STILL THINKING OR STOPPED TALKING? DIALOGUE SILENCE INTENTION CLASSIFICATION USING MULTIMODAL LARGE LANGUAGE MODEL

Muyun Wu, Zi Haur Pang, Koji Inoue, Tatsuya Kawahara
School of Informatics, Kyoto University, Kyoto, Japan
{wu, pang, inoue, kawahara}@sap.ist.i.kyoto-u.ac.jp

ABSTRACT

Silence during conversation is a crucial element. Silence may indicate thinking, memory retrieval, or the decision-making process, not just the end of the utterance. Deep understanding of silence is a key to developing natural Spoken Dialogue Systems (SDSs). We frame silence understanding as a multimodal classification task, aiming to distinguish whether a user's long pause indicates they are **Thinking** or they have **Stopped** the utterance. To support this research, we construct a new multimodal dataset collected from 63 speakers interacting with an "attentive listening" system designed to elicit natural pauses. Each silent interval over two seconds is labeled based on preceding Japanese linguistic cues, non-verbal behaviors such as gaze during the pause, and subsequent user actions. Using this dataset, we develop a Multi-modal Large Language Model (MLLM) named SilenceLLM for this audio-visual classification task. Experiments on our dataset show that our model achieves 0.857 marco F1 score after comparing several different Audio Encoders and Visual Encoders, accompanied by a Q-former as a fusion model.

Index Terms— Spoken Dialogue Systems, Silence, Dataset, Multi-modal Large Language Models

1. INTRODUCTION

In human communication, silence is not merely the absence of speech but a socially and pragmatically meaningful phenomenon [1]. Conversational analysts have long shown that silences play a crucial role in turn-taking, emotional states, or interactional stances [2]. However, most current SDSs either disregard silence or treat it as an error state [3], thereby missing essential cues about the speakers' intentions and missing its communicative value. Without the ability to interpret silence, conversational agents risk producing responses that feel unnatural or socially inappropriate. Under this discussion, a long silence does not just indicate that the conversation has stopped. It may also mean that the user is thinking. We simplify this problem by framing silence understanding as a classification problem: determining if a user's long pause indicates they are **Thinking** or have **Stopped** and are awaiting a response.

Understanding of someone's silence is based on a lot of contextual information. In the presence of silence, meaning must be inferred from the preceding dialogue (audio context) and the user's non-verbal cues (visual information). This makes it an ideal audio-visual classification problem, architecturally similar to models used for audio-visual video captioning that also capture complementary information from different streams. Previous research on SDSs has employed linguistic, acoustic [4], and visual [5, 6] features, often combining them to improve performance [7, 8]. More recently, MLLMs have integrated these modalities to enhance predictions [9]. Furthermore, recent studies [10, 11, 12] have demonstrated that large language models (LLMs) can act as effective multi-modal speech learners by leveraging their contextual modeling capabilities to generate tokens in an autoregressive manner. Motivated by this, in this work, we develop and evaluate a MLLM architecture, combining vision and audio encoders. By training this model on our dataset, we compare its performance with other MLLMs in classifying user intent during the conversational silence to improve the interactional quality of SDSs.

2. DATASET

2.1. Dataset Construction

In this work, we build a dataset from interactions between 63 Japanese native speakers and a semi-autonomous attentive listening system. The system mainly responds with backchannels and short utterances automatically, and an operator will intervene in the conversation when the user is disengaged. The dataset includes video recordings of the full dialogue (face and upper body) and audio recordings containing only user speech, with system responses excluded to focus on user silence analysis.

