TU Delft

Delft University of Technology

Multimodal Conversational Events Estimation in Complex Social Scenes

Li, Litian

DOI
10.1145/3716553.3750821

Publication date
2025

Document Version
Final published version

Published in
ICMI 2025 - Proceedings of the 27th International Conference on Multimodal Interaction

Citation (APA)
Li, L. (2025). Multimodal Conversational Events Estimation in Complex Social Scenes. In R. Subramanian, Y. I. Nakano, T. Gedeon, M. Kankanhalli, T. Guha, J. Shukla, G. Mohammadi, & O. Celiktutan (Eds.), *ICMI 2025 - Proceedings of the 27th International Conference on Multimodal Interaction* (pp. 716-720). ACM. https://doi.org/10.1145/3716553.3750821

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository
as part of the Taverne amendment.

More information about this copyright law amendment
can be found at https://www.openaccess.nl.

Otherwise as indicated in the copyright section:
the publisher is the copyright holder of this work and the
author uses the Dutch legislation to make this work public.

Multimodal Conversational Events Estimation in Complex Social Scenes

Litian Li
Socially Perceptive Computing Lab
Delft University of Technology
Delft, Netherlands
Socially Perceptive Computing Lab
Delft University of Technology
Delft, Netherlands
l.li-11@tudelft.nl

Abstract

Conversational events, such as speaking turns, backchannels, topic changes, and laughter, are central to the structure of multiparty interaction and play a key role in shaping its dynamics. However, detecting such events in real world social settings remains challenging due to perceptual ambiguity, visual occlusion, signal noise, and limitations in acquiring high quality audio data. This work addresses these challenges by focusing on spontaneous interactions in socially complex and privacy sensitive environments, exploring multimodal, nonverbal cues that do not rely on audio. The goal is to develop a novel modeling approach for group context awareness to infer conversational events and support social scene understanding under real world constraints.

Keywords

Social Signal Processing, Multimodal Machine Learning, Multimodal interaction, Human-centered computing

ACM Reference Format:
Litian Li. 2025. Multimodal Conversational Events Estimation in Complex Social Scenes. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3716553.3750821

1 Introduction and Problem Motivation

As intelligent systems become increasingly integrated into everyday environments, their ability to interpret complex, situated social behaviors becomes essential. A foundational step toward this capability is the automatic detection of conversational events—temporally situated behaviors such as speaking status, backchannels, topic transitions, and laughter—which help structure and regulate multiparty interaction [55]. From a broader perspective, contextual conversational events also serve as building blocks for modeling higher level dynamics such as participant involvement, rapport, and the divergence or convergence of dialogue. These dynamics, in turn, support the inference of higher level intentions [35], such as initiating a new conversational trajectory, leaving a group, or terminating a topic.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ICMI '25, Canberra, ACT, Australia
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1499-3/25/10
https://doi.org/10.1145/3716553.3750821

In many real world situations, access to high quality audio and front view visual information commonly used to detect conversational signals such as speaking activity, backchannels, laughter, and topic transitions—is often limited, unreliable, or ethically constrained. At the same time, detecting these events under such conditions presents substantial challenges. In Complex Conversational Scenes (CCS) [19], such as informal mingling or networking situations, interactions are spontaneous, multiparty, and typically lack clearly structured turn taking. This spontaneity, combined with noisy sensory conditions, overlapping speech, and frequent occlusion, results in a loose and variable alignment between observable behavioral cues and the conversational events they represent [45]. As a result, it becomes difficult to robustly interpret the underlying structure of interaction

My research aims to reduce reliance on audio and detailed front view information of individuals by leveraging non-verbal multimodal cues that remain accessible even under challenging sensing conditions. It uses a top down view to capture more robust visual information across participants, enabling scalable deployment while sacrificing fine grained individual detail.

Specifically, my work focuses on modeling non-verbal behaviors such as posture, gestures, and movement at the group level, supplemented by inertial data. It explores how individuals coordinate with one another in conversational dynamics, aiming to develop a multimodal approach for estimating social conversational events that aligns with how humans intuitively perceive interactional structure.

