I^{2}TTS: Image-indicated Immersive Text-to-speech Synthesis with Spatial Perception

Jiawei Zhang*[1], Tian-Hao Zhang*, Jun Wang†, Jiaran Gao*, Ruijie Tao‡, Xinyuan Qian*[2], and Xu-Cheng Yin*
* University of Science and Technology Beijing, Beijing, China
† Tencent AI Lab, Shenzhen, China
‡ National University of Singapore, Singapore
E-mail: javezone2@gmail.com, qianxy@ustb.edu.cn

***Abstract*—Controlling the spatial and stylistic characteristics of synthesized speech is essential for immersive and personalized applications such as virtual reality, gaming, and human-computer interaction. While recent Text-to-speech (TTS) systems have explored multi-modal conditioning, they often suffer from poor reverberation fidelity or degraded audio quality due to reliance on external vocoders. In this paper, we propose Image-indicated Immersive Text-to-speech Synthesis (I^{2}TTS), an end-to-end multi-modal TTS framework that synthesizes high-quality, immersive speech from text and visual scene prompts. Our model leverages a CLIP-based image encoder with an adaptive adapter to extract scene-aware features, a Speech Reverberation Classifier (SRC) for refining acoustic-visual alignment during training, and a speaker encoder to enable zero-shot speaker generalization. Built upon a VITS backbone, I^{2}TTS generates reverberant speech directly without requiring a separate vocoder. Experimental results demonstrate that our approach produces spatially accurate and natural-sounding speech, achieving superior performance in both subjective and objective evaluations. Project demo page: https://spatialTTS.github.io/**

I. Introduction

Speech synthesis has seen remarkable advancements in recent years, driven by breakthroughs in deep learning and neural network architectures [1]–[4]. These innovations have enabled the generation of high-quality, natural-sounding speech across various applications, such as virtual assistants, accessibility technologies, and interactive media. Recently, numerous studies [5]–[8] have explored the use of prompts to control specific speech characteristics, such as emotion, style, or speaker identity, enabling personalized and dynamic speech synthesis.

Reverberation is also a critical auditory feature that defines how sound interacts with its environment, profoundly influences the perception and intelligibility of speech [9]. While reverberation plays a vital role in creating immersive auditory experiences, most existing speech synthesis models are designed for neutral or non-reverberant environments, limiting their applicability in spatially aware or immersive contexts. To address this, recent efforts have begun integrating environmental context into the synthesis process. For instance, Environment-Aware TTS [10] introduces an environment embedding extractor that learns environmental features from reference speech, improving the synthesis of speech tailored to specific acoustic settings by reducing intra-environment embedding distance and increasing inter-environment separation. Similarly, VoiceLDM [11] focus on mapping textual or audio descriptions into environmental feature vectors, effectively controlling the reverberation aspects of the generated audio. Furthermore, ViT-TTS [12] presents a groundbreaking multi-modal TTS task that generates speech with reverberation characteristics closely aligned with the acoustic properties of specific visual scenes. This innovation marks a significant advancement by integrating visual context into TTS. In addition, MS2KU-VTTS [13] explores immersive TTS by leveraging multi-source spatial knowledge, including image, depth, and semantic segmentation, to enhance spatial understanding. However, they depend on an external vocoder for waveform generation, which introduces synthesis artifacts, such as distortions and noise, and struggles to preserve fine-grained acoustic details like reverberation. Given that reverberation is inherently linked to environmental context, achieving robust and natural reverberation modeling necessitates a more integrated and end-to-end approach within the TTS pipeline.