To extract silence intervals, we employ a Voice Activity Detection (VAD) model. We first identify all silence segments longer than 0.1 seconds and calculate their durations. The average silence duration in our dataset is 1.50 seconds. Based on this observation, we retain only two-second silence periods; if it lasts longer than two seconds, we cut it off. Shorter pauses like backchannels or brief sentence-final pauses are typically insignificant, speakers usually resume talking quickly after such breaks. In addition, the average

	Stopped	Thinking	#Speakers
train	624	274	36
val	168	105	12
test	221	122	15

Table 1. Statistics of train, validation, and test sets.

speech rate in our dataset is 3.78 words per second, meaning that approximately five seconds of speech provides sufficient contextual information to determine the type of silence. Therefore, we segment each video into clips consisting of five seconds of speech preceding a two-second silence.

2.2. Annotation

During the data annotation process, we listened to the user's speech preceding the silence while simultaneously observing their body movements and facial expressions. Additionally, we examine the user's behavior within five seconds before the silence of two seconds to annotate each video segment. Specifically, in Japanese, sentence-ending expressions such as "masu", "desu", and "deshita" are strong indicators that the user has likely concluded the current utterance, while conjunctions such as "kedo", "de", and "soshite" suggest continuation. Building upon this, we further relied on users' gaze and actions during silence for interpretation. For instance, gazing directly at the camera often indicates an expectation of response, whereas looking downward or displaying a wandering gaze suggests a lower willingness to continue the exchange. In order to maintain objectivity, we also check the user's behavior after a two-second silence to ensure our judgment is accurate. If the user continues to speak or is suddenly interrupted by the system (which can be inferred from their gaze and abrupt movements indicating their thoughts were interrupted) after a two-second silence, we will annotate it as "Thinking". If the user still maintains silence and acts like waiting for a response, or the system starts to give a response, this situation will be judged as they have "Stopped" speaking. The statistics of our dataset are shown in Table 1. The speakers are not overlapped between the training, validation, and test sets

3. PROPOSED METHODS

As shown in Fig.1, the architecture of our works includes five components: a vision encoder, an audio encoder, a selectable speech rate predictor, an AV Q-former, and an LLM decoder, which will predict sentence-like outputs from multimodal tokens. The vision encoder will convert videos or each frame into fixed-length features at 25Hz. When the audio inputs are provided, audio encoders will also encode them into fixed-length features at 25Hz and truncate them to align with the length of video features. Then we tried several early feature fusion methods to combine different features, and finally,

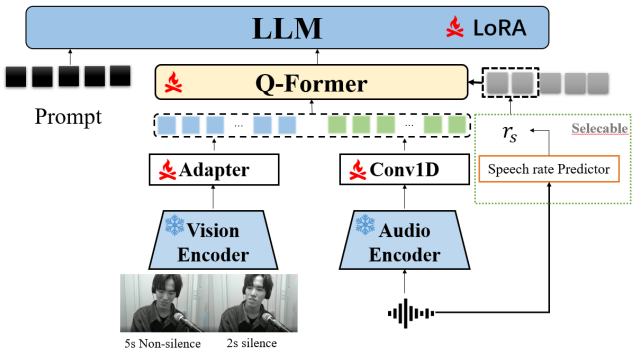

Fig. 1. Model Architecture. Our model is composed of a vision encoder, an audio encoder, Q-former, and an LLM decoder. The speech rate predictor used to generate the speech rate r_s is optional. Input data contains 5s of non-silence interval and 2s of silence interval. We use prompts to guide the MLLM to focus on the silence intervals.

we chose concatenation as the fusion method. For aligning multimodal tokens with the input of LLMs, we also add several parts: Adapter, Q-former, and Speech Rate Predictor. To transfer the features with different lengths into a fixed length, the Q-former use a learnable query to tokenize multimodal features.

3.1. Video Encoder

Our dataset primarily consists of recordings that capture participant's face and upper body. Consequently, the visual encoder must be capable of recognizing both facial expressions and body movements. Moreover, video data inherently contains spatial as well as temporal information. Therefore, a video encoder, rather than a frame-level image encoder, is better suited for our task. Based on this consideration, we evaluate four different vision encoders: CLIP [13], SigLip2 [14], AV-HuBERT [15], and Marlin [16]. CLIP and SigLip2 are popular vision-language encoders, exhibiting strong capabilities in image understanding. AV-HuBERT [33], building upon Audio HuBERT, leverages self-supervised learning to jointly model audio and video features generated by the respective encoders. Marlin [34], designed specifically for extracting facial expressions from video, demonstrates strong transferability across a variety of facial analysis tasks.