An overview of the challenges, task objectives, and proposed approach is presented in fiture 1

To this end, my research is guided by the following questions:

(1) How can interactional coordination, such as posture, movement, and head orientation be modeled to detect key conversational events, including speaking activity, backchannels, topic changes, and laughter, during spontaneous and unstructured group interactions (e.g., social mingling)?
(2) How can multimodal models remain robust to missing modalities at inference time, for example detecting conversational events without audio at test time, even if audio was available during training?
(3) How can structured, temporally grounded representations be learned from multimodal group behavior to support conversational event detectionwhile ensuring consistent alignment of cues across individuals and modalities?

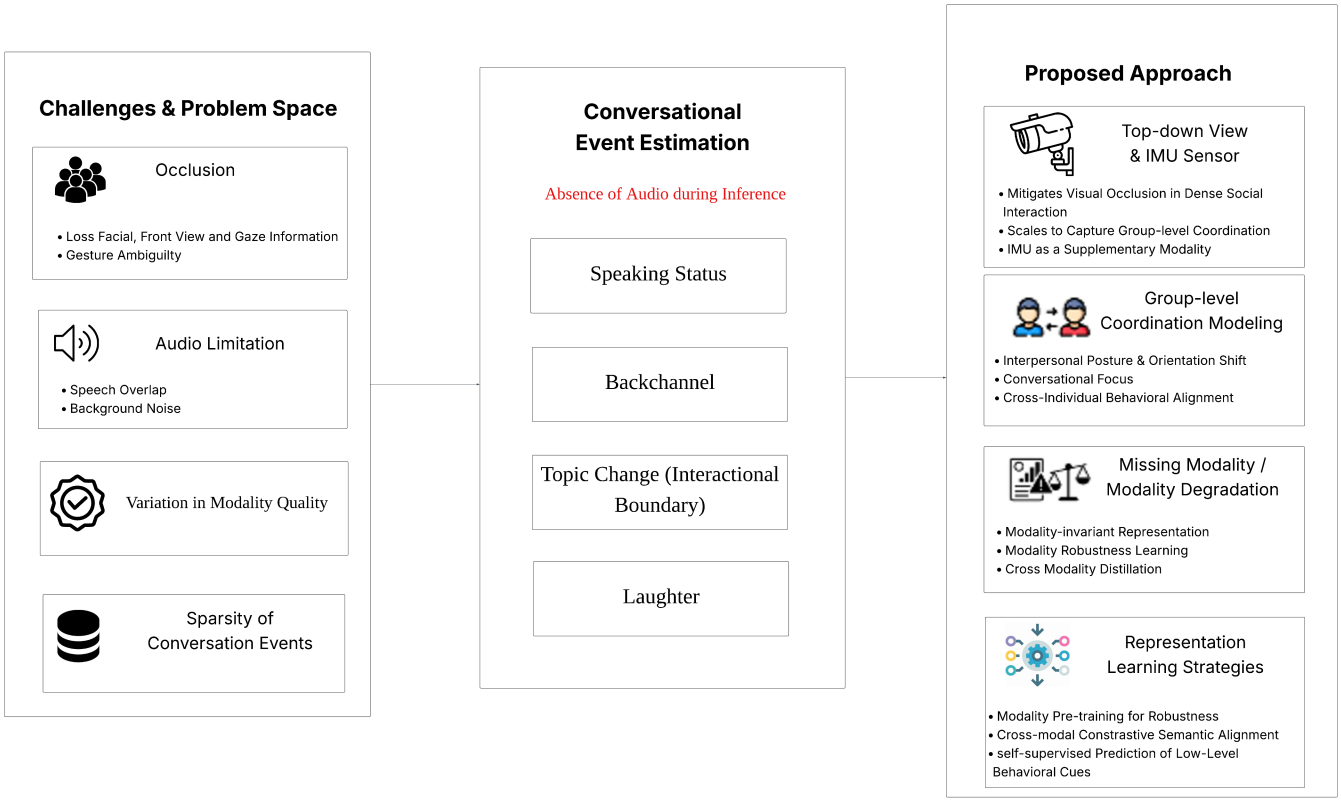

Figure 1: Overview of the problem space and proposed solution.

