Multimodal Voice Activity Projection for Social Robot Mediation: Expected Behavior and Deployment Constraints

Antonio Cano[1,2], Guillermo Pérez[1], Luis Merino[2] and Randy Gomez[3]

***Abstract*— Turn-taking prediction is especially relevant for social robots that act as mediators in human-human interaction, where the expected action is often not to speak, but to orient, wait, avoid interruption, or prepare a balanced intervention. This paper presents Multimodal Voice Activity Projection (MM-VAP) as a human state-aware perception layer for future robot mediation behavior. The model estimates the future evolution of the conversational floor from synchronized audio-visual evidence and derives turn-taking events such as Hold, Shift, Shift prediction, Backchannel prediction, and overlap-related states. The approach uses VA-related pretrained audio-visual encoders, LoRA adaptation, inter-speaker attention, and zero-shot event inference from future voice activity projections. Experiments on NoXi, NoXi+J, and Haru EDR support the feasibility of this formulation, especially for floor management events that can be connected to gaze preparation, active listening, and conservative intervention. Finally, the paper defines the expected robot output interface and discusses the main deployment constraints, including real-time inference, preprocessing latency, multimodal synchronization, and input-quality monitoring.**

The source codes and pretrained models are available at https://github.com/acano15/MM-VAP.

I. Introduction

Turn-taking is the conversational mechanism that structures spoken interaction and coordinates the roles of speaker and listener [1]. In social robotics, this mechanism is especially relevant because the robot must not only recognize the end of an utterance, but also interpret the evolution of the conversational floor over time. This requirement becomes more demanding when the robot is placed in a human-human interaction as a background mediator, where an incorrect intervention may interrupt a speaker, reinforce an overlap, or reduce the naturalness of the dialogue. Most practical dialogue systems still rely on reactive end-of-turn decisions [2] based on silence or voice activity thresholds. These strategies are useful as low-level safeguards, but they are limited for interaction management because human speakers may pause inside their own turn, produce short backchannels, overlap with another interlocutor, or indicate a transition before silence is actually observed [3]. In addition, it is well established that speech production takes longer than typical turn-transition gaps, which means that humans rely on prediction to prepare their responses before the current speaker's turn is complete. Human turn-taking is therefore better described as a predictive process in which participants continuously estimate whether the current floor will be held, shifted, closed, or briefly supported by listener feedback.

The present work builds on the MM-VAP formulation proposed for the Haru social robot mediation scenario [4], [5]. Haru is not conceived here as one of the main interlocutors, but as an embodied social mediator that observes a dyadic exchange and may support the interaction by regulating conversational flow, balancing participation, managing silences, and expressing socio-emotional behaviors such as active listening. This motivates a representation of turn-taking events not only as benchmark labels, but also as human state estimates that can be consumed by a future robot behavior layer. This work focus on the applications of a MM-VAP model in the context of robot mediation. First, it summarizes the MM-VAP architecture as a compact audio-visual predictive model for conversational floor estimation. Second, it defines an explicit interface between predicted turn-taking events and expected mediation outputs, with special emphasis on gaze and intervention timing. Third, it discusses the real-time deployment constraints that must be solved before the model can be used as an online module in a physical social robot.

II. Related Work

Turn-taking prediction has been extensively studied in spoken dialogue systems and human-robot interaction [2]. Classical approaches usually detect the end of speech from acoustic activity and silence duration, but such decisions are inherently delayed and sensitive to pauses, hesitations, and overlapping speech. Predictive turn-taking models (PTTMs) [6] address this limitation by estimating future voice activity at frame level, allowing the system to prepare a response before the end of the current turn is fully observed. Voice Activity Projection (VAP) [7] formulated turn-taking prediction as a self-supervised task that projects the future binary voice activity (VA) of two speakers over a short temporal horizon. The original audio-based model combines a pre-trained CPC encoder on raw waveforms with a voice-activity history encoder and a transformer that predicts future activity states. Downstream events such as shift, hold, backchannel, and end-of-turn prediction are derived by aggregating probability mass over interpretable future activity patterns. This formulation been extended to multi-party interaction,

[1] Antonio Cano and Guillermo Perez are with 4i Intelligent Insights, Seville, Spain. a.cano@4i.ai, g.perez@4i.ai

[3] Antonio Cano is with Universidad de Sevilla, Seville, Spain. acano4@us.es

[2] Luis Merino is with Universidad Pablo de Olavide, Seville, Spain. lmercab@upo.es

[3] Randy Gomez is with Honda Research Institute Japan, Saitama, Japan. r.gomez@jp.honda-ri.com