3.2. Audio Encoder

In our dataset, speech dominates over non-verbal vocalizations. Since speech conveys critical information about users' intentions, which can underlie silence, it is essential to select an appropriate audio encoder. To extract features from speech, we experiment with two representative encoders: Whisper and HuBERT. Both encoders downsample the raw audio input into features with a temporal resolution of 50

Hz. To further reduce the sequence length, we adopt a 1D convolutional layer to downsample the extracted features to 25 Hz.

3.3. AV Q-former

The Q-former module was originally introduced in BLIP-2 [17] as a mechanism to bridge the gap between a frozen vision encoder and an LLM decoder. In our framework, after the early fusion of audio and video embeddings, we employ the AV Q-former to facilitate multimodal learning by connecting the encoder outputs to the LLM decoder. Inspired by MMS-LLaMA [18], we adopted the pretrained models. Then we further explore following query strategies to enhance alignment:

- Length-variable queries based on speech rate.

$$N_{alloc} = \left\lfloor f_Q \times \frac{T_v}{F_v} \times r_s \right\rfloor. \quad (1)$$

  where F_v is the fps of our videos, T_v is the length of our video embeddings, f_Q is a hyper-parameter which is the number of queries per second, r_s is the speech rate prediction generated from speech rate predictor.

- Fixed-rate query per second. Here, we choose 3 queries per second

3.4. Spatial Temporal Pooling Connector

To adapt the image encoders (CLIP [13] and SigLip2 [14]) for video understanding tasks while preserving both spatial and temporal information, we incorporate a Spatial-Temporal Pooling (STP) Connector. The STP Connector, originally proposed in VideoLLaMA2 [12] composed of 3D convolution together with the ReStage block, reduces the number of spatial-temporal tokens while maintaining their sequential order. This design has shown effectiveness in our experiments, yielding strong performance in multimodal silence classification.

4. TRAINING

4.1. Processing

The original videos in our dataset are recorded at 30 frames per second (fps), and the audio features are sampled at 16 kHz. We resample all videos to 25 fps while keeping the audio sampling rate unchanged at 16 kHz. For CLIP, SigLip2, Whisper, and HuBERT, we employ their official preprocessing pipelines. In contrast, for AV-HuBERT and Marlin, we implement a custom preprocessing procedure: each video frame is first converted to grayscale, then center-cropped, and finally normalized before feature extraction. Then, we downsample the audio features from Whisper and Hubert to 25Hz

4.2. Architechtures

We employ a video encoder and an audio encoder. After concatenating the embeddings from the two modalities, the output representations are passed into the Q-former for cross-modal alignment. For the large language model (LLM) decoder, we adopt variants of Qwen3, Llama3.2-3B, and PerveiverIO [19], which generate sentence-like outputs conditioned on the fused multimodal features.

4.3. Training and Evaluation

We employ the Adam [20] optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.98$ together with a cosine learning rate scheduler. The initial learning rate is set to 10^{-5}, with a warm-up ratio 0f 10%. All projector layers and the Q-former are trained from scratch. For fine-tuning the LLM decoder, we adopt the LoRA [21] approach with a rank of 16, an α of 32, and a dropout rate of 0.05. LoRA is applied to the query, key, value, and output projection layers. During evaluation, we generate outputs using beam search decoding with a temperature parameter of 0.3. As for the prompt, the use of a structured JSON output format not only facilitates the generation process but also enforces strict syntactic constraints, thereby ensuring consistency in the model's responses. The prompt is defined as: *"Please analyze the silent section at the end of the video and determine which of the following two categories best describes it: class: 0, content: stopped class: 1, content: thinking Return your answer as a single JSON object, choosing only one of the two categories above."*