2 Background and Related Work

2.1 Audio-Centric Approaches and Limitation

Conversational events such as speaking turns, backchannels, laughter, and topic changes play a critical role in shaping the dynamics of human interaction. Traditionally, their automatic detection has relied almost exclusively on the acoustic channel. Voice activity detection techniques infer speaking status and turn taking from energy and prosodic cues [12, 48], and similar prosodic features remain standard for backchannel detection [32, 58]. Laughter is likewise recognized from prosodic and spectral–temporal patterns [2, 25].

Topic segmentation has been treated primarily as a lexical problem, where reliable transcripts presuppose high quality audio [17, 60]. Prosody based boundaries offer an alternative [47], but they also depend on clean acoustic input.

In real world multiparty gatherings, however, audio streams are frequently corrupted by overlapping speech and background noise. These factors significantly degrade the performance of acoustic feature based detection methods [7, 14], motivating a shift toward complementary or alternative modalities.

2.2 Audiovisual progress and its fragility

To overcome audio fragility, audiovisual methods have been developed to integrate facial and vocal signals. These approaches consistently outperform unimodal baselines in controlled environments. For instance, [50] proposed a bimodal recurrent neural network that integrates audio and facial features for robust speech activity detection, significantly outperforming audio only models in noisy environments. Similarly, [52] combined vocal and visual cues such as gestures and facial expressions using LSTM and RNNs to predict head nods and turn taking events, highlighting the benefits of multimodal integration in dyadic interactions.

In laughter detection, audiovisual fusion has also shown promise. Studies such as [3] and [36] demonstrated that combining facial and acoustic signals leads to more accurate recognition of spontaneous smiles and laughter, particularly in distinguishing different levels of expressive intensity.

However, these approaches still rely to some extent on high quality audio and clear frontal facial information. In natural multiparty settings, participants may be partially occluded, facial information may be unavailable due to privacy preserving measures, and environmental noise often persists. Consequently, audiovisual based approaches can be affected when shifting from controlled environments to real world deployment.

2.3 Nonverbal, audio free approaches

When both audio and frontal facial data are unreliable, researchers have explored fully nonverbal alternatives. These methods use body motion, pose, and inertial sensing to detect conversational events. Full frame methods such as S-VVAD localise speakers from raw video by weakly supervised motion segmentation [46], and thoracic optical flow traces of respiration co-vary with speech onsets [29]. Wearable IMUs combined with pose filtered video have proven

effective even with occluded faces [54], and earlier work has shown tight coupling between body motion and speaking activity [13, 53]. Hand motion bursts also flag speaking or listening roles [10]. Beyond turn taking, transformer fusion of pose and facial key points yields audio free backchannel cues [1], and full body dynamics alone can reveal laughter [31, 39].

Despite these advancements validating audio free event detection feasibility, current methods still assume stable upper body visibility and rarely address multimodal coordination among multiple interlocutors. Explicitly modeling such coordination remains a significant research gap, underscoring the motivation behind the present study.

3 Methodology

3.1 Conceptual and Theoretical Foundation

Before addressing conversational event modeling, it is crucial to establish a solid conceptual foundation that supports both current and future research directions. Given the challenges of privacy concerns and unreliable audio or face-based sensing in real-world settings, this work shifts focus toward group-level nonverbal behavioral cues and wearable sensors. These modalities provide a robust, privacy-preserving alternative for capturing conversational structure in socially complex environments.

Understanding dynamic interactions, where participants' intentions may conflict or evolve, requires insights from social signal processing and multimodal communication research. Prior studies highlight how gesture, posture, and speech emerge from a unified cognitive system [28], how rhythmic nesting of modalities links motor behavior to discourse [37], and how subtle nonverbal "honest signals" encode trust and engagement [35]. Social signal processing frameworks model these cues to infer social roles and group structure [55], while gestures and bodily orientation are recognized as integral to interaction organization [24].

Conversational events such as laughter and backchannels are inherently embedded in social and temporal contexts, often aligning with speaking dynamics and topic transitions [9, 16, 18, 22, 61]. These observations suggest that events should be modeled as coordinated, context-sensitive phenomena rather than isolated instances. Multimodal behavioral signals are not mere proxies for missing audio but meaningful indicators of conversational flow, reflecting latent dynamics like alignment, hesitation, or conflict.