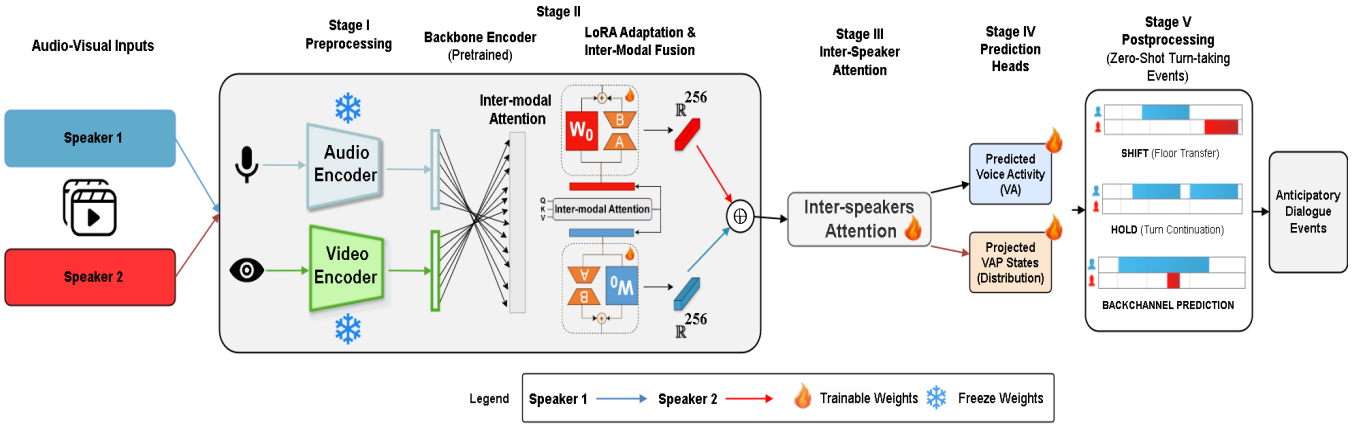

Fig. 1: Proposed MM-VAP architecture for dyadic interaction overview. Blue and red paths indicate the flow from raw inputs to latent representations for Speaker 1 and Speaker 2.

backchannel prediction, multilingual settings, and real-time implementations. Recent multimodal extensions of VAP [8]–[11] have shown that non-verbal information can improve the prediction of conversational events. Previous approaches introduced gaze, facial features, head pose, action units, or generic visual encoders as additional sources of evidence. However, most approaches still rely on engineered descriptors or visual backbones not explicitly pretrained for voice-activity-related dynamics, which limits their suitability for modeling audio-visual turn-taking dependencies. In addition, when audio and visual streams are encoded independently, the model must learn cross-modal relations only during VAP training, often from relatively limited conversational data.

III. Proposed Method: Multimodal Voice Activity Projection (MM-VAP)

This section describes the proposed MM-VAP framework and its complete pipeline, from synchronized audio-visual input sequences to VA and VAP state projection estimation and final inference of turn-taking events.

VAP was formally defined as the task of predicting, or projecting, the future binary VA state of each interlocutor in a dialogue, where each speaker is represented as either active or inactive. MM-VAP extends this formulation by incorporating synchronized audio and visual signals for each interlocutor. The projection is defined over a finite temporal horizon Δt, spanning from the current frame t to a future time $t+\Delta t$. Within this interval, the future binary VA of both speakers is modeled jointly as a single projection window, $w_t^{\Delta t}$, which represents the expected future activity trajectory of the dialogue. The length of this projection window must be selected to capture relevant turn-taking dynamics without representing excessively long or complex future interaction patterns. Following the original VAP implementation, $w_t^{\Delta t}$ is divided into four heterogeneous temporal bins per speaker. The first bins cover the immediate and near future, whereas the later bins represent longer-term regions of the future activity trajectory. Since the interaction involves two speakers, the combination of the four bins for both interlocutors defines a joint output space of $2^{2N} = 256$ possible states. Each state compactly represents one possible future configuration of the interaction horizon between the two interlocutors.

As shown in the Fig. 1 the pipeline is composed of preprocessing, backbone encoding, inter-speaker attention mechanism, prediction heads and turn-taking event zero shot inference. First, the raw audio-visual streams are segmented into speaker-specific context window centered and sincronized at each frame step. For each speaker, the audio chunk waveform is transformed into the spectral representation and the faces sequences are extracted by the face detection module. Both modalities are then synchronized to the same temporal resolution and prepare for the input's requirement of the selected backbone. The backbone selection was based on two criteria: first, the backbone should have achieved high accuracy on its original task related to VA, and second, it must include a robust internal attention mechanism between audio-visual modalities. Therefore, inside the encoder, each modality is processed by its corresponding modality branch, yielding speaker representations in their respective latent spaces. The resulting audio and visual embeddings are fused with the backbone's original inter-modality attention, producing one multimodal embedding per speaker.

