
LLM
Prompt Template
Visual Description
Visualization
Text2Image

The image depicts a cheerful teenage girl in ....
The color palette is bright and lively, reflecting the girl's positive energy.
She appears youthful, dressed in a casual yet artsy outfit, ...
Her facial expression exudes happiness, with a wide smile and bright, sparkling eyes that convey genuine delight. ....
The setting and her joyful demeanor .... perfectly encapsulating the joyous undertaking she describes.

Figure 5: Dataset construction pipeline for the HarAvaSpeech. LLM coverts vocal descriptions with transcripts into visual descriptions. Then a text-to-image model visualizes the portrait with more details. Together, the speech audio and the synthesized images form a multimodal speech dataset with enriched momentary visual contexts.

5.1 Modality Augmentation

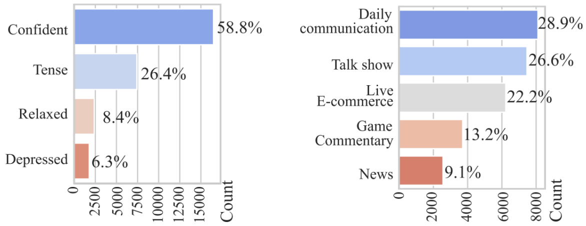
The emergence of emotional speech datasets and natural language-prompted speech datasets [25, 28, 30, 39] has garnered considerable attention recently. These datasets utilize expert systems based on audio understanding models to annotate pseudo-labels of high-quality speech and re-paraphrase the labels into natural languages. However, unlike [25, 28] and [30], we employ a generative model to create high-quality human portraits for each audio sample after retrieving the label-based vocal descriptions. This not only ensures the diversity of visual contexts, but also improves the generalizability across different speakers and contexts, making it more robust in real-world applications. Specifically, we design the following automatic pipeline for dataset collection as shown in Fig. 5.

- **Textual Modality Transferring:** Given a carefully crafted prompt template, LLMs first convert the vocal descriptions (textual descriptions for speech characteristics) to visual descriptions (image captions describing the scenarios of the speech).
- **Visualization:** Given the visual descriptions, text-to-image models synthesize the image capturing the essence of audio context and speaking style, completing the vision piece for the language-audio-vision speech dataset.

Before the first stage, we employ the tools from SpeechCraft [30] to annotate the pseudo labels and intermediate vocal descriptions. In the first stage, we employ GPT-3.5 turbo [7] as the textual modality converter, as previous works [25, 28] have proved its effectiveness in language processing. The prompt template of textual modality transferring are provided App. C (supplementary materials). As for the image generation, we employ DALL-E 3 for its state-of-the-art prompt following ability [4]. As for the data quality control, great efforts have been made during the post-processing to filter out the low-quality and harmful content.

5.2 Multimodal Dataset HarAvaSpeech

We obtain a bilingual multimodal image-speech dataset HarAvaSpeech based on an internal expressive speech dataset. HarAvaSpeech has 28,929 image-audio pairs, with 28,063 samples for training split, and 866 samples for testing split. We visualize the diversity of partial attributes of the HarAvaSpeech in Fig. 6. Originating from large-scale multi-speaker corpora, HarAvaSpeech has superiority both in audio and visual diversities compared to existing multimodal avatar datasets.



(a) Body posture. **(b) Scene.**

Figure 6: Visualization for the body posture and scene distribution of HarAvaSpeech.

6 Experiments

This section conducts experiments and shows results. The quantitative and qualitative evaluations have proved the effectiveness of the dataset and the superiority of the HarmoniVox and HarAvaSpeech.

6.1 Experimental Settings

Baseline Settings. To validate the performance of HarmoniVox, we adopt a fair but straight-forward design as the baseline method, such that the Q-former is trained with cosine similarity loss between image embedding and audio embedding (without contrastive learning and MMCL strategy).