To operationalize these perspectives, this research emphasizes group-level coordination signals, patterns of bodily movement, spatial arrangement, and orientation emerging from interaction among participants [24, 55]. Mutual body orientation, shared posture dynamics, and temporal synchrony in gestures provide rich indicators of conversational engagement.

While gaze direction is fundamental for conveying attention and turn-taking, privacy constraints and our top-down view setting often prevent its fine-grained tracking. Therefore, this study employs head orientation as a practical proxy, modeling relative head orientation dynamics to infer shared focus.

Temporal synchrony in upper-body movements, such as co-occurring postural shifts or gestures, signals interactional attunement and group-level engagement [11, 28, 35, 37]. These distributed coordination cues remain valuable when audio and facial data are limited, supporting robust conversational event inference under real-world constraints.

3.2 Dataset

To support this research, I utilize existing multimodal datasets that meet key requirements: availability of audio (at least during training) for event annotation, top-down or wide angle visual perspectives to reduce occlusion and capture group coordination, joint keypoints for modeling posture and gestures, and group-level labels indicating conversational participation. Datasets with finer annotations of nonverbal behaviors are especially valuable for analyzing spontaneous interaction dynamics.

Three datasets are central to this work. MatchNMingle [5] provides overhead video, audio, inertial, and proximity data with annotations of social actions and F-formations in mingling scenarios. PopGlass offers high-quality individual audio, multi-angle overhead video, periodic F-formation labels, and inertial data. ConfLab [42] includes top-down video, full-body keypoints, synchronized IMUs, and detailed annotations of speaking status and laughter. Additionally, I am contributing to a new dataset collection featuring high quality audio, top-down video, and multiple IMUs, designed to support topic change detection and conversational event modeling under privacy-sensitive, multimodal conditions.

These datasets vary in social settings, sensor quality, and annotation density. While this diversity introduces challenges for domain adaptation, it also allows for evaluating representation-level generalizability across different data sources.

3.3 Speaking Status and Backchannel Estimation

The modeling is grounded in observations from multiparty communication, where the relative orientation of individuals' heads and bodies—along with accompanying hand gestures—often reflects the current speaker. Shifts in these orientations typically signal speaker changes. These patterns and transitions also provide valuable cues for identifying individuals likely to produce backchannel responses. However, backchannels are brief in duration and may not always coincide with distinct gestures, making them difficult to detect. To address this, the initial work leverages high frequency speaking status labels from the ConfLab dataset to distinguish short non-speaking intervals or brief turns when backchannels are more likely to occur. Approximate head behavior will be inferred using keypoints following [34], while self-supervised representations developed in [41], supplemented by inertial data, will support group context based estimation of speaking status and backchannel activity.

3.4 Topic Change and Laughter Detection

This research jointly models topic change and laughter, as these phenomena often co-occur and fulfill similar interactional roles [15, 18, 27, 33]. Subtle, individually produced laughter often lacking clear group-level synchrony is difficult to detect from movement cues alone, but incorporating topic dynamics is expected to improve its recognition.

The proposed method leverages group-level posture, gestures, movements, and head-body orientation, based on the hypothesis that shifts in these nonverbal cues co-vary with conversational

topic transitions [6, 38, 43, 45]. Rather than modeling absolute positions, it focuses on relative behavioral changes among participants, employing rotation-invariant representations to ensure spatial robustness, especially from top-down views [21].

Beyond boundary detection, the approach aims to distinguish successful vs. unsuccessful topic shift attempts, and to capture finer sub-topic transitions, informed by conversational topic dynamics literature [4, 44]. Detailed modeling of coordination patterns is also expected to enhance sensitivity to subtle laughter moments occurring during topic re-alignments.

Given the limited annotated data, self-supervised and contrastive learning methods will be employed to extract structured behavioral patterns from sparse supervision [8, 41]. Transformer-based architectures with strong pretrained backbones (e.g., ViT, DINOv2) will support flexible multimodal fusion and robust representation learning, especially under missing modality conditions. A BiLSTM-CRF sequence labeling framework [57] will be implemented as a baseline, with future architectural improvements planned as the project progresses.

