Unveiling Subtle Cues: Backchannel Detection Using Temporal Multimodal Attention Networks

Kangzhong Wang
cskwang@comp.polyu.edu.hk
Department of Computing
Hong Kong Polytechnic University
Hong Kong, China

MK Michael Cheung
mk.cheung@polyu.edu.hk
School of Optometry
Hong Kong Polytechnic University
Hong Kong, China

Youqian Zhang
you-qian.zhang@polyu.edu.hk
Department of Computing
Hong Kong Polytechnic University
Hong Kong, China

Chunxi Yang
chunxyang@polyu.edu.hk
Department of Rehabilitation Sciences
Hong Kong Polytechnic University
Hong Kong, China

Peter Q. Chen
qichen@polyu.edu.hk
Department of Computing
Hong Kong Polytechnic University
Hong Kong, China

Eugene Yujun Fu*
eugene.fu@polyu.edu.hk
Department of Rehabilitation Sciences
Hong Kong Polytechnic University
Hong Kong, China

Grace Ngai
csgngai@comp.polyu.edu.hk
Department of Computing
Hong Kong Polytechnic University
Hong Kong, China

ABSTRACT

Automatic detection of backchannel has great potential to enhance artificial mediators, which indicate listeners' attention and agreement in human communication. It is often expressed by subtle non-verbal cues that occur briefly and sparsely. Focusing on identifying and locating these subtle cues (i.e., their occurrence moment and the involved body parts), this paper proposes a novel approach for backchannel detection. In particular, our model utilizes temporal- and modality-attention modules to determine and lead the model to pay more attention to both the indicative moment and the accompanying body parts at that specific time. It achieves an accuracy of 68.6% on the testing set in *MultiMediate'23* backchannel detection challenge, outperforming the counterparts. Furthermore, we conducted an ablation study to thoroughly understand the contributions of our model. This study underscores the effectiveness of our selection of modality inputs and the importance of the two attention modules in our model.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

backchannel detection; attention models; visual cues

1 INTRODUCTION

Backchannels often occur in human-human interaction, expressed by subtle verbal and non-verbal cues [6]. These subtle cues serve as indicators of a listener's attention and agreement without causing interruptions, yet often difficult to discern [21]. The ability to accurately detect backchannel occurrence has profound implications across a wide range of domains, such as artificial mediation [9, 12, 22]. This paper proposes a novel approach for the *MultiMediate* backchannel detection challenge [15, 16].

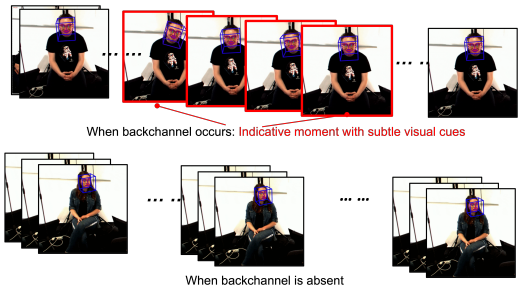

Figure 1: Backchannel is often conveyed via subtle non-verbal cues (e.g., head nods). The brief and sparseness of these cues challenge the detection. Leading the model to the indicative moments and body parts (i.e., when and where the indicative cues occur) contributes in better detection.

Extensive research has been conducted on applying machine learning techniques for backchannel detection, which leverage different types of modalities [4, 14, 20]. The audio signal is one of the primary modalities, from which verbal cues can be extracted for backchannel detection [13]. Though, it may not be available in certain situations, such as when the audience has the camera on while muting themselves to avoid interrupting the speakers in video conferences. This paper particularly explores subtle visual cues from video signals for backchannel detection.

Recent studies [1, 16] have also highlighted the significance of visual cues in backchannel detection. However, the brief and sparse occurrence of these cues challenges the existing methods. They are limited in the capability to identify either when (at which frames) or where (accompanied by which body parts) the indicative visual cues occur, thus are easily misled by irrelevant information. Prior studies

*Corresponding author. E-mail address: eugene.fu@polyu.edu.hk (E.Y. Fu)