To address this challenge, we have integrated scene-aware reverberation directly into the end-to-end speech synthesis process. To capture the acoustic characteristics of a given environment, we first use an image encoder to extract visual features from the scene's image. Specifically, we employ the Contrastive Language-Image Pretraining (CLIP) [14] model for our image encoding process, as it effectively aligns visual and textual information, enabling the model to extract meaningful features representing the scene's environment. In addition to scene understanding, personalization is also crucial for immersive experience. We introduce a speaker encoder module that enables zero-shot speaker adaptation, allowing the model to synthesize speech with the voice characteristics of an unseen speaker based on a short reference speech. This module extracts speaker embeddings from reference speech and conditions the speech generation process accordingly, ensuring that the synthesized output reflects the target speaker identity. The entire pipeline is built on Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) [3], with modifications to incorporate the features that are relevant to the scene. Once immersive speech is generated, we further refine it using a speech reverberation classifier (SRC) model,

[1]Work done during internship at Tencent AI Lab.
[2]Corresponding author

979-8-3315-7206-8/25/$31.00

which adjusts the reverberation and refines the reverberation characteristics to ensure accurate alignment with the prompt scene.

Our contributions are summarized as follows:

1) We propose the first end-to-end multi-modal TTS framework that synthesizes immersive and high-quality reverberant speech directly from text and visual scene prompts, built upon a VITS-based backbone without requiring a separately trained vocoder.
2) We introduce a novel architecture that includes a CLIP-based image encoder to extract scene-aware visual representations and an SRC (Scene Relevance Classifier) module to refine the acoustic consistency between the synthesized speech and the visual environment, thereby improving scene precision.
3) Our system supports zero-shot speaker adaptation through a speaker encoder, enabling personalized and spatially grounded speech generation that realizes specific speakers in specific scenes.
4) Extensive objective and subjective evaluations demonstrate that our method achieves state-of-the-art performance in generating spatially aligned, high-fidelity speech across diverse visual environments.

II. RELATED WORK

TTS has evolved significantly over the past few decades, moving from early rule-based systems [15]–[17] to modern deep learning-based approaches [1]. This section reviews key developments in TTS, with a focus on areas relevant to our proposed task: reverberation modeling, context-aware speech synthesis, and the integration of visual cues in speech generation.

A. Standard Speech Synthesis

Early speech synthesis systems, such as formant synthesis and concatenative synthesis, focused on generating speech by stringing together pre-recorded phonemes or sound units. Although these methods were able to produce intelligible speech, they often lacked naturalness and flexibility, particularly in varying acoustic environments. The advent of deep learning brought about a paradigm shift, with models like Tacotron [1] that enables end-to-end TTS to produce highly natural and expressive speech. Since then, TTS has entered an explosive development with many models being explored, such as FastSpeech2 [18], Diff-TTS [19] and ProDiff [20]. Recently, discrete token-based audio synthesis approaches, such as CLaM-TTS [21], VALL-E [4], AudioLM [22] and NaturalSpeech3 [23] formulate the generation process as a conditional language modeling task. Using extensive training data and the power of large language models, they significantly enhance speech quality and naturalness. However, these models typically generate speech in a neutral and non-reverberant environment, without considering the acoustic characteristics of the scene.

B. Spatial Perception in Speech Synthesis

Reverberation modeling has traditionally been addressed in the field of audio signal processing, where techniques such as convolution with room room impulse responses (RIR) are used to simulate the effects of different acoustic environments on speech. In the context of speech synthesis, reverberation is often applied as a post-processing step, where the synthesized speech is convolved with an RIR to achieve the desired effect [24]–[26]. Although effective, this approach treats reverberation as an integral part of the synthesis process rather than an afterthought. To address this limitation, recent research has started to integrate environmental cues directly into the synthesis pipeline. For instance, ViT-TTS [12] introduces a multimodal speech synthesis paradigm where visual scenes are used to condition the generation of speech with appropriate reverberation. By leveraging visual prompts, it attempts to model the spatial context inherent in the image. Extending this direction, MS2KU-VTTS [13] takes a more holistic approach to immersive TTS by incorporating multi-source spatial information, including RGB images, depth maps, and semantic segmentations. However, due to their reliance on external vocoders, they may suffer from synthesis artifacts and limited control over fine-grained reverberation.

