The 3rd Workshop on **Nonverbal Cues for Human-Robot Cooperative Intelligence**

August 24-28, 2026

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

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

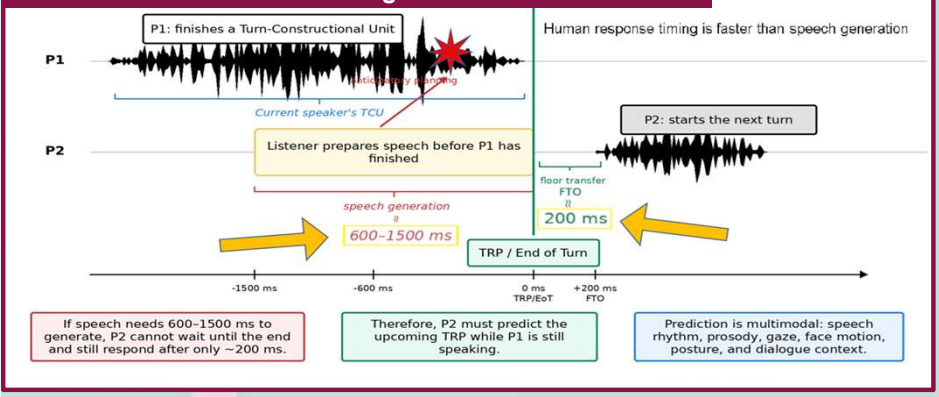
* corresponding author : aantcan@upo.es, a.cano@4i.ai

[1]4i Intelligent Insights company, Seville, Spain [2]Universidad Pablo de Olavide, Seville, Spain

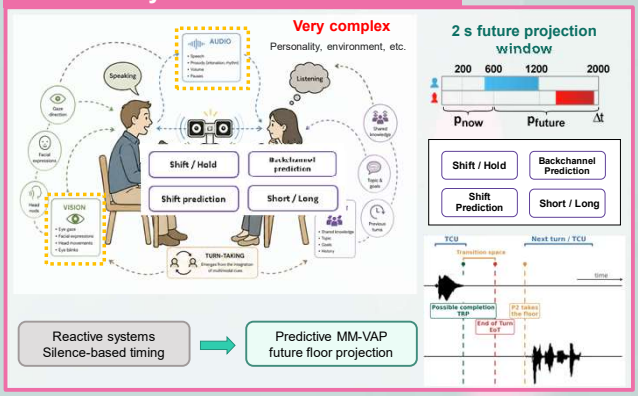
[3]Honda Research Institute, Wako-Shi, Japan

2026 IEEE RO-MAN Kitakyushu, Fukuoka, JAPAN.

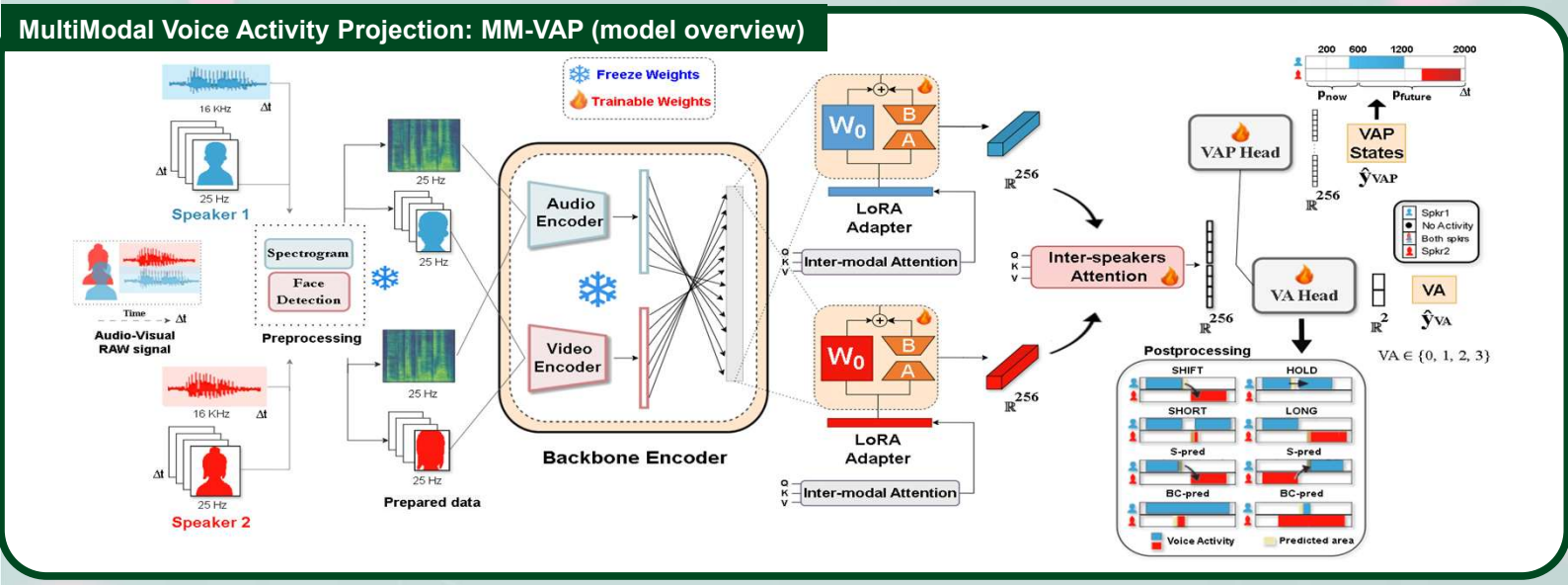
Introduction: Natural Turn-Taking is Predictive & Multimodal



