



Towards Human-like Multimodal Conversational Agent by Generating Engaging Speech

Taesoo Kim, Yongsik Jo, Hyunmin Song, Taehwan Kim

Artificial Intelligence Graduate School, UNIST, Republic of Korea

{taesoo0630, josik, hyunminsong, taehwankim}@unist.ac.kr

Abstract

Human conversation involves language, speech, and visual cues, with each medium providing complementary information. For instance, speech conveys a vibe or tone not fully captured by text alone. While multimodal LLMs focus on generating text responses from diverse inputs, less attention has been paid to generating natural and engaging speech. We propose a human-like agent that generates speech responses based on conversation mood and responsive style information. To achieve this, we build a novel multi-sensory conversation dataset focused on speech to enable agents to generate natural speech. We then propose a multimodal LLM-based model for generating text responses and voice descriptions, which are used to generate speech covering para-lingual information. Experimental results demonstrate the effectiveness of utilizing both visual and audio modalities in conversation to generate engaging speech.

Index Terms: human-computer interaction, computational paralinguistics

1. Introduction

In real life, people communicate through multimodal signals and interpret others' non-verbal cues. This highlights the importance of multimodal understanding, where words, facial expressions, and speech vibes contribute to interpreting meaning. Furthermore, individuals adapt their responses based on these cues, not only in what they say but also in how they express it, reflecting subtle differences in tone, emphasis, and delivery.

Recently, communication with machines has made significant progress due to the remarkable success of large language models (LLMs), which demonstrated a high level of common knowledge [1]. For instance, text-based QA systems [2, 3], visual QA systems [4, 5], video QA systems [6, 7], audio-video QA systems [8] can interpret text, video and audio inputs. Despite these advances, these models are currently only capable of generating text responses. There have also been attempts to generate other modalities using LLMs. These models try to retain the semantic information of the input but often struggle with cross-modal consistency [9, 10] or lose acoustic details in generated speech due to usage of speech tokens [11, 12]. Integrating a text-to-speech (TTS) module with LLMs is a straightforward approach that enables effective speech interaction. However, current TTS modules [13, 14, 15] are inadequate for human-like communication that considers paralingual information reflecting the mood of communication.

Developing the proposed conversational agent requires a large-scale corpus of multimodal interactive conversation data. However, this presents a significant challenge due to the limitations of existing datasets, which are often constrained by their smaller size or the lack of certain modalities, such as audio. To

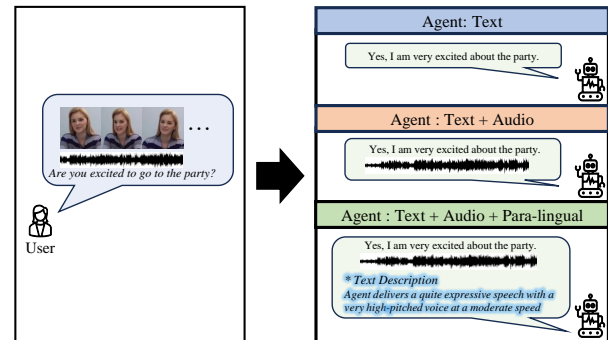


Figure 1: A conversational agent with (Top) text, (Middle) text and audio, (Bottom) text, audio, and paralinguistic signals.

overcome these limitations, we present a new dataset, *Multi-Sensory Conversation (MSenC)* dataset. Our dataset is a carefully curated collection of about 31,000 utterances extracted from daily conversation YouTube videos. The creation of such a conversational model relies on exposure to this diverse range of multimodal conversation dataset and requires the seamless integration of textual, visual, and acoustic elements. To comprehend multimodal information in conversations, we adopt the BLIP-2 [16] approach to ensure efficient cross-modal training. Finally, to communicate with paralinguistic components derived from the overall communication mood, we utilize LLM and instruction tuning which can guide our model in generating voice descriptions. By generating responsive voice descriptions that consider the conversation history, we can enhance the naturalness and contextual appropriateness in dialogue systems, as illustrated in Figure 1.