C. Summary

While speech synthesis has advanced significantly in producing natural and expressive outputs, the integration of environmental and spatial context remains a relatively nascent area. Post-processing methods for spatial effects, such as RIR convolution, have been standard but lack deep integration with the synthesis process. Recent efforts, have begun to explore multi-modal approaches that utilize scene prompts, yet challenges in speech quality and acoustic alignment persist due to their reliance on external vocoders and limited reverberation modeling fidelity.

Our approach addresses these limitations by tightly integrating scene-aware reverberation into the TTS pipeline. By treating environmental acoustics as an integral part of synthesis rather than a post-processing effect, our model achieves superior spatial alignment and speech quality, offering a more immersive and contextually accurate auditory experience.

III. PROPOSED METHOD

The overall architecture of our proposed framework is shown in Fig 1. Our proposal is expected to generate speech that not only sounds natural, but also conveys a strong sense of spatial presence and speaker identity, guided by the scene and speaker prompts.

A. Image Encoder

In order to capture the acoustic characteristics of a given scene prompt, we first utilize the CLIP model and an adapter to extract visual features from the scene image that represent the scene's likely acoustic properties. The adapter module in our model is designed to efficiently integrate visual features

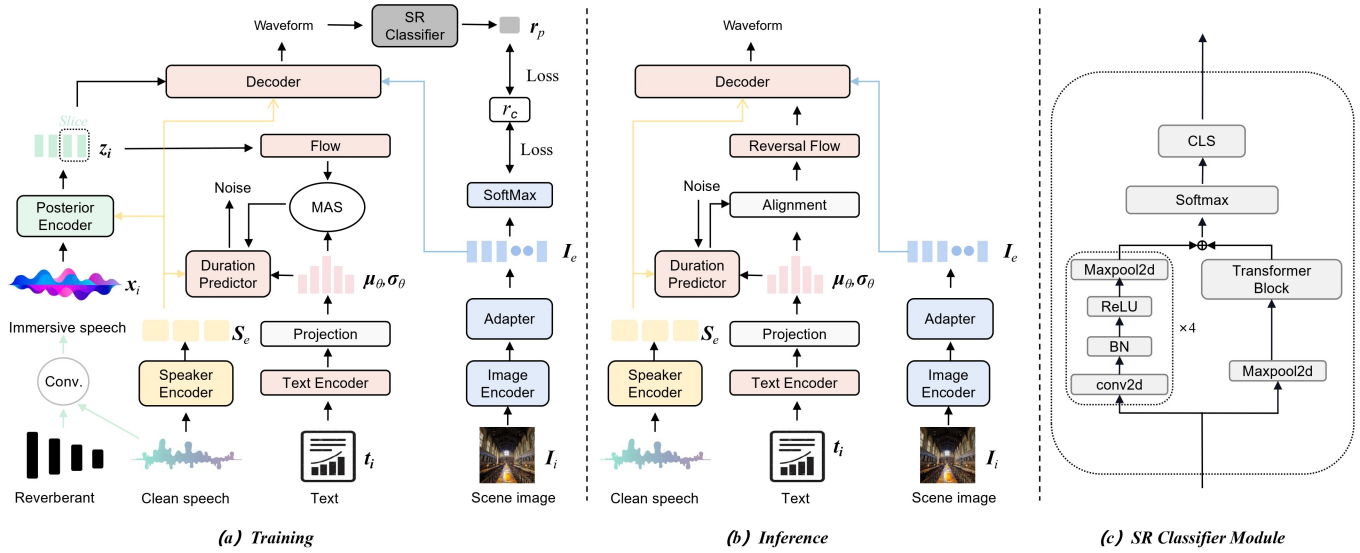

Fig. 1. Block diagram of the I^{2}TTS model ($\oplus$ indicates element-wise addition).

of the scene into the speech synthesis pipeline. This module is simply comprised of a MLP network.