Table 2: Quantitative and qualitative results for intra-domain and out-of-domain settings. For qualitative evaluations, sixteen (-I) and ten participants (-O) were invited to rate the speech naturalness (QMOS) and audiovisual consistency (RMOS). The suffixes '-I' and '-O' respectively represent the intra-domain settings and out-of-domain settings.

	Intra-domain Settings									Out-of-domain Settings				
	MCD	**Age**	**Gender**	**Pitch**	**Energy**	**Emotion**	**Speed**	**QMOS-I**	**CMOS-I**	**Age**	**Gender**	**Emotion**	**QMOS-O**	**CMOS-O**
Ground-truth	-	-	-	-	-	-	-	4.41±0.95	3.99±1.13	-	-	-	-	-
Baseline	10.5943	53.76	**100.00**	64.52	54.84	44.09	88.17	2.91±1.31	2.90±1.22	27.91	**90.80**	38.42	4.00±0.93	3.53±0.86
Proposed w/o descs	9.5823	**58.06**	98.92	63.44	64.52	50.54	**96.77**	**3.32±1.14**	**3.04±1.06**	**28.90**	89.98	37.60	4.21±0.92	3.53±0.92
Proposed	**9.3991**	56.99	97.85	**65.59**	**65.59**	**58.06**	**96.77**	3.23±1.10	3.02±1.10	23.81	90.64	**40.39**	**4.31±0.91**	**4.08±0.91**
Proposed w/o anneal	9.4266	**58.06**	97.85	58.06	63.44	49.46	95.70	-	-	-	-	-	-	-

Training Settings. In stage I, the VITS backbone with reference speech encoder is pretrained on an internal Mandarin speech dataset and two English speech datasets: TextrolSpeech [28] and Gigaspeech-m [8]. At the early training stage, we additionally three large-scale corpora to establish the training: AISHELL-3 [55], ZHVOICE[3], LibriTTS-R [36]. In Stage II, the model is trained on HarAvaSpeech with textual description guidance.

Evaluation Settings. To evaluate the performance and generalization ability of our model, the metrics we employed includes:

- **Mel Cepstral Distortion** (MCD)[4] aims to evaluate the overall distances between Mel frequency cepstral coefficient (MFCC) vectors of ground-truth speech and synthesized speech.
- **Classification Accuracy for speech attributes** aims to measure the capability of speech style controlling from images in detail. The classification of age and gender is based on the pretrained wav2vec 2.0 model [2] with a linear classification head on top. The classification of emotion is achieved by Emotion2Vec model [46], a universal self-supervised emotion recognition model open-sourced. The classification of pitch, energy, and speed is based on the audio processing tool *librosa*.
- **Quality Mean Opinion Scale** (QMOS) asks human raters to evaluate the naturalness of speech on a scale from one to five, ranging from *quite unnatural* to *quite natural*.
- **Consistency Mean Opinion Scale** (CMOS) asks human raters to assess the audiovisual harmony between the visual image and speech, ranging from *quite unmatching* to *quite matching*.

We conduct experiments in the following settings:

- Intra-domain settings. For the intra-domain settings, we adopt the image and ground-truth audio from HarAvaSpeech. The quantitative experiments in this setting utilize metrics including MCD, and feature accuracy, while the qualitative metric QMOS and CMOS with the ten samples randomly picked out.
- Out-of-domain settings. For the quantitative experiments, we collect 550 real-human images from the web to test the generalization capability. Due to the lack of ground-truth audio, only the accuracy of age, gender, and emotion is computed based on the image labels. For the qualitative experiments, we collect 20 real-human images from the web for inference.