In this work, TalkNet and WhisperFlamingo were considered suitable backbone candidates because they satisfy both requirements and provide well-proved pretrained multimodal encoders for ASD and AVSR, respectively, which are closely enough to VA and audio-visual speech understanding. Nevertheless, the original pretrained weights remain specialized for their original task and fully fine-tuning the complete backbone is computationally expensive and may also modify useful pretrained representations. For this reason, LoRA is applied to redirect the capabilities acquired during the original training phase toward future VA projection. LoRA keeps the original weights frozen and introduces trainable low-rank matrices in selected layers reducing the number of training parameters while preserving the original structure and capabilities of the backbone.

Once each speaker embedding is obtained, the model applies a second multihead cross-attention mechanism between both speakers, referred to as inter-speaker attention, to learn how the current multimodal state of one participant influences the other. This stage allows the model to encode turn-taking as a relational process, so the prediction is conditioned on both the speaker's current state and the other speaker's state. The resulting joint representation is then fed to two prediction heads. The first head estimates the multiclass VAP state distribution over the 256 future activity states, while the second head predicts the binary VA probabilities for each speaker. These two outputs are complementary: the VA head provides a direct activity estimation, whereas the VAP head preserves the future-projection structure needed to infer turn-taking events. Finally, the predicted VAP distribution is post-processed to obtain the target turn-taking event categories in a zero-shot manner by aggregating the probability mass of the states associated with each conversational outcome. This inference follows the original self-supervised VAP strategy, where the model learns from future VA patterns derived directly from the dialogue signal and derives higher-level turn-taking events from the predicted VAP state distribution. Shift and Hold describe whether the conversational floor is expected to move to the other speaker or remain with the current one. Shift prediction estimates whether the other speaker is likely to take the floor soon, even while the current speaker is still active. Backchannel prediction estimates whether the listener is likely to produce a short feedback response without taking the floor. Overlap-related states indicate that both speakers may become active within the prediction horizon.

The loss combines VA estimation, future VAP state projection, and semantic regularization of the 256-state space:

$$\mathcal{L} = \mathcal{L}_{VA} + \mathcal{L}_{VAP} + \mu\mathcal{L}_{sem}. \quad (1)$$

Here, $\mathcal{L}_{VA}$ is the binary VA loss, $\mathcal{L}_{VAP}$ is the cross-entropy loss over the 256 future states, and $\mathcal{L}_{sem}$ groups semantically equivalent states, reducing redundant microstates influence.

IV. MM-VAP for Haru Mediation: Expected Outputs and Deployment Constraints

In a social robot mediation scenario, the robot requires an intermediate module that connects conversational floor estimation with observable robot behavior. MM-VAP can provide this connection by producing event-level predictions about the current and future state of the interaction. These predictions can then be translated by the robot controller into gaze, waiting, or mediation actions according to the dialogue context. In this formulation, MM-VAP is treated as a predictive perception layer that estimates who is speaking, who is likely to speak next, and whether the conversational floor is stable, changing, overlapping, or supported by listener feedback.

The most direct output is active-speaker-oriented gaze. When one participant is detected as the active speaker and the model predicts that the floor will be held, the corresponding robot response should be maintain gaze toward that participant, supporting active listening and keeping the robot visually aligned with the current speaker. If a Shift is estimated, the same interface can activate the preparation of a gaze transition toward the expected next speaker. Overlapping speech is associated with a different output: when both speakers become active and the predicted state suggests an interruption-like transition, the proposed gaze target becomes the participant who is taking or contesting the floor. In contrast, backchannel predictions are interpreted as listener feedback, so the current floor configuration is preserved and intervention is suppressed. Together, these gaze updates make the robot's attention consistent with the participant who becomes interactionally relevant and contribute to the perception of the robot as a real engaged mediator.

When the robot has an opportunity to mediate, the expected output should depend on the dialogue history. After sustained silence, repeated floor dominance, or clear participation imbalance, the robot work is to orient toward the less talkative participant before prompting or yielding the floor. When both participants have spoken with similar balance, the robot can alternate gaze between both interlocutors or address the group without privileging one speaker. Therefore, the MM-VAP event stream would get benefits from the combination with longer-term interaction variables such as accumulated speaking time, number of floor acquisitions, silence duration, and overlap frequency. The perception module estimates the conversational floor dynamics, and the robot controller decides whether the robot should maintain gaze, prepare a gaze transition, wait, suppress intervention, or initiate a mediation turn.