In practice, we first input the scene image I_i into the CLIP-based image encoder to extract the scene prompt features. After being fed to the adapter, we transform the scene prompt features into an acoustic space embedding I_e. To make this embedding close to the reverb category r_c, we guide it with a SoftMax layer and compute a Cross-Entropy (CE) loss function with r_c. Once obtaining the output embedding from the adapter module, we directly feed it into the TTS backbone. This integration allows the scene-aware information, encapsulated in the embedding, to influence various stages of the TTS pipeline.

B. Speaker Encoder

To enable zero-shot speaker adaptation and ensure that the synthesized speech maintains consistent speaker identity, we introduce a dedicated speaker encoder module. This module extracts speaker embeddings from a short reference speech, allowing our model to imitate unseen speaker voices without retraining.

Concretely, inspired by YourTTS [5], we adopt the StyleEncoder architecture from [27] to encode speaker-specific characteristics into a compact embedding S_e. This embedding is then projected via a linear layer and injected into the TTS backbone through adaptive layer modulation, allowing the system to synthesize speech in the voice of an unseen speaker in a zero-shot manner.

C. Speech Reverberation Classification

To ensure that the reverberation characteristics of the synthesized speech are accurately matched to the visual scene, once the initial mel-spectrogram is generated, we further refine it using an SRC model that classifies and adjusts the reverberation to ensure that it matches the scene.

For the architecture illustrated in Fig. 1 (b), the SRC reveives the mel-spectrogram from the synthesis speech and predicts the class of reverberation. Specifically, the main network comprises four 2-D convolution blocks to capture local time-frequency features and a transformer [28] block to obtain global dependencies across time frames. Then the features of the two blocks are concatenated to form a comprehensive feature representation fed into a softmax layer for further transformation, followed by a classification layer that outputs the final predictions. The predictions of the reverb class r_p are finally computed by a CE loss.

D. TTS Backbone

We adopt VITS [3] as the backbone of our system for its end-to-end design and high-quality output. It consists of a Posterior Encoder, Text Encoder, and Transformer-based flow for expressive latent modeling, with MAS and Duration Predictor for alignment and phoneme timing, and a Decoder to reconstruct the final waveform.

During training, the model receives text, immersive speech, speaker embedding, and scene embedding as input. The scene embedding, extracted from the CLIP-based image encoder and adapted via an MLP, guides the model to learn scene-aware reverberation characteristics. Moreover, speaker embedding S_e, obtained via a speaker encoder, allows the system to generalize to unseen speakers in a zero-shot fashion. Both embeddings are injected into the model through conditional modules, enabling joint learning of content, speaker identity, and spatial acoustics. Finally, the SRC module verifies and refines the reverberation characteristics in the mel-spectrogram to better align with the input scene.

At inference time, the system takes input text, an image prompt, and an optional speaker reference. The adapted CLIP feature and speaker embedding modulate the synthesis process, allowing the decoder to generate immersive speech with

appropriate reverberation and speaker identity.

IV. EXPERIMENT

A. Implementation Details

For our experiments, we used three primary datasets: LJSpeech [29] , VCTK [30] and Image2Reverb [24]. These datasets provide the necessary diversity in speech, reverberated sound, and scene images to train and evaluate our proposed method. Specifically, we randomly convolve the clean speech data from the LJSpeech dataset and VCTK dataset with impulse responses from the Image2Reverb dataset. This process generates convolutional speech samples with reverberation effects that correspond to the visual scenes depicted in the images, which finally serves as the ground truth for training, providing a benchmark for evaluating the accuracy and naturalness of the synthesized output.

Additionally, we pretrained the SRC model on the convolutional speech samples over 300 epochs. Once integrated into the TTS framework, we froze its parameters and only used the model to calculate the classification loss of the generated speech. The TTS framework was trained for 200K iterations using AdamW optimizer on 3 NVIDIA GeForce RTX 3090 GPUs.