While topic shifts are inherently semantic, nonverbal behaviors, such as posture shifts, head orientation changes, and group coordination can signal attempts to steer conversation flow. This study focuses on detecting such topic shift attempts through observable social cues, with future work exploring cross-modal semantic representations where language and audio signals aid inference from nonverbal modalities when audio is unavailable at test time.

3.5 Handling Missing Modalities and Temporal Disruptions

This research considers two types of modality missingness. The first is modality absence at inference, where certain modalities like high-quality audio are present during training but unavailable at test time, common in real-world deployments like topic change detection under sensor limitations. The second is temporal degradation, where occlusion, noise, or sensor failures cause intermittent gaps in nonverbal signals such as posture or gestures.

To address these challenges, two main strategies will be explored. One applies modality dropout during training to encourage robustness and modality-invariant representations [26, 30, 51]. The other uses full modality supervision during training to compensate for missing inputs at inference, through techniques such as cross-modal distillation [20, 56], modality hallucination [23, 59], and parameter-efficient adaptation of pretrained models [40].

Additionally, temporal smoothness priors will be applied to estimate occluded cues like head orientation and posture, leveraging sequential dynamics as proposed in [49].

4 Preliminary Result and Remaining Work

To examine how modality affects annotation quality, I analyzed annotators' laughter labeling decisions under three conditions—video only, audio only, and audiovisual (AV)—using the Conflab dataset [42]. Inter-annotator agreement varied across conditions, echoing prior findings [39]. While AV and audio-only labels were mostly consistent, discrepancies arose from timing mismatches or ambiguous durations. Visual cues sometimes caused source misattribution, while gestures compensated for weak audio signals. Video-only annotations showed the greatest divergence, often reflecting "perceived laughter" inferred from visible behaviors.

As an exploratory step, I computed synchrony metrics to test whether laughter aligns more strongly within conversational groups than across them. Within-group pairs consistently exhibited higher synchrony, supporting the role of laughter as an in-group bonding cue [15]. Cluster-based permutation tests further revealed that real laughter events form denser temporal clusters than chance, with group size influencing clustering patterns [11, 18].

Preliminary correlations between upper-body keypoints and laughter episodes indicate that certain movements are strongly associated with laughter, despite individual expressive differences. Synchrony in posture and gestures also appears linked to co-laughter dynamics, warranting further investigation.

These findings provide a foundation for detecting laughter through group-level nonverbal cues and offer insights into its role in multiparty interactions. They also highlight how modality-dependent annotation biases can inform the design of contrastive learning strategies, laying the groundwork for jointly modeling topic shifts and laughter.

5 Research Contributions

This research develops methods for estimating conversational events in real-world multiparty settings, particularly when speech signals are missing or unreliable. By leveraging nonverbal cues like movement, posture, gestures, and orientation, it focuses on modeling group-level coordination to support robust inference of interaction dynamics. The approach produces temporally grounded, context-aware representations that bridge low-level behavioral signals and higher-level constructs such as social involvement, topic shifts, and communicative intent, enabling downstream modeling of complex social phenomena.

Acknowledgments

This paper is funded by ERC Consolidator Grant NEON: 1010891.

References

[1] Ahmed Amer, Chirag Bhuvaneshwara, Gowtham K Addluri, Mohammed M Shaik, Vedant Bonde, and Philipp Müller. 2023. Backchannel detection and agreement estimation from video with transformer networks. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[2] Jo-Anne Bachorowski, Moria J Smoski, and Michael J Owren. 2001. The acoustic features of human laughter. *The journal of the Acoustical Society of America* 110, 3 (2001), 1581–1597.

[3] Hugo Bohy, Kevin El Haddad, and Thierry Dutoit. 2022. A new perspective on smiling and laughter detection: Intensity levels matter. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[4] Erik Brockbank, Tobias Gerstenberg, Judith E Fan, and Robert Hawkins. 2025. How do we get to know someone? Diagnostic questions for inferring personal traits. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 47.

[5] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. 2018. The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing* 12, 1 (2018), 113–130.

[6] Justine Cassell, Yukiko I Nakano, Timothy W Bickmore, Candace L Sidner, and Charles Rich. 2001. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 114–123.