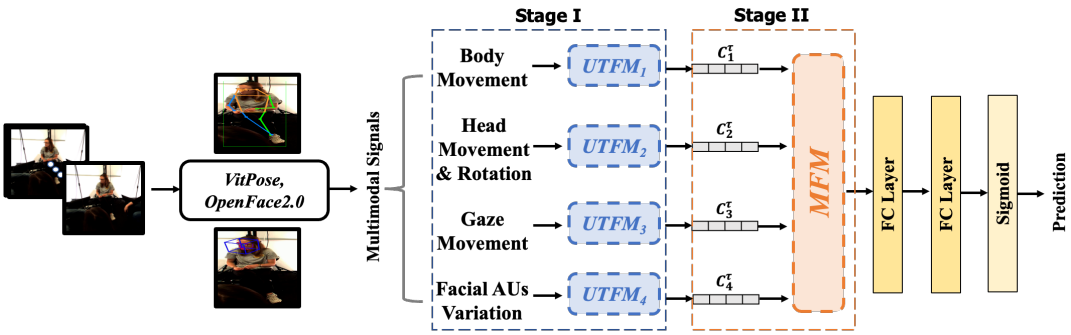

Figure 2: Overview of our proposed method: Multimodalities are extracted from the raw video frames, which are then further processed and encoded by our proposed two-stage attention-based model for backchannel detection.

stressed the incorporation of attention mechanisms to determine the salience of the inputs and enable a model to attend to the most salient ones [2, 7, 19, 25]. Inspired by that, this paper proposes a neural network model with attention modules, namely Temporal Multimodal Attention Network (TMAN), to tackle the challenge.

In particular, TMAN incorporates two attention-based modules: the Unimodal Temporal Fusion Module (UTFM), which locates salient frames and encodes temporal information within each visual modality (e.g., head and gaze); and the Multimodal Fusion Module (MFM), which identifies salient modalities and encode efficient cross-modal features. They enable the model to determine and lead the model to focus more on both the indicative frames and the accompanying body parts at those frames.

We evaluated the proposed model on the backchannel detection dataset in the MultiMediate'23 challenge [15] that was originally introduced in MultiMediate'22 [16]. The experimental results demonstrate the significance of attending to the salient moments and visual body parts, and the effectiveness of attention modules. Specifically, an accuracy of 68.6% is achieved when applying TMAN on the test set in the challenge, outperforming all the competitors.

2 METHODOLOGY

TMAN extracts multiple body modalities from the given frames, and then encodes the temporal and cross-modality features with attention modules for backchannel detection (Fig. 2).

2.1 Multi-modalities

We employ VitPose [27] for body pose estimation and Openface 2.0 [3] for the extraction of facial- and head-related features (Fig. 2). In particular, features from four modalities are extracted as follows.

Body pose: the coordinates of certain key joints extracted from each video frame. Since all participants were seated in the experiments, we specifically focus on the 11 upper body points: a pair of eyes, ears, hands, elbows, and shoulders, and a single nose point.

Head pose: the position and orientation of the head. We extract six head-related features from the detected face in each video frame. These include three-dimensional position vector (i.e., x, y, z) that represents the head's relative position to the camera in millimeters, as well as three-dimensional orientation vector that measures the head's rotation in radians with respect to the camera.

Eye gaze: the gaze direction of the eyes. In particular, eight gaze related features are extracted for each frame. These comprise three-dimensional position vectors (i.e., x, y, z) indicating the gazing position for each eye, as well as two averaged eye gaze directions in radians for both eyes.

Facial action units (AUs): the representative facial AUs of each individual. We specifically extract the intensity features from 17 representative AUs selected by the OpenFace 2.0 toolkit [3].

Furthermore, previous work demonstrated the significance of movement patterns in group behavior analysis [8]. We hypothesize that the raw modality features themselves do not provide adequate information associated with the movement nature of these signals. We thus propose to pre-process the input data by computing the variations between consecutive frames for each modality. This results in four additional modality inputs, namely, *Body Movement*, *Head Movement & Rotation*, *Gaze Movement*, and *Facial Action Units Variation*. The investigation of using raw or variation inputs, allows us to understand the data and our model more effectively.

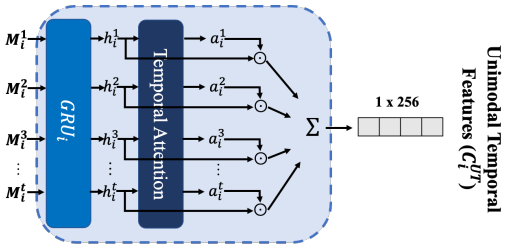

Figure 3: Unimodal Temporal Fusion Module (UTFM).