For the comparative analysis, we conducted experiments on the following systems: 1) GT: the ground-truth convolutional speech. 2) Baseline: the VITS backbone trained on the clean speech followed by convolution with RIR. 3) ViT-TTS. 4) MS2KU-VTTS. 5) Proposed method w/o SR Classifier: our proposed framework that without SRC module. 6) Proposed method w/o CLIP: Our proposed framework utilizing a convolution network replaces the CLIP encoder.

B. Evaluation Metrics

To assess the effectiveness of our proposed method, we first evaluate the quality of the generated speech and the correspondence between the scene and the reverberation using some objective metrics including Word Error Rate (WER), mel cepstral distortion (MCD) [31], Speaker Encoder Cosine Similarity (SECS) [32] and Space Recognition Error (SRE). WER measures the accuracy of the synthesized speech in terms of word recognition. MCD measures the spectral distance between the synthesized speech and the ground truth speech. SECS is used to assess the speaker similarity between the input and generated speech. SRE measures the correctness of the room acoustics of the generated speech.

We also conduct subjective evaluations to assess perceptual aspects that are difficult to capture quantitatively. We use a Mean Opinion Score (MOS) [33] with 95 $\%$ confidence intervals to assess speech quality, where listeners are asked to rate the audio on a scale from 1 to 5. We employ the Naturalness-Mean Opinion Score (NMOS) to gauge the naturalness of synthesized speech to human listeners, evaluating its overall fluidity, expressiveness, and resemblance to human speech. Simultaneously, we utilize the Similarity-Mean Opinion Score (SMOS) to measure the perceived similarity between the speaker's voice in the input and the generated speech. Furthermore, We test Immersive-Mean Opinion Score (IMOS) for matching between scene prompt and spatial characteristics of speech to evaluate the effectiveness of our model in aligning with human perceptions of the scene's acoustics. We engaged 20 listeners each evaluating clarity, reverberation accuracy, and naturalness on 6 samples of test set. Standard deviations across these scores were small ($<$0.25), indicating consistent evaluations.

C. Experimental Results

Table I presents the objective and subjective evaluation results across different models. Compared to the baseline and prior scene-aware TTS systems (ViT-TTS and MS2KU-VTTS), our proposed method consistently achieves superior performance across nearly all metrics.From the objective perspective, our full model attains the lowest WER (7.6%) and MCD (4.22), indicating enhanced intelligibility and spectral fidelity. Furthermore, it shows a notable improvement in SRE, reducing the mismatch between predicted and target scene acoustics to 27.2%, a significant gain over ViT-TTS (64.2%) and MS2KU-VTTS (57.1%). The SECS also improves to 0.62, demonstrating better preservation of speaker identity.In terms of subjective quality, our model achieves the highest scores on NMOS (3.98), SMOS (4.21), and IMOS (3.96), reflecting improved naturalness, speaker similarity, and perceived scene-spatial alignment. Notably, even ablation variants (w/o CLIP or SR Classifier) outperform existing methods, highlighting the robustness of our architecture.These results confirm the effectiveness of integrating both visual scene prompts and speaker identity encoding into an end-to-end TTS framework, enabling the generation of immersive, high-quality, and scene-consistent reverberant speech.

D. Environments analysis

To further evaluate our model's robustness across diverse acoustic conditions, we analyze performance separately in wide and narrow environments. As shown in Table II, our proposed method significantly outperforms ViT-TTS and MS2KU-VTTS in both scenarios. In wide environments, which typically exhibit longer reverberation tails and more diffuse reflections, our model achieves the lowest Scene Reverb Error (SRE) at 25.3% and the highest IMOS score of 4.01. This suggests that the model is more capable of capturing complex spatial cues from wide scenes, leading to more immersive and perceptually accurate reverberation. In narrow environments, which feature shorter reverberation times and tighter spatial constraints, our method again yields superior results with an SRE of 28.6% and an IMOS of 3.90. Compared to the SREs of 67.1% and 62.8% from ViT-TTS and MS2KU-VTTS respectively, our approach demonstrates better adaptability and scene-awareness. These findings confirm that the integration of visual scene prompts and the refinement by the SRC module enable our model to generalize well across spatial contexts. The consistently higher IMOS scores also reflect that human listeners perceive the reverberation in our outputs to be more faithful to the corresponding visual scenes, regardless of environmental scale.