6.2 Multi-modal Contrastive Learning Success

Analysis of Result. The quantitative and qualitative results of both intra-domain and out-of-domain settings are illustrated in Tab. 2. These results demonstrate that the proposed method surpasses the baseline method in both the naturalness and visual-vocal consistency. Compared to the baseline method, our model achieves lower MFCC distances between the synthesized audio and ground-truth audio, showing a stronger reconstruction capability. Further, given a static portrait, the proposed model shows a finer speech style control, leading to higher classification accuracy for age, pitch, energy, emotion, and speed. ***Most importantly, in terms of out-of-domain settings, our method demonstrates better generalizability over the baseline method, with higher values in emotion accuracy and subjective metrics.*** We credit this to the MMCL strategy, aligning the speech-related visual representation to the speech style space. Though the baseline method could establish a connection between the visual embedding and the audio embedding, the generative model with such a large amount of trainable parameters (110M for BERT-base) would meet the problem of *mode collapse* [20]. On the contrary, the uniformity of latent representation brought by the contrastive learning prevents the generative model from the mode collapse problem[61]. As a result, the proposed model achieves better performance with unseen cases.

Ablation Study. The proposed model demonstrates more intricate emotional controllability and generalizability, validating the effectiveness of MMCL strategy. However, its underlying mechanism remains unclear, prompting us to conduct the following ablation studies specifically targeting the MMCL strategy: (1) *Proposed w/o descs*: The model is trained without the supervision of textual descriptions, i.e., set $\lambda_{it} = \lambda_{at} = 0$. (2) *Proposed w/o anneal*: The model is trained without annealing function, i.e. set fixed $\lambda_{ia} = 1$.

As shown in Tab. 2, despite outperforming the baseline method, ***the absence of the descriptions in cross-modal alignment leads to a performance downgrade***, especially in the emotion, pitch, energy accuracy of intra-domain settings and the speech naturalness and ***audio-visual consistency of out-of-domain settings***. This indicates the effectiveness of textual descriptions in alleviating the complicated audiovisual relations as expected. When learned with textual supervision, the image embedding from the Q-former first moves closer to the speech style subspaces, which guarantees the stability of the gradient direction. Meanwhile, the model training without the annealing function shows a decrease in accuracy for pitch, energy, emotion, and speed. ***This verifies the annealing function for efficiently integrating the multiple learning objectives at each step.***

Overall, these results indicate that our proposed method HarmoniVox, especially the MMCL strategy, exhibits excellent performance in stylizing speech harmonious with the visual presentation.

[3] https://github.com/fighting41love/zhvoice.
[4] https://sypi.org/project/pymcd/.