The deployment protocol should also indicate the quality of the input signals. Missing faces, unstable tracking, robot self-noise, room noise, occlusions, or speakers outside the camera view can reduce the reliability of the prediction. For this reason, the robot controller should receive information about which modalities are available, together with the predicted event. When the input is incomplete or degraded, the behavior controller should apply more conservative mediation rules.

Real-time deployment is the most challenging aspect of this proposal. To date, no previous work has demonstrated a similar multimodal turn-taking projection system operating online in a physical robot scenario. In a real Haru setting, the system must deal with the natural behavior of users, which makes the interaction less controlled than in offline experiments. At the same time, several demanding modules must run in parallel, including video acquisition, face tracking, audio processing, modality synchronization, event filtering, and robot control. This creates computational and hardware constraints that may compromise inference time, preprocessing latency, memory usage, synchronization, and robustness. Therefore, future work must evaluate the complete online interaction pipeline and assess its real applicability.

V. Experiments and results

The experimental protocol follows the original VAP evaluation. The model is trained on synchronized audio-visual context windows and predicts a 2 s future VA. The main evaluation metrics are accuracy and F1-score for event categories associated with floor development, such as shift/hold, short/long responses, shift prediction, and backchannel prediction. NoXi [12] and NoXi+J [13] provide multilingual dyadic interaction data with expert-novice conversations. These corpora contain natural two-party interaction and synchronized multimodal recordings. The Haru EDR corpus [4] provides the robot-mediated scenario of interest, with semi-structured speaker-listener exchanges conducted in the presence of Haru.

The results indicate that VA-related pretrained backbones provide a clear advantage over the CPC+3DResNet baseline reference. On NoXi, WhisperFlamingo achieved F1 values of 0.97 for both S/H and S-pred, while TalkNet obtained 0.94 for S/H and 0.93 for S/L. On the complete NoXi+J setting, TalkNet provided the strongest BC-pred score among the reported configurations, whereas WhisperFlamingo achieved the strongest S-pred score. These patterns support the hypothesis that pretrained audio-visual representations associated with speech activity can be effectively transferred to future conversational floor projection.

Corpus	Backbone	S/H	S/L	S-pred	BC-pred
NoXi	CPC+3DResNet	0.64	0.70	0.65	0.33
NoXi	TalkNet	0.94	0.93	0.91	0.42
NoXi	WhisperFlamingo	0.97	0.91	0.97	0.45
NoXi+J (Chin.+Jap.)	CPC+3DResNet	0.50	0.61	0.51	0.55
NoXi+J (Chin.+Jap.)	TalkNet	0.84	0.82	0.87	0.62
NoXi+J (Chin.+Jap.)	WhisperFlamingo	0.85	0.80	0.93	0.66
NoXi+J (All)	CPC+3DResNet	0.55	0.59	0.58	0.39
NoXi+J (All)	TalkNet	0.85	0.85	0.89	0.62
NoXi+J (All)	WhisperFlamingo	0.81	0.86	0.90	0.44

TABLE I: F1-score results on NoXi and NoXi+J.

The Haru EDR evaluation was restricted to mediation-related events because backchannel and short/long distinctions provide less direct value for the current mediation objective. The best-performing English configuration was used for each backbone. WhisperFlamingo reached 0.92 F1 for Hold, 0.85 F1 for Shift, and 0.91 F1 for S-pred, while TalkNet reached 0.84, 0.79, and 0.87, respectively. These results suggest that the model captures the floor maintenance and floor transition events that are most relevant for mediation-oriented behavior.

Backbone	Hold	Shift	S-pred
TalkNet	0.83 / 0.84	0.78 / 0.79	0.87 / 0.87
WhisperFlamingo	0.92 / 0.92	0.83 / 0.85	0.91 / 0.91

TABLE II: Haru EDR results Accuracy / F1.

VI. Discussion and conclusions

The proposed formulation treats turn-taking prediction as a human state-aware perception problem for social robot mediation. MM-VAP estimates the future evolution of the conversational floor from multimodal evidence. This is particularly useful for a mediator robot because the relevant action is often not to speak, but to orient, wait, avoid interruption, or prepare a balanced intervention. The explicit mapping between predicted events and expected outputs clarifies how the model can be integrated into a future behavior layer. Active speaker detection supports gaze toward the current floor holder. Shift prediction supports anticipatory gaze toward the next speaker. Overlap detection supports gaze toward the interrupter or new floor holder. Robot turn prediction can support mediation toward the less talkative participant, while backchannel prediction should normally suppress intervention. This mapping remains a design interface and must be validated with users before being treated as a complete mediation policy.