5. EXPERIMENT RESULTS

5.1. Comparison with different Multi-modal LLM

To evaluate which model configuration best captures silence using audio-visual information, we conduct extensive experiments based on several Multi-modal LLMs, as shown in Table 2. The results indicate that Qwen3-1.7B + SigLip2 (with STPConnector) + Whisper yields the best performance in our experiments. By leveraging Whisper's strong capability in speech understanding together with the Siglip2 and STPConnector's effectiveness in modeling spatial-temporal information, this configuration achieves good results. Then we used McNemar's test to assess whether the difference in classification accuracy between our model and MMS-LlaMA. McNemar's test (with continuity correction) yields $p = 4.31e^{-7} (< 0.05)$, indicating that our model performs significantly better than the baseline (MMS-LlaMA).

5.2. Effect of multi-modal information

To examine whether audio and visual modalities are better than single modality, we conduct ablation studies on these two modalities, respectively. In Table 2, the results indicate that

Model	LLM decoder	Audio Encoder	Visual Encoder	Visual Projector	length-variable	macro F1	weighted F1
Video-LlaMA2 [12]	Llama3.2-3B	Whisper	CLIP	STPConnector	✗	0.855	0.866
MMS-LlaMA [18]	Llama3.2-3B	Whisper	AV-Hubert	-	✗	0.849	0.861
	Llama3.2-3B	Whisper	AV-Hubert	-	✓	0.841	0.854
SilenceLLM (ours)	Qwen3-1.7B	Whisper	SigLip2	STPConnector	✗	**0.859**	**0.870**
- *Audio-only*	Qwen3-1.7B	Whisper	-	-	✗	0.662	0.678
- *Video-only*	Qwen3-1.7B	-	SigLip2	STPConnector	✗	0.392	0.505

Table 2. Comparison with different multi-modal LLMs. We report each architecture (Decoder, Audio Encoder, Visual Encoder, and Visual Projector). And we also show the effect of different modalities. "Length-variable" means whether using length-variable queries. We also test the contribution of audio modality and visual modality.

LLM decoder	Audio Encoder	Visual Encoder	Visual Projector	Length-variable	macro F1	weighted F1
Qwen3-0.6B	Whisper	SigLip2	STPConnector	✗	0.839	0.852
Qwen3-1.7B	Whisper	AV-Hubert	-	✗	0.819	0.837
	Whisper	Marlin	-	✗	0.843	0.855
	Whisper	SigLip2	STPConnector	✗	**0.859**	**0.870**
	Whisper	SigLip2	STPConnector	✓	0.848	0.861
	Hubert	SigLip2	STPConnector	✗	0.491	0.568
Llama3.2-3B	Whisper	AV-Hubert	-	✗	0.849	0.861
	Whisper	Marlin	-	✗	0.818	0.834
	Whisper	SigLip2	STPConnector	✗	0.849	0.861

Table 3. Ablation study. We compare different Audio Encoders, Visual Encoders, and Decoders.

the visual modality leads to poor performance, and the audio modality contributes more. However, we confirm that the synergy of using visual information can lead to some performance improvement from 0.662 to 0.859 of marco F1. This result shows that although classifying the state of silence depends not only on speech and utterance, it also depends on human action and facial expression information.

5.3. Ablation study

To demonstrate that our model can effectively classify silence, we conduct comparisons across multiple audio encoders, visual encoders, and LLM decoders. As shown in Table 3, Whisper significantly outperforms HuBERT in this task, probably because Whisper understands speech content better. For visual encoders, AV-HuBERT pretrained on the multi-modal MuAViC dataset [22] shows strong capability, but SigLip2 + STPConnector is more powerful in capturing spatial-temporal information, because SigLip2 developed techniques into a unified recipe. However, the use of length-variable queries, whose lengths are determined by a speech rate predictor, does not lead to any noticeable improvement in this setting, because there are only slight differences in speech rate in our dataset. Qwen3-1.7B, a lightweight LLM, achieves better performance than Llama3.2-3B. Qwen3-0.6B also has a strong performance, despite its relatively small number of parameters. This result indicates that Qwen3, as an LLM with a thinking mode and an MoE architecture, possesses strong reasoning capabilities and can effectively infer silence states.