[7] G Charan, G Sivateja, M Karthik Sarma, and Sabyasachi Kumar. 2025. Unveiling the challenges of speech recognition in noisy environments: A comprehensive review of issues and solutions. *Challenges in Information, Communication and Computing Technology* (2025), 407–412.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[9] Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science* 13, 2 (1989), 259–294.

[10] Marco Cristani, Anna Pesarin, Alessandro Vinciarelli, Marco Crocco, and Vittorio Murino. 2011. Look at who's talking: Voice activity detection by automated gesture analysis. In *International Joint Conference on Ambient Intelligence*. Springer, 72–80.

[11] Rick Dale, Gregory A Bryant, Joseph H Manson, and Matthew M Gervais. 2020. Body synchrony in triadic interaction. *Royal Society open science* 7, 9 (2020), 200095.

[12] Thomas Drugman, Yannis Stylianou, Yusuke Kida, and Masami Akamine. 2015. Voice activity detection: Merging source and filter-based information. *IEEE Signal Processing Letters* 23, 2 (2015), 252–256.

[13] Ekin Gedik, Laura Cabrera Quiros, and Hayley Hung. 2019. No-Audio Multimodal Speech Detection Task at MediaEval 2019.. In *MediaEval*.

[14] Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman. 2021. Robust Laughter Detection in Noisy Environments.. In *Interspeech*, Vol. 2021. 2481–2485.

[15] Emer Gilmartin, Francesca Bonin, Carl Vogel, and Nick Campbell. 2013. Laugher and topic transition in multiparty conversation. In *Proceedings of the SIGDIAL 2013 Conference*. 304–308.

[16] Phillip Glenn. 2003. *Laughter in interaction*. Vol. 18. Cambridge University Press.

[17] Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* 23, 1 (1997), 33–64.

[18] Elizabeth Holt. 2010. The last laugh: Shared laughter and topic termination. *Journal of Pragmatics* 42, 6 (2010), 1513–1525.

[19] Hayley Hung, Ekin Gedik, and Laura Cabrera Quiros. 2019. Complex conversational scene analysis using wearable sensors. In *Multimodal Behavior Analysis in the Wild*. Elsevier, 225–245.

[20] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. 2024. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16006–16015.

[21] Arjun Jain, Jonathan Tompson, Mykhaylo Andriluka, Graham W Taylor, and Christoph Bregler. 2013. Learning human pose estimation features with convolutional networks. *arXiv preprint arXiv:1312.7302* (2013).

[22] Gail Jefferson. 1984. 14. On the organization of laughter in talk about troubles. (1984).

[23] Xilin Jiang, Junkai Wu, Vishal Choudhari, and Nima Mesgarani. 2025. Bridging Ears and Eyes: Analyzing Audio and Visual Large Language Models to Humans in Visible Sound Recognition and Reducing Their Sensory Gap via Cross-Modal Distillation. *arXiv preprint arXiv:2505.06803* (2025).

[24] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

[25] Mary Tai Knox and Nikki Mirghafori. 2007. Automatic laughter detection using neural networks.. In *Interspeech*. Citeseer, 2973–2976.

[26] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131* (2022).

[27] Bogdan Ludusan and Barbara Schuppler. 2022. To laugh or not to laugh? The use of laughter to mark discourse structure. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 76–82.

[28] David McNeill. 2000. *Language and gesture*. Vol. 2. Cambridge University Press.

[29] Arnab Mondal and AP Prathosh. 2020. RespVAD: Voice activity detection via video-extracted respiration patterns. *IEEE Sensors Letters* 4, 9 (2020), 1–4.

[30] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. 2015. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2015), 1692–1706.

[31] Radoslaw Niewiadomski, Maurizio Mancini, Giovanna Varni, Gualtiero Volpe, and Antonio Camurri. 2015. Automated laughter detection from full-body movements. *IEEE Transactions on Human-Machine Systems* 46, 1 (2015), 113–123.

[32] Hiroaki Noguchi and Yasuharu Den. 1998. Prosody-based detection of the context of backchannel responses.. In *ICSLP*.

[33] Nick O'Donnell-Trujillo and Katherine Adams. 1983. Heheh in conversation: Some coordinating accomplishments of laughter. *Western Journal of Communication (includes Communication Reports)* 47, 2 (1983), 175–191.