The contributions of our work can be summarized as follows: To the best of our knowledge, we are the first to study a dialogue model incorporating para-lingual output in responses. We generate a response with paralinguistic information reflecting multimodal factors in conversation. We introduce the Multi-Sensory Conversation dataset, which will be publicly available to advance research in multimodal conversational agents. Our model effectively utilizes both visual and auditory modalities, producing contextually appropriate speech responses, as validated by both quantitative metrics and qualitative assessments.

2. MSenC Dataset

Most existing multimodal conversation datasets [17, 18, 19] focus on single-speaker utterances and lack comprehensive multimodal features. Another dataset [20] provides facial images and audio in communication but shows fixed spatial information such as a green screen background. Notably, a dataset [21] covers many requirements but is designed for emotional analysis,

leading to imprecise audio splitting and background noise from the audience. To effectively communicate in a more human-like way, a dataset that encompasses conversing with human faces, rich visual context, and high-quality voice is desirable.

To address these limitations, we have taken the initiative to develop our novel dataset, the MultiSensory Conversation (MSenC) dataset depicted in Figure 2. This dataset, sourced from YouTube and designed for daily conversation, ensures that there is no background music and the spoken English is clear and high quality, preventing overlaps, disfluencies and non-speech vocalisations. Each scene includes a person, presenting natural conversations with rich visual and auditory elements that help enhance the contextual understanding of dialogue situations. These videos offer a diverse range of voice features and interactions across various scenarios and contexts, crucial for developing robust models. Consequently, we divided the video content into 1,120 dialogues and 31,409 utterances. The total video length is 21.5 hours and the average duration of an utterance is 2.46 seconds.

2.1. Preprocessing

2.1.1. Dialogue Split

Manually segmenting over 36 hours of videos based on speech is a challenging task. However, it is crucial to carefully check for any unnecessary parts to ensure the content is suitable for learning conversations. So we proceeded to manually segment and filter out the dialogue by human to ensure fairness and accuracy following criteria: 1) When multiple dialogues occurred within the same context such as the individuals involved changed. 2) When the scene transitioned to a different setting during the conversation.

2.1.2. Utterance Split

To efficiently segment the dialogue into individual utterances, we can employ speaker diarization, which identifies speakers in an audio recording and assigns timestamps to their speech. While this approach faces challenges, such as difficulty in accurately distinguishing speakers and a tendency to overly fragment. To address these issues, we incorporated automatic speech recognition (ASR) with timestamp capabilities. In our approach, we utilized a pre-trained ASR model¹ that trains OpenAI’s Whisper-large-v3 [22] on English-only data, providing more accurate and faster inference speeds. However, since this model is trained for audio clips up to 25 seconds long, it struggles to accurately timestamp longer clips. To overcome this, we applied a scene detector² to divide longer audio into shorter clips. For clips that are still longer than 25 seconds, we employed speaker diarization³. This method allowed us to more effectively segment the entire video into distinct speech units, each corresponding to individual speakers.

2.2. Metadata Processing

2.2.1. Speaker Assignment

We assign a speaker ID to each video clip according to dialogue units. While speaker diarization is the desirable method for indexing speakers to utterances, it has limitations in performance. We take an alternative approach to address this limitation: cluster the speech embedding. We obtain speech embeddings from