6. CONCLUSION

We have proposed SilenceLLM, equipped with SigLip2 as the visual encoder, incorporating STPConnector as a video embedding projector, and Whisper as the audio encoder, together with early audio-visual fusion and an AV Q-former. Furthermore, we construct a new multimodal dataset in which each video is annotated based on contextual information before and after the silence period, thereby ensuring the objectivity in the labeling process. On this dataset, our model achieves a macro F1-score of 0.859, highlighting the effectiveness of the proposed framework.

In the future, we will collect more data and make better use of the capabilities of LLMs, not just for classification but also for captioning. At the same time, we aim to build a fast streaming system and apply it in real-world scenarios.

7. ACKNOWLEDGEMENTS

This study was supported by the JST Moonshot R&D Goal 1 AvatarSymbiotic Society Project (JPMJMS2011) and JSPS KAKENHI 25H01142.

8. REFERENCES

[1] Rui Zhe Goh, Ian B Phillips, and Chaz Firestone, "The perception of silence," *Proceedings of the National Academy of Sciences*, vol. 120, no. 29, pp. e2301463120, 2023.

[2] Michal Ephratt, "The functions of silence," *Journal of pragmatics*, vol. 40, no. 11, pp. 1909–1938, 2008.

[3] Shuo-yiin Chang, Bo Li, Tara N Sainath, Chao Zhang, Trevor Strohman, Qiao Liang, and Yanzhang He, "Turn-taking prediction for natural conversational speech," *arXiv preprint arXiv:2208.13321*, 2022.

[4] Nigel G Ward, *Prosodic patterns in English conversation*, Cambridge University Press, 2019.

[5] Ziedune Degutyte and Arlene Astell, "The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings," *Frontiers in Psychology*, vol. 12, pp. 616471, 2021.

[6] Kristiina Jokinen and et al Furukawa, "Gaze and turn-taking behavior in casual conversational interactions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 3, no. 2, pp. 1–30, 2013.

[7] Divesh Lala, Koji Inoue, and Tatsuya Kawahara, "Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 78–86.

[8] Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro, "Turn-taking estimation model based on joint embedding of lexical and prosodic contents.," in *Interspeech*, 2017, pp. 1686–1690.

[9] Takeshi Saga and Catherine Pelachaud, "Voice activity projection model with multimodal encoders," *arXiv preprint arXiv:2506.03980*, 2025.

[10] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang, "video-salmonn: Speech-enhanced audio-visual large language models," *arXiv preprint arXiv:2406.15704*, 2024.

[11] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 110805–110853, 2024.

[12] Boqiang Zhang and et al Li, "Videollama 3: Frontier multimodal foundation models for image and video understanding," *arXiv preprint arXiv:2501.13106*, 2025.

[13] Alec Radford and et al Kim, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[14] Michael Tschannen and et al Gritsenko, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[15] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[16] Zhixi Cai and et al Ghosh, "Marlin: Masked autoencoder for facial video representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1493–1504.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven 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. 19730–19742.

[18] Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, and Yong Man Ro, "Mms-llama: Efficient llm-based audio-visual speech recognition with minimal multimodal speech tokens," *arXiv preprint arXiv:2503.11315*, 2025.

[19] Andrew Jaegle and et al Borgeaud, "Perceiver io: A general architecture for structured inputs & outputs," *arXiv preprint arXiv:2107.14795*, 2021.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Edward J Hu and et al Shen, "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, pp. 3, 2022.

[22] Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang, "Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation," *arXiv preprint arXiv:2303.00628*, 2023.