2.2 Unimodal Temporal Fusion Module (UTFM)

In the first stage of the proposed model, we encode representative temporal features for each modality respectively with an attention-based module, named Unimodal Temporal Fusion Module (UTFM). Fig. 3 shows the structure of the proposed UTFM. For a particular modality (e.g., head movement), it initially applies a standard Gated Recurrent Units (GRU) to encode the latent features (h^t) for each frame. It then encodes the temporal information across all the frames in the given time window. Our choice of GRU is informed by

studies that underscore its effectiveness and robustness in encoding temporal sequences across various tasks [5, 24]. Since backchannel cues often occur briefly and sparsely, it is important to identify and make the model focus on the most likely frames where backchannel cues are prone to occur. To this end, UTFM employs an attention layer to compute the saliency weights a^t of the frames, based on their latent features (h^t). The overall temporal representation of the target modality is attained by computing the weighted summation of the latent features across all the frames. Frames with higher attention weight contribute more to the temporal representation, i.e., assisting the model to attend to the most salient frames.

Mathematically, UTFM attention computation and modelling are achieved as follows:

$$h_i^t = GRU_i(M_i^t) \tag{1}$$

$$a_i^t = \text{Softmax}(\text{ReLU}(h_i^t W_t + b_t)) \tag{2}$$

$$c_i^{UT} = \sum_{t=1}^{T} a_i^t h_i^t \tag{3}$$

where M_t^i represents the input at the timestep (frame) t of modality i, h_T^i denotes the latent features encoded by GRU. W_t and b_t are trainable parameters for attention computation, a_i^t denotes the attention weight of the frame t for modality i, undergo ReLU and softmax function. Lastly, c_i^{UT} is the attention-weighted sum of all latent features for the modality i, which serves as the overall temporal features of that modality for further modelling.

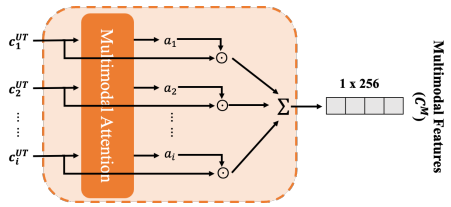

Figure 4: Multimodal Fusion Module (MFM).

2.3 Multimodal Fusion Module

We then encode the cross-modality features for backchannel detection based on the temporal representations extracted in the previous stage. As shown in Fig. 4, we apply another attention-based module, Multimodal Fusion Module (MFM), to discern the significance of different modalities and guide the model to concentrate on the most crucial ones. MFM computes the attention weights of the modalities based on their temporal representations. Attention mechanisms incorporated with sigmoid and tanh functions showed their particular effectiveness for multimodal data [10, 26]. Inspired by that, we tailored the following attention computation to accommodate the unique characteristics of our task.

$$a_i = \text{Sigmoid}(\text{Tanh}(c_i^{UT} W_1 + b_1) W_2 + b_2) \tag{4}$$

$$c^M = \sum_{i=1}^{N} a_i c_i^{UT} \tag{5}$$

where W_1, W_2, b_1, and b_2 denote trainable parameters of the attention mechanism. a_i is the computed attention weight for the modality i. Notably, this attention computation yielded better performance in our pilot experiments than other commonly adopted methods [1, 11, 18], consistent with observations from [26]. Lastly, c^{MF} is the integrated representation of all modality features, reflecting the weighted contribution of each modality to the final multimodal feature vector.

3 EXPERIMENTS

3.1 Dataset

We evaluated the proposed model on the backchannel detection dataset in MultiMediate'23 challenge [16]. This dataset is generated from a public group conversation dataset MPIIGroupInteraction [17], which consists of 24 group conversations last about 20 minutes captured in a laboratory setup. The original MPIIGroupInteraction videos were cut into multiple non-overlapping 10-second windows forming the data samples in the backchannel dataset. The ground truth of each sample was annotated based on the existence of backchannel behavior at the end of the 10-second window. The task is thereby framed as a binary classification task: the presence or absence of a backchannel. In total, the dataset has 6,716 instances for training, 2,854 for validation, and 4,898 for testing.

3.2 Experiment Settings

Previous work [23] and our pilot study observe the significance of the last-second context window for backchannel detection. We thus apply this setting in our experiments: inputting the last second of the context window to our model for backchannel detection.