TABLE I
EXPERIMENTAL RESULTS FOR DIFFERENT MODELS.

	Objective metrics				Subjective metrics		
Model	WER(%) ↓	MCD ↓	SECS↑	SRE(%) ↓	NMOS ↑	SMOS↑	IMOS ↑
GT	4.7	/	0.76	21.4	4.54	4.46	4.63
Baseline	7.8	4.25	/	31.6	3.83	/	3.92
ViT-TTS	8.6	4.58	/	64.2	3.77	/	3.71
MS2KU-VTTS	8.2	4.49	/	57.1	3.79	/	3.86
Proposed method w/o SR Classifier	8.0	4.29	0.53	39.7	3.82	4.16	3.81
Proposed method w/o CLIP	7.8	4.27	0.57	30.8	3.85	4.19	3.83
Proposed	7.6	4.22	0.62	27.2	3.98	4.21	3.96

TABLE II
EXPERIMENTAL RESULTS FOR DIFFERENT ENVIRONMENT.

	Wide environment		Narrow environment	
Model	SRE(%) ↓	IMOS ↑	SRE(%) ↓	IMOS ↑
ViT-TTS	58.2	3.73	67.1	3.68
MS2KU-VTTS	54.7	3.89	62.8	3.81
Proposed	25.3	4.01	28.6	3.90

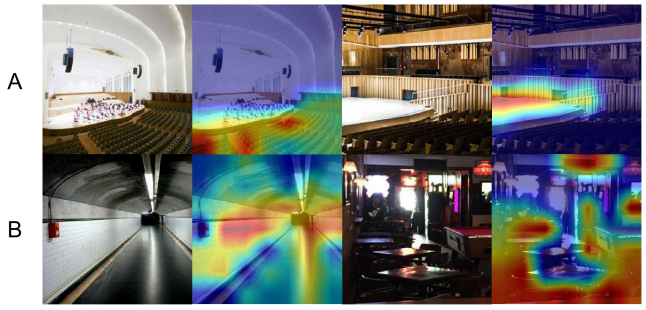

Fig. 2. Gradient-weighted Class Activation Mapping (Grad-CAM)s for images passed the scene prompt encoder, showing movement towards more textured areas for (A) a wide, and (B) a narrow environment.

E. Visualization

To gain insight into which visual features are most influential in our encoder, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [34], a widely used technique applied to visually interpret the image regions that contribute the most to scene understanding. Grad-CAM allows us to visualize the areas of the input image that contribute most significantly to the scene-aware feature extraction process. We produce such maps for our test images with the scene prompt encoder as shown in Fig. 2. The original images are shown on the left, with the corresponding Grad-CAM heatmaps on the right. These heat maps highlight areas the model considers crucial for extracting image features relevant to the acoustic environment. The color scale, which ranges from blue to red, indicates increasing levels of importance. For example, in the wide environment (A), the highlighted regions emphasize broad structural elements such as walls and seating arrangements. In contrast, in the narrow environment (B), attention is focused on confined linear features such as hallways. Experimental results demonstrate that the detected visual objects aid in TTS systems with reverberation.

V. CONCLUSION

In this work, we presented a novel approach to end-to-end multi-modal TTS system that incorporates visual scene prompts to guide the generation of immersive, contextually appropriate speech. Our method extends the capabilities of traditional TTS by introducing a framework that integrates spatial and scene-aware acoustic features into the synthesis process, addressing limitations in existing models. We demonstrated that our model achieves high-quality scene and reverb matching without deteriorating the naturalness of the speech. The combination of objective metrics and subjective evaluations shows that our model successfully balances speech quality with scene awareness, paving the way for more immersive and context-sensitive applications in virtual environments, interactive media, and other fields requiring spatially-aware audio.