The results obtained on NoXi, NoXi+J, and Haru EDR support the feasibility of using VA-related pretrained audio-visual encoders for MM-VAP. Nevertheless, real-time deployment remains the central open problem. The final system must demonstrate that the predictive gain is not lost through computation, buffering, synchronization, or unstable event postprocessing. Future work will therefore focus on online inference, lightweight adaptation, behavior-level validation, and integration with the complete Haru multimodal dialogue architecture. In this direction, the model is not expected to decide the complete social strategy of the robot. Instead, it should provide an estimation of the conversational floor and a small set of candidate outputs that are meaningful for mediation. Also, the more complex cases, such as deciding when the robot should verbally mediate or whether an overlap requires intervention, should remain managed by additional rules and future user studies.

This also defines a practical discussion point for the next iteration of the work: whether predictive floor-state estimation provides enough temporal advantage to produce better human understanding and fewer inappropriate interventions than a reactive end-of-speech baseline, first in a simple 1:1 human-robot interaction and later in mediated human-human interaction.

Overall, the presented work provides a unified path between multimodal conversational perception and adaptive social robot behavior. The model is not designed to replace the robot dialogue manager, but to provide a predictive representation of the human conversational floor that can be used by a dialogue manager to decide when to observe, when to prepare, and when to mediate.

Acknowledgments: This work was partially funded by the Spanish Ministry of Science and Innovation and the State Research Agency (MCIN/AEI/10.13039/501100011033) under the projects **TIFON** [MIG-20232039 / PLEC2023-010251] (A.C., G.P.), **PICRAH4.0** [PLEC2023-010353] (L.M.) and **LIPTACON**, by CDTI Innovation through the *Programa Tecnológico Espacial (PTE) 2024* via project [PTEP-20241001] (A.C., G.P.).

REFERENCES

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 50, p. 696, Dec. 1974.

[2] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review," *Computer Speech & Language*, vol. 67, p. 101178, May 2021.

[3] A. S. Meyer, "Timing in Conversation," *Journal of Cognition*, vol. 6, p. 20, Apr. 2023.

[4] C. A. Cruz, E. C. Montiel-Vazquez, C. Maeda, and R. Gomez, "When and How to Express Empathy in Human-Robot Interaction Scenarios," in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, (Eindhoven, Netherlands), pp. 1070–1077, IEEE, Aug. 2025.

[5] S. Cooper, R. Gomez, D. Szapiro, and L. Merino, "Design of Social Features for Robot-mediated Cross-cultural Interaction," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, (Boulder CO USA), pp. 351–355, ACM, Mar. 2024.

[6] M. J. Pinto and T. Belpaeme, "Predictive Turn-Taking: Leveraging Language Models to Anticipate Turn Transitions in Human-Robot Dialogue," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, (Pasadena, CA, USA), pp. 1733–1738, IEEE, Aug. 2024.

[7] E. Ekstedt and G. Skantze, "Voice Activity Projection: Self-supervised Learning of Turn-taking Events," May 2022. arXiv:2205.09812 [eess].

[8] K. Onishi, H. Tanaka, and S. Nakamura, "Multimodal Voice Activity Prediction: Turn-taking Events Detection in Expert-Novice Conversation," in *International Conference on Human-Agent Interaction*, (Gothenburg Sweden), pp. 13–21, ACM, Dec. 2023.

[9] K. Onishi, H. Tanaka, and S. Nakamura, "Multimodal Voice Activity Projection for Turn-Taking and Effects on Speaker Adaptation," *IEICE Transactions on Information and Systems*, vol. E108.D, pp. 445–453, June 2025.

[10] T. Saga and C. Pelachaud, "Voice Activity Projection Model with Multimodal Encoders," June 2025. arXiv:2506.03980 [cs].

[11] S. O. Russell and N. Harte, "Visual Cues Enhance Predictive Turn-Taking for Two-Party Human Interaction," May 2025. arXiv:2505.21043 [cs].

[12] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The NoXi database: multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, (Glasgow UK), pp. 350–359, ACM, Nov. 2017.

[13] M. Funk, S. Okada, and E. André, "Multilingual Dyadic Interaction Corpus NoXi+J: Toward Understanding Asian-European Non-verbal Cultural Characteristics and their Influences on Engagement," in *International Conference on Multimodel Interaction*, pp. 224–233, Nov. 2024. arXiv:2409.13726 [cs].