Mediation dyadic multimodal scenario



MultiModal Voice Activity Projection: MM-VAP (model overview)

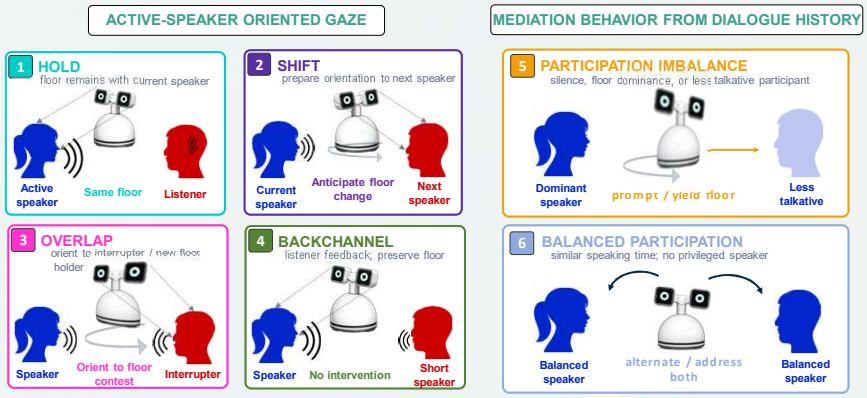


Results

Turn-Taking Events (F1-Score)

Method	Corpus	Subset	S/H	S/L	S-pred	BC-pred
Previous work						
Russell et al.	Candor	—	**0.92**†	—	**0.74**‡	—
Saga et al.	NoXi	French	0.72	0.67	0.75	0.46
Onishi et al.	NoXi	French	0.87	0.80	0.69	0.69
Onishi et al.	NoXi	English	0.95	0.87	0.71	0.79
Onishi et al.	NoXi	German	0.90	0.77	0.69	0.73
Onishi et al.	NoXi+J	Japanese	0.82	0.74	0.65	0.48
Onishi et al.	NoXi	All	0.90	0.81	0.72	0.71
Ours						
CPC+3DResNet (no pretraining)	NoXi	All	0.64	0.70	0.65	0.33
CPC+3DResNet (no pretraining)	NoXi+J	All	0.55	0.59	0.58	0.39
TalkNet	NoXi	All	0.94	**0.93**	0.91	0.42
TalkNet	NoXi+J	All	**0.85**	0.85	0.89	**0.62**
WhisperFlamingo	NoXi	French	0.84	**0.89**	**0.96**	0.52
WhisperFlamingo	NoXi	English	0.82	**0.89**	**0.96**	0.45
WhisperFlamingo	NoXi	German	0.82	**0.91**	**0.97**	0.43
WhisperFlamingo	NoXi+J	Japanese	**0.87**	**0.80**	**0.85**	0.52
WhisperFlamingo	NoXi	All	**0.97**	**0.91**	**0.97**	0.45
Haru EDR results (Accuracy / F1)						
TalkNet	Haru	All	0.83 / 0.84	0.78 / 0.79	0.87 / 0.87	—
WhisperFlamingo	Haru	All	0.92 / 0.92	0.83 / 0.85	0.91 / 0.91	—

Haru expected outputs



Deployments constraint

PARALLEL MODULES
Camera acquisition, face tracking, audio processing, synchronization, model inference, event filtering, and robot control run at the same time

NATURAL USER BEHAVIOR
The Haru setting is less controlled than the datasets: users pause, overlap, interrupt, move, look away, or leave the camera view

COMPUTATIONAL LOAD
Pretrained audio-visual backbones increase inference cost, memory demand, and latency under robot-side hardware limitations

INPUT QUALITY
Missing faces, unstable tracking, robot self-noise, room noise, and occlusions can reduce prediction reliability

MODALITY AWARENESS
The controller needs the predicted event together with modality availability and confidence, so incomplete inputs are visible to the behavior layer

STABLE ROBOT OUTPUT
Frame-level predictions need smoothing before action. Degraded or unstable signals require conservative rules: wait, keep gaze, or suppress intervention

Sensing → Sync → Inference → Filtering → Control

Conclusions & next steps

- VA-related pretrained backbones **outperform** generic encoders and precomputed visual features
- **VA pretrained encoders** transfer effectively from speech-activity tasks to future voice activity projection
- MM-VAP predicts the conversational-floor events most relevant to mediation: Hold, Shift, Overlap and speaker-future activity
- MM-VAP connects multimodal conversational perception with adaptive robot **behavior**
- Current evidences remain **offline**

Future line

- Add **text** modality with semantic completeness prediction
- Unify audio,visual and semantic evidece in a new trimodal **MM-VAP**
- **Full-duplex** interaction with continous listening and **barge-in**
- Online inference and confidence-aware control
- Validate timing in 1:1 dialogue and two participants
- **Real time** implementation with lighter backbones
- **Ablation** design for single-modality unreilable
- Robot real world technology validation and effectiveness in mediation

Acknowledgements

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 Tecnol´ogico Espacial (PTE) 2024 via project [PTEP-20241001] (A.C., G.P.).