¹<https://huggingface.co/distil-whisper/distil-large-v3>

²<https://github.com/Breakthrough/PySceneDetect>

³<https://huggingface.co/pyannote/speaker-diarization-3.1>

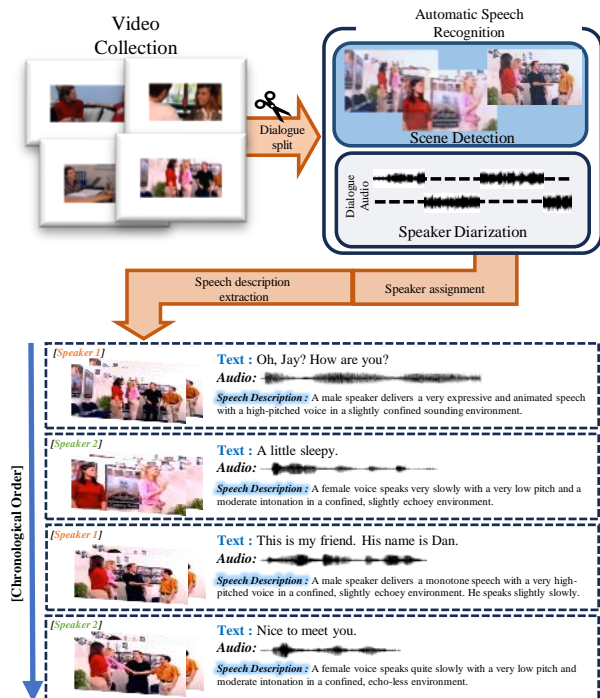


Figure 2: The illustration depicts the creation process of the MultiSensory Conversation dataset.

each speech clip using WeSpeaker [23], a model focused on learning speaker embedding, particularly for speaker verification tasks. Specifically, we use the HDBSCAN [24], an algorithm capable of handling varying densities and does not require predefining the number of clusters. This flexibility is particularly important in our environment, where the number of participants is variable. We employ cosine distance as the distance measure since most speaker verification systems utilize cosine similarity for evaluation. To assess the quality of our process, we calculated the accuracy of our method by manually labeling 20 dialogues, which included a total of 602 samples, resulting in an accuracy of 95.49%. This demonstrates the effectiveness of our speaker assignment method.

2.2.2. Speech Description

Since our goal is generating engaging speech, we extracted speech descriptions that accurately capture the characteristics of the speech. Parler-TTS [25] is a text-to-speech system that transforms text into speech, incorporating detailed paralingual descriptions. This system provides methods⁴ for extracting annotations of speaking style and generating audio descriptions, derived from these annotations.

For processing the MSenC dataset, we extract annotations including gender, pitch, speech monotony, speaking pace, and reverberation. Especially for gender, which is needed for generating speech descriptions but cannot be derived directly, we perform gender recognition⁵ from raw speech, achieving an F1 score of 99.93%. Subsequently, the LLaMa-3 [2] generates natural language descriptions that effectively convey the conversation mood based on these annotations.

⁴<https://github.com/huggingface/dataspeech>

⁵<https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech>

Modality	MSenC				MELD [21]			
	B@1	B@3	METEOR	ROUGE	B@1	B@3	METEOR	ROUGE
Text	12.30	4.11	5.81	11.90	7.99	1.60	4.47	8.09
Text + Audio	12.96	4.82	6.27	11.83	9.10	2.11	4.35	8.24
Text + Video	14.62	4.78	6.63	13.38	5.62	1.00	2.53	4.03
Text + Audio + Video	15.11	5.25	6.89	14.12	10.23	2.19	4.74	9.88

Table 1: Ablation study on different modalities across two datasets. The text-only modality model represents a pure LLM that has been fine-tuned with each dataset. "B" stands for BLEU score.

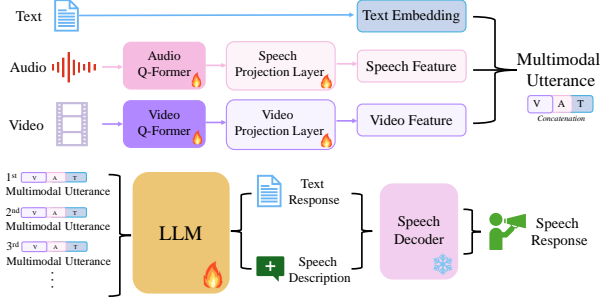


Figure 3: Overview of model architecture. The multimodal utterances are composed of text, audio, and video features. Then LLM generates text response and speech description with them.

3. Method

We develop a model capable of processing multiple modalities and generating engaging speech within a large language model. Figure 3 shows the overview of our architecture. Our model takes a set of images, audio, and text as a single utterance input and generates a responsive textual sentence output along with a voice description. We denote our dataset as $D = \{d^a, d^v, d^l\}$ where a represents the acoustic modality, v the visual modality, and l the linguistic modality, with d indicating dialogue. Each dialogue consists of a set of utterances. Let $d^m = \{u_1^m, u_2^m, \dots, u_t^m, u_{t+1}^m\}$ represent a single dialogue, where t denotes time step, and m presents a certain modality.