[34] Kazuhiro Otsuka and Masahiro Tsumori. 2020. Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks. *IEEE Access* 8 (2020), 217169–217195.

[35] Alex Pentland. 2010. *Honest signals: how they shape our world*. MIT press.

[36] Stavros Petridis, Maelle Leveque, and Maja Pantic. 2013. Audiovisual detection of laughter in human-machine interaction. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 129–134.

[37] Wim Pouw, Shannon Proksch, Linda Drijvers, Marco Gamba, Judith Holler, Christopher Kello, Rebecca S Schaefer, and Geraint A Wiggins. 2021. Multilevel rhythms in multimodal communication. *Philosophical Transactions of the Royal Society B* 376, 1835 (2021), 20200334.

[38] Francis Quek, David McNeill, Robert Bryll, Cemil Kirbas, Hasan Arslan, Karl E McCullough, Nobuhiro Furuyama, and Rashid Ansari. 2000. Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, Vol. 2. IEEE, 247–254.

[39] Jose David Vargas Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. 2023. Impact of annotation modality on label quality and model performance in the automatic assessment of laughter in-the-wild. *IEEE Transactions on Affective Computing* 15, 2 (2023), 519–534.

[40] Abduljalil Radman and Jorma Laaksonen. 2024. AV-PEA: Parameter-Efficient Adapter for Audio-Visual Multimodal Learning.. In *VISIGRAPP (2): VISAPP*. 730–737.

[41] Chirag Raman, Hayley Hung, and Marco Loog. 2021. Social processes: Self-supervised forecasting of nonverbal cues in social conversations. (2021).

[42] Chirag Raman, Jose Vargas Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. 2022. Conflab: A data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild. *Advances in Neural Information Processing Systems* 35 (2022), 23701–23715.

[43] Albert E Scheflen. 1964. The significance of posture in communication systems. *Psychiatry* 27, 4 (1964), 316–331.

[44] Helen Schmidt, Claire Augusta Bergey, Changyi Zhou, Chelsea Helion, and Robert D Hawkins. 2025. Dynamics of topic exploration in conversation. *Annotation* 40 (2025), 38.

[45] Daniel Schulman and Timothy Bickmore. 2011. Posture, relationship, and discourse structure: Models of nonverbal behavior for long-term interaction. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. Springer, 106–112.

[46] Muhammad Shahid, Cigdem Beyan, and Vittorio Murino. 2021. S-VVAD: Visual voice activity detection by motion segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2332–2341.

[47] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication* 32, 1-2 (2000), 127–154.

[48] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE signal processing letters* 6, 1 (1999), 1–3.

[49] Stephanie Tan, David MJ Tax, and Hayley Hung. 2022. Head and Body Orientation Estimation with Sparse Weak Labels in Free Standing Conversational Settings. In *Understanding Social Behavior in Dyadic and Small Group Interactions*. PMLR, 179–203.

[50] Fei Tao and Carlos Busso. 2019. End-to-end audiovisual speech activity detection with bimodal recurrent neural models. *Speech Communication* 113 (2019), 25–35.

[51] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, Vol. 2019. 6558.

[52] Bekir Berker Türker, Engin Erzin, Yücel Yemez, and T Metin Sezgin. 2018. Audio-Visual Prediction of Head-Nod and Turn-Taking Events in Dyadic Interactions.. In *Interspeech*. 1741–1745.

[53] Jose Vargas and Hayley Hung. 2019. CNNs and fisher vectors for no-audio multimodal speech detection. In *Working Notes Proceedings of the MediaEval 2019 Workshop*. 11–13.

[54] Jose Vargas-Quiros, Laura Cabrera-Quiros, and Hayley Hung. 2022. No-audio speaking status detection in crowded settings via visual pose-based filtering and wearable acceleration. *arXiv preprint arXiv:2211.00549* (2022).

[55] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.

[56] Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 216–226.

[57] Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. *arXiv preprint arXiv:1808.09147* (2018).

[58] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics* 32, 8 (2000), 1177–1207.

[59] Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. 2024. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*.

[60] Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. Tiage: A benchmark for topic-shift aware dialog modeling. *arXiv preprint arXiv:2109.04562* (2021).

[61] Victor H Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*. 567–578.