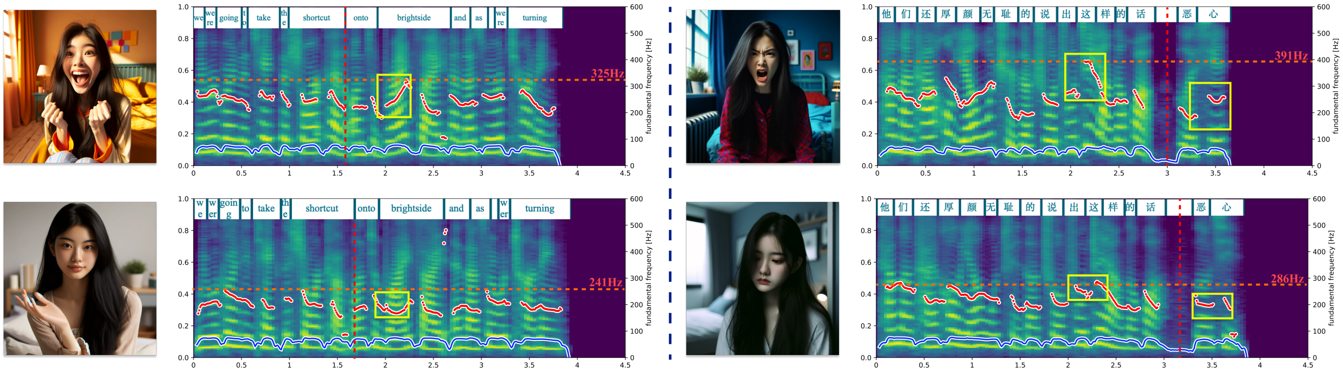


Figure 7: Visualization of acoustic features of the synthesized speech along with the corresponding conditional images. For every figure, the background demonstrates the *mel-spectrogram*, the brighter color, the larger amplitude. The fragmented red curve outlined in white shows the *pitch contour*. The continuous blue curve outlined in white shows the *energy intensity*. The horizontal orange dash line and the number above the line display the *peak fundamental frequency* (F0). The verticalwred dash line displays the exact timestamp of one specific break in the transcripts. The light yellow bounding box indicates a worth noting shift of F0 in the corresponding interval. The time-aligned transcripts are floating on the top of the figure.

6.3 Effectiveness of Synthesized Data

Collecting real-world data for training is challenging for our task. However, to verify the effectiveness and superiority of HarAvaSpeech, we have conducted a comparison experiment on the MEAD-TTS[1] data with our proposed model. To make the comparison more fair, the evaluation is conducted on the out-of-domain test set from the web, which is completely orthogonal (non-overlapping) with both MEAD-TTS [23] and HarAvaSpeech. The quantitative results presented in the Tab. 3 indicate that the model trained on HarAvaSpeech exhibits significantly ***better performance than MEAD-TTS in addressing explicit consistency, such as predicting harmonious timbre and emotion for avatars***, verifying the robustness of our dataset.

Table 3: Quantitative results for Dataset Comparison with the proposed method on out-of-domain testset.

	Age	Gender	Emotion
MEAD-TTS	**24.63**	82.10	35.14
HarAvaSpeech (Ours)	23.81	**90.64**	**40.39**

6.4 Case Study

To illustrate the performance of our model, we select the synthesized speech conditioned by four female portraits and visualize the acoustic features (mel-spectrogram, F0, intensity) in Fig. 7. All the portraits describe a potentially-talking scene where a young female with sleek black hair is sitting in the room yet the *hand gestures*, *body posture*, and *facial expressions* of the subject differ. For comparison, as shown in Fig. 7, we use the images of happy laughing (the left top) and neutral states (the left bottom) to condition the generation of the same English sentence, and the images of angry shouting (the right top) and vaguely distracted (the right bottom) to condition the generation of the same Mandarin sentence.

For the English sentence (the left samples), the top speech has a higher pitch and faster speed compared to the bottom speech generally, as the word break between 'shortcut' and 'onto' appears earlier and the peak F0 is much higher as shown in the top figure. Most importantly, the word 'brightside' is heavily emphasized in the top speech, demonstrating a sharp increase in the pitch as shown in the yellow bounding box. ***This is consistent with the widely open mouth and the joyful expression in the conditional image.***

For the Mandarin sentence (the right samples), the top speech also has a higher pitch and faster speed in general, with an early word break for the comma and a higher peak F0. In addition to the sharp pitch jump in the first bounding box of the right top figure, ***there's a noticing pitch uplift at the tailing of the sentence, conveying the furious emotion and contemptuous tone.***

As shown in Fig. 7, our model exhibits a promising capability in audiovisual harmony for virtual avatar. To be noted, in the two top cases, the model automatically emphasizes the words 'brightside' and '这样' (translation:'this'), indicating that the model has learned human habits of prosodic emphasis to intensify the emotion from the large scale speech corpora.

7 Conclusion

In this paper, we focus on the crafting harmonious speech in consistent with visual presentations for virtual avatar animation. First, we systematically define the key audiovisual harmony framework via empirical studies. Then, we proposed a multi-modal consistency modeling method HarmoniVox, which outperforms the baseline method by demonstrating a more profound visual control capability in speech style control. To support the method, we employed generative AI to establish a multi-modal dataset HarAvaSpeech via modality augmentation, tackling the data scarcity issue. Leveraging HarAvaSpeech dataset, extensive experiments demonstrate that the models trained on the synthesized dataset show generalization ability for out-of-domain photos including emojis.