3.1. Multimodal Understanding

We utilize Q-Former following Video-LLaMA [8], where video and audio modalities are processed using Q-Formers with the same structure as Blip-2 [16]. This Q-Former has demonstrated strong performance by enhancing computational efficiency and model stability. It produces fixed-size features regardless of the length of input video and audio, which simplifies the integration of multimodal data and ensures consistent input sizes for subsequent processing. To initialize the Q-Former, we adopt the pretrained model from Blip-2 [16]. These models are then fine-tuned to enable our model to capture visual context and auditory information effectively.

For single utterance $u_t^m = \{u_t^a, u_t^v, u_t^l\}$, the video and audio inputs are processed separately. For video sampling, we uniformly extract three frames per second and consider the list of images as a conversation scene. In contrast, the audio processing takes the entire speech as input. While video sampling is conducted to reduce redundant information and improve efficiency, the same method cannot be applied to audio due to significant information loss. The sampled data are applied to their respective Q-Former and then projected into the text embedding space of a large language model through the linear projection layer. The resulting features are concatenated with those from other modalities and used as utterance representations.

3.2. Speech Description Generation

The processed utterances are combined with the conversation history $\{u_1^m, u_2^m, \dots, u_{t-1}^m, u_t^m\}$ and fed into the LLM to understand the context comprehensively. To provide richer communication, we train our model to incorporate not only linguistic information but also paralinguistic cues by describing voice. This is achieved through instruction tuning, a new process where voice descriptions are created after the language model generates responses. Ultimately, the model generates the response and speech description in textual format. We also provide instructions to specify which speaker delivers each utterance, enabling the model to respond or continue the previous utterance.

3.3. Training Loss

We use a target response sentence paired with its corresponding audio description. The cross-entropy loss is then computed between the target $\{u_{t+1}^l, desc_{t+1}\}$ and the model output $\{\hat{u}_{t+1}^l, \hat{desc}_{t+1}\}$, as illustrated in Equation 1, using the concatenation operation denoted by \parallel .

$$Loss = CE(u_{t+1}^l \parallel desc_{t+1}, \hat{u}_{t+1}^l \parallel \hat{desc}_{t+1}) \quad (1)$$

Furthermore, we fine-tune the LLM backbone with parameter efficient fine tuning [26] to specialize the model specifically for generating paralinguistic descriptions in the conversation.

4. Experiment

4.1. Experimental Setup

In our experiments, we evaluate our model on the MSenC dataset and MELD [21]. We extract visual features using CLIP-ViT [27]. The acoustic features are obtained from WavLM [28]. We utilize Mistral-7B [29] as our LLM backbone and utilize Parler-TTS [25] as our speech decoder. Our experimental environment was conducted using a single NVIDIA A100 80G GPU. We use a batch size of 6 and training has spent 30 hours.

4.2. Text Analysis

4.2.1. Modality Ablation

Since our model processes multimodal input, we assess how each modality impacts performance by examining changes in metrics. METEOR [30] measures not just word overlap but also semantic meaning, and ROUGE [31] measures coherence and flow. We calculate the score only on the text response, excluding the voice description. Table 1 shows the impact of audio and video modality processed through the Q-Former. According to the MSenC dataset result, incorporating additional modalities enhances the quality of the responses, indicating a positive effect on multimodal understanding. Specifically, combining audio, video, and text yields the highest performance, suggesting that responses are more appropriate, contextually accurate, and natural-sounding. Similar results are observed with the MELD

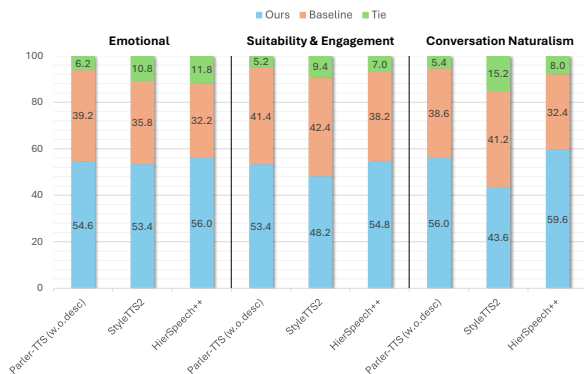


Figure 4: User study results on the MSenC test dataset.