VI. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 62306029, the Beijing Natural Science Foundation under Grants L233032, Shenzhen Research Institute of Big Data under Grant No. K00120240007 and CCF-Tencent Rhino-Bird Fund.

REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

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

[4] C. Wang, S. Chen, Y. Wu, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[5] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Proc. of Int. Conf. on Machine Learning*, PMLR, 2022, pp. 2709–2720.

[6] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "Prompttts: Controllable text-to-speech with text descriptions," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, IEEE, 2023, pp. 1–5.

[7] D. Yang, S. Liu, R. Huang, C. Weng, and H. Meng, "Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, 2024.

[8] W. Guan, Y. Li, T. Li, *et al.*, "Mm-tts: Multi-modal prompt based style transfer for expressive text-to-speech synthesis," in *AAAI Conf. on Artificial Intelligence*, vol. 38, 2024, pp. 18 117–18 125.

[9] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, 2002.

[10] D. Tan, G. Zhang, and T. Lee, "Environment aware text-to-speech synthesis," *arXiv preprint arXiv:2110.03887*, 2021.

[11] Y. Lee, I. Yeon, J. Nam, and J. S. Chung, "Voiceldm: Text-to-speech with environmental context," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 566–12 571.

[12] H. Liu, R. Huang, X. Lin, *et al.*, "Vit-tts: Visual text-to-speech with scalable diffusion transformer," *arXiv preprint arXiv:2305.12708*, 2023.

[13] S. He and R. Liu, "Multi-source spatial knowledge understanding for immersive visual text-to-speech," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[14] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. of Int. Conf. on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[15] J. Allen, M. S. Hunnicutt, D. H. Klatt, R. C. Armstrong, and D. B. Pisoni, *From text to speech: The MITalk system*. Cambridge University Press, 1987.

[16] A. Black, P. Taylor, R. Caley, and R. Clark, *The festival speech synthesis system*, 1998.

[17] K. Tokuda, H. Zen, and A. W. Black, "An hmm-based speech synthesis system applied to english," in *IEEE speech synthesis workshop*, Citeseer, 2002, pp. 227–230.

[18] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[19] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," *arXiv preprint arXiv:2104.01409*, 2021.

[20] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "Prodiff: Progressive fast diffusion model for high-quality text-to-speech," in *Proc. of ACM Int. Conf. on Multimedia*, 2022, pp. 2595–2605.

[21] J. Kim, K. Lee, S. Chung, and J. Cho, "Clam-tts: Improving neural codec language model for zero-shot text-to-speech," in *Proc. ofInt. Conf. on Learning Representations*, 2023.

[22] Z. Borsos, R. Marinier, D. Vincent, *et al.*, "Audiolm: A language modeling approach to audio generation," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, 2023.

[23] Z. Ju, Y. Wang, K. Shen, *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *arXiv preprint arXiv:2403.03100*, 2024.

[24] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2reverb: Cross-modal reverb impulse response synthesis," in *Proc. of Int. Conf. on Computer Vision*, 2021, pp. 286–295.

[25] C. Chen, R. Gao, P. Calamia, and K. Grauman, "Visual acoustic matching," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 18 858–18 868.

[26] A. Somayazulu, C. Chen, and K. Grauman, "Self-supervised visual acoustic matching," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[27] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *International Conference on Machine Learning*, PMLR, 2021, pp. 7748–7759.

[28] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[29] K. Ito and L. Johnson, "The lj speech dataset," 2017.

[30] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, vol. 6, p. 15, 2017.

[31] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, IEEE, vol. 1, 1993, pp. 125–128.

[32] E. Casanova, C. Shulby, E. Gölge, *et al.*, "Sc-glowtts: An efficient zero-shot multi-speaker text-to-speech model," *arXiv preprint arXiv:2104.05557*, 2021.

[33] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 29, pp. 132–157, 2020.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. of Int. Conf. on Computer Vision*, 2017, pp. 618–626.