We also acknowledge the limitations of our research, such as the multi-modal bias in dataset distributions. Future work includes the balancing multi-modal distribution and integration of our approach with avatar motion generation. Addressing these aspects constitutes our future research directions, with the ultimate goal of achieving a seamless, unified and harmonious user experience.

Acknowledgments

This work is supported by the National Key R&D Program of China under Grant No. 2024QY1400, and the National Natural Science Foundation of China No. 62425604. This work is also supported by Tsinghua University Initiative Scientific Research Program, Shenzhen Science and Technology Program (JCYJ20220818101014030) and the Institute for Guo Qiang at Tsinghua University.

References

[1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. 2021. Partial FC: Training 10 Million Identities on a Single Machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 1445–1449.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12449–12460.

[3] Maciej Behnke, Nadia Bianchi-Berthouze, and Lukasz D. Kaczmarek. 2021. Head movement differs for positive and negative emotions in video recordings of sitting individuals. *Scientific Reports* 11, 1 (April 2021), 7405.

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2 (2023), 3.

[5] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*. ACM Press/Addison-Wesley Publishing Co., USA, 187–194. doi:10.1145/311535.311556

[6] Margaret M Bradley and Peter J Lang. 2000. Emotion and motivation. *Handbook of psychophysiology* 2 (2000), 602–642.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901.

[8] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909* (2021).

[9] Minsoo Choi, Alexandros Koilias, Matias Volonte, Dominic Kao, and Christos Mousas. 2023. Exploring the Appearance and Voice Mismatch of Virtual Characters. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 555–560. doi:10.1109/ISMAR-Adjunct60411.2023.00118

[10] J. S. Chung, A. Nagrani, and A. Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8440–8451.

[12] Mark Coulson. 2004. Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence. *Journal of Nonverbal Behavior* 28, 2 (June 2004), 117–139.

[13] Jiahao Cui, Hui Li, Yun Zhang, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. 2024. Hallo3: Highly Dynamic and Realistic Portrait Image Animation with Video Diffusion Transformer. _eprint: 2412.00733.

[14] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-Training with Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3504–3514. arXiv:1906.08101 [cs].

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (May 2019). 101 citations (INSPIRE 2024/4/13) 101 citations w/o self (INSPIRE 2024/4/13) arXiv:1810.04805 [cs.CL].

[16] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2020. 3D Morphable Face Models—Past, Present, and Future. *ACM Trans. Graph.* 39, 5 (June 2020). doi:10.1145/3395208 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[17] Cathy Ennis, Ludovic Hoyet, Arjan Egges, and Rachel McDonnell. 2013. Emotion Capture: Emotionally Expressive Characters for Games. In *Proceedings of Motion on Games (MIG '13)*. Association for Computing Machinery, New York, NY, USA, 53–60. doi:10.1145/2522628.2522633 event-place: Dublin 2, Ireland.

[18] Sarah Evans, Nick Neave, and Delia Wakelin. 2006. Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology* 72, 2 (2006), 160–163.

[19] Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. 2021. Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (IVA '21)*. Association for Computing Machinery, New York, NY, USA, 76–83. doi:10.1145/3472306.3478338 event-place: Virtual Event, Japan.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (October 2020), 139–144. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[21] 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*. 1321–1325.

[22] Shreyank N. Gowda, Dheeraj Pandey, and Shashank Narayana Gowda. 2023. From Pixels to Portraits: A Comprehensive Survey of Talking Head Generation Techniques and Applications. doi:10.48550/arXiv.2308.16041 arXiv:2308.16041 [cs].

[23] Wenhao Guan, Yishuang Li, Tao Li, Hukai Huang, Feng Wang, Jiayan Lin, Lingyan Huang, Lin Li, and Qingyang Hong. 2024. MM-TTS: Multi-Modal Prompt Based Style Transfer for Expressive Text-to-Speech Synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (March 2024), 18117–18125.