Model	Accuracy
Ours	15.10%
Parler-TTS [25] (w.o. description)	11.20%
StyleTTS2 [14]	13.72%
HierSpeech++ [15]	12.54%

Table 2: Result of evaluating emotional continuity in conversations on the MSenC test dataset.

[21], where incorporating audio and video inputs also results in the highest performance.

4.3. Speech Analysis

4.3.1. User Study

We conducted a human evaluation how to assess the conversational speech style generated by our system contributes to expressiveness, focusing on the impact of paralinguistic information. We used Amazon Mechanical Turk for the assessment, which involved 5 judges and 100 generated samples. The conversational history was limited to a maximum of five entries, presented through video content. The evaluation focused on three criteria: emotional, suitability & engagement, and conversational naturalness. We compared our system with StyleTTS2 [14], which applies a suitable speaking style to the input text, HierSpeech++ [15], a zero-shot speech synthesis framework that enhances robustness and expressiveness, and Parler-TTS [25], which can generate speech from a natural language prompt, and in our comparison, we evaluated it without such a prompt. Comparisons are made against their official checkpoints. The speech sample was generated from the testset of the MSenC dataset. As in Figure 4, our model shows superior results on every criterion, which demonstrates the effectiveness of our approach to generate expressive speech reflecting conversation mood.

4.3.2. Emotion Classification

We classify the emotions of each utterance and calculate accuracy based on whether the emotions match. The process assumes that if an utterance’s emotion aligns with the previous one, it is considered empathetic, highlighting continuity in dialogue. Using the MSenC dataset and a pretrained speech emotion classification model⁶, we classify one of eight emotions: angry, calm, disgust, fearful, happy, neutral, sad, or surprised. Our model generates each speech and then compares it with a prior speech to assess emotional consistency. Table 2 demon-

⁶<https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition>

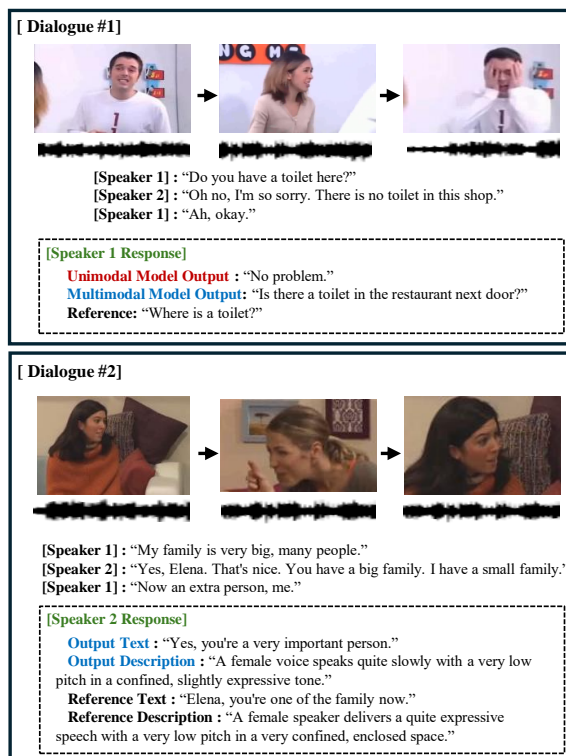


Figure 5: Qualitative analysis samples evaluated on the MSenC test dataset.

strates that our model outperforms the baseline models in maintaining consistent emotional expression across the conversation.

4.4. Qualitative Analysis

Our previous analysis highlights our model’s ability to understand multimodality and determine how the text should be spoken. However, it is worth noting that metrics alone might not capture the full essence in an open-domain scenario. Consequently, we present a comparative analysis illustrated in Figure 5. Dialogue # 1 compares the multimodal model with the text-based unimodal model. The speaker’s gestures in the video and tone of voice in the audio convey an urgent situation. These additional modalities enable our model to generate more contextually appropriate responses. In Dialogue #2, the output shows that our model generates speech descriptions with similar characteristics to the reference, including pace, pitch, and tone. This leads to more engaging and contextually suitable speech responses. Overall, our model has a better understanding of multimodal inputs, generating engaging responses that closely match the context and improve relevance. The generated speech samples are included in the supplementary material.

5. Conclusion

We study a dialogue model with visual and audio inputs from a speaker and generate engaging speech. We propose a novel dataset that is suitable and curated for training such models. Then we propose a novel conversation model that outperforms the baseline in experiments and thus shows its effectiveness. Our model cannot replicate a speaker’s exact voice from historical recordings, but this does not affect inference since the agent consistently uses a single voice. We believe our approach contributes to more natural and human-like conversation and our proposed dataset may promote subsequent research.

6. Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-II220608/2022-0-00608, Artificial intelligence research about multimodal interactions for empathetic conversations with humans, No.IITP-2025-RS-2024-00360227, Leading Generative AI Human Resources Development & No.RS-2020-II201336, Artificial Intelligence graduate school support(UNIST)) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00219959).

7. References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [2] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [3] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, “Stanford alpaca: An instruction-following llama model,” <https://github.com/tatsu-lab/stanford-alpaca>, 2023.
- [4] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, 2024.
- [5] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” *arXiv preprint arXiv:2310.03744*, 2023.
- [6] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” *arXiv preprint arXiv:2409.12191*, 2024.
- [7] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, “Llava-onevision: Easy visual task transfer,” *arXiv preprint arXiv:2408.03326*, 2024.
- [8] H. Zhang, X. Li, and L. Bing, “Video-LLaMA: An instruction-tuned audio-visual language model for video understanding,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Y. Feng and E. Lefever, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 543–553. [Online]. Available: <https://aclanthology.org/2023.emnlp-demo.49/>
- [9] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, “Next-gpt: Any-to-any multimodal llm,” *arXiv preprint arXiv:2309.05519*, 2023.
- [10] Z. Tang, Z. Yang, M. Khademi, Y. Liu, C. Zhu, and M. Bansal, “Codi-2: In-context interleaved and interactive any-to-any generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 425–27 434.
- [11] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li *et al.*, “Anygpt: Unified multimodal llm with discrete sequence modeling,” *arXiv preprint arXiv:2402.12226*, 2024.
- [12] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, “Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities,” *arXiv preprint arXiv:2305.11000*, 2023.
- [13] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, “Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers,” *arXiv preprint arXiv:2304.09116*, 2023.
- [14] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, “Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [15] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, “Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis,” *arXiv preprint arXiv:2311.12454*, 2023.
- [16] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.
- [17] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” *arXiv preprint arXiv:1804.03619*, 2018.
- [18] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, “Mead: A large-scale audio-visual dataset for emotional talking-face generation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 700–717.
- [19] H. Chu, D. Li, and S. Fidler, “A face-to-face neural conversation model,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7113–7121.
- [20] S. Park, C. Kim, H. Rha, M. Kim, J. Hong, J. Yeo, and Y. Ro, “Let’s go real talk: Spoken dialogue model for face-to-face conversation,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 16 334–16 348. [Online]. Available: <https://aclanthology.org/2024.acl-long.860/>
- [21] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “Meld: A multimodal multi-party dataset for emotion recognition in conversations,” *arXiv preprint arXiv:1810.02508*, 2018.
- [22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.
- [23] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, “Wespeaker: A research and production oriented speaker embedding learning toolkit,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [24] L. McInnes, J. Healy, S. Astels *et al.*, “hdbscan: Hierarchical density based clustering,” *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017.
- [25] D. Lyth and S. King, “Natural language guidance of high-fidelity text-to-speech with synthetic annotations,” *arXiv preprint arXiv:2402.01912*, 2024.
- [26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [28] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [29] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, “Mistral 7b,” *arXiv preprint arXiv:2310.06825*, 2023.
- [30] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
- [31] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.