[24] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. 2024. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. arXiv:2407.03168 [cs].

[25] Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable Text-To-Speech With Text Descriptions. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Rhodes Island, Greece, 1–5.

[26] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654 [cs].

[27] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu. 2018. Face-voice matching using cross-modal embeddings. In *Proceedings of the 26th ACM international conference on Multimedia*. 1011–1019.

[28] Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024. TextrolSpeech: A Text Style Control Speech Corpus with Codec Language Text-to-Speech Models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 10301–10305.

[29] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. 2024. Loopy: Taming Audio-Driven Portrait Avatar with Long-Term Motion Dependency. arXiv:2409.02634 [cs].

[30] Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. 2024. SpeechCraft: A Fine-Grained Expressive Speech Dataset with Natural Language Description. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. Association for Computing Machinery, New York, NY, USA, 1255–1264. event-place: Melbourne VIC, Australia.

[31] 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 (2003), 1709–1714. Publisher: Elsevier.

[32] Wassily Kandinsky, Franz Marc, and Klaus Lankheit. 1974. *The Blaue Reiter almanac* (new documentary ed. / edited and with an introduction by klaus lankheit ed.). Thames and Hudson, New York, NY, USA. https://ci.nii.ac.jp/ncid/BA21725515

[33] Spencer D. Kelly and Quang-Anh Ngo Tran. 2023. Exploring the Emotional Functions of Co-Speech Hand Gesture in Language and Communication. *Topics in cognitive science* (2023).

[34] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*. PMLR, 5530–5540.

[35] Shigenobu Kobayashi. 1998. *Colorist : a practical handbook for personal and professional Use.* Kodansha International.

[36] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802* (2023).

[37] Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin Kim, Dohee Kong, and Sangjin Kim. 2023. VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design. arXiv:2307.16430

[cs, eess].

[38] Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. 2023. Imaginary Voice: Face-styled Diffusion Model for Text-to-Speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

[39] Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiang-Yang Li, Sheng Zhao, Tao Qin, and Jiang Bian. 2023. PromptTTS 2: Describing and Generating Voices with Text Prompt. arXiv:2309.02285 [cs, eess].

[40] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19730–19742.

[41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12888–12900.

[42] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (Nov. 2017). doi:10.1145/3130800.3130813 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (Oct. 2015). doi:10.1145/2816795.2818013 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[44] Hsiao-Han Lu, Shao-En Weng, Ya-Fan Yen, Hong-Han Shuai, and Wen-Huang Cheng. 2021. Face-based Voice Conversion: Learning the Voice behind a Face. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, Virtual Event China, 496–505.

[45] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. 2023. DreamTalk: When Expressive Talking Head Generation Meets Diffusion Probabilistic Models. arXiv:2312.09767 [cs].

[46] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. *arXiv preprint arXiv:2312.15185* (2023).

[47] Lawrence E. Marks. 1978. The Unity of the Senses: Interrelations Among the Modalities. https://api.semanticscholar.org/CorpusID:27335285

[48] Lauren W Mavica and Elan Barenholtz. 2013. Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance* 39, 2 (2013), 307. Publisher: American Psychological Association.

[49] Jan Ondřej, Cathy Ennis, Niamh A. Merriman, and Carol O'sullivan. 2016. FrankenFolk: Distinctiveness and Attractiveness of Voice and Motion. *ACM Trans. Appl. Percept.* 13, 4 (July 2016). doi:10.1145/2948066 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[50] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748 [cs, stat].

[51] Suzanne Oosterwijk, Mark Rotteveel, Agneta H. Fischer, and Ursula Hess. 2009. Embodied emotion concepts: How generating words about pride and disappointment influences posture. *European Journal of Social Psychology* 39, 3 (2009), 457–466. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.584.

[52] Wim Pouw, Alexandra Paxton, Steven J. Harrison, and James A. Dixon. 2020. Acoustic information about upper limb movement in voicing. *Proceedings of the National Academy of Sciences* 117, 21 (2020), 11364–11367. _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2004163117.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763.

[54] HENDRIK N. J. SCHIFFERSTEIN and CHARLES SPENCE. 2008. 5 - MULTISENSORY PRODUCT EXPERIENCE. In *Product Experience*, Hendrik N. J. Schifferstein and Paul Hekkert (Eds.). Elsevier, San Diego, 133–161. doi:10.1016/B978-008045089-6.50008-3

[55] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567* (2020).

[56] Charles Spence and Nicola Di Stefano. 2022. Crossmodal Harmony: Looking for the Meaning of Harmony Beyond Hearing. *i-Perception* 13 (2022). https://api.semanticscholar.org/CorpusID:246766300

[57] Xusen Sun, Longhao Zhang, Hao Zhu, Peng Zhang, Bang Zhang, Xinya Ji, Kangneng Zhou, Daiheng Gao, Liefeng Bo, and Xun Cao. 2023. VividTalk: One-Shot Audio-Driven Talking Head Generation Based on 3D Hybrid Prior. arXiv:2312.01841 [cs].

[58] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. 2024. EMO: Emote Portrait Alive-Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions. *arXiv preprint arXiv:2402.17485* (2024).

[59] Haoyu Wang, Haozhe Wu, Junliang Xing, and Jia Jia. 2023. Versatile Face Animator: Driving Arbitrary 3D Facial Avatar in RGBD Space. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 7776–7784. doi:10.1145/3581783.3612065 event-place: Ottawa ON, Canada.

[60] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation. In *Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*. Springer-Verlag, Berlin, Heidelberg, 700–717. event-place: Glasgow, United Kingdom.

[61] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*. PMLR, 9929–9939.

[62] Zixuan Wang, Jia Jia, Haozhe Wu, Junliang Xing, Jinghe Cai, Fanbo Meng, Guowen Chen, and Yanfeng Wang. 2022. GroupDancer: Music to Multi-People Dance Synthesis with Style Collaboration. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. Association for Computing Machinery, New York, NY, USA, 1138–1146. doi:10.1145/3503161.3548090 event-place: Lisboa, Portugal.

[63] Haozhe Wu, Jia Jia, Junliang Xing, Hongwei Xu, Xiangyuan Wang, and Jelo Wang. 2023. MMFace4D: A Large-Scale Multi-Modal 4D Face Dataset for Audio-Driven 3D Face Animation. arXiv:2303.09797 [cs].

[64] Haozhe Wu, Songtao Zhou, Jia Jia, Junliang Xing, Qi Wen, and Xiang Wen. 2023. Speech-Driven 3D Face Animation with Composite and Regional Facial Movements. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 6822–6830. event-place: Ottawa ON, Canada.

[65] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. 2021. Modeling clothing as a separate layer for an animatable human avatar. *ACM Trans. Graph.* 40, 6 (Dec. 2021). doi:10.1145/3478513.3480545 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[66] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. SECap: Speech Emotion Captioning With Large Language Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19323–31. Issue: 17.

[67] Zhihan Yang, Zhiyong Wu, Ying Shan, and Jia Jia. 2023. What Does Your Face Sound Like? 3D Face Shape towards Voice. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 11 (June 2023), 13905–13913. Number: 11.

[68] Zijie Ye, Jia Jia, and Junliang Xing. 2023. Semantics2Hands: Transferring Hand Motion Semantics between Avatars. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 9282–9290. doi:10.1145/3581783.3612703 event-place: Ottawa ON, Canada.

[69] Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. 2020. ChoreoNet: Towards Music to Dance Synthesis with Choreographic Action Unit. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 744–752. doi:10.1145/3394171.3414005 event-place: Seattle, WA, USA.

[70] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8652–8661.

[71] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3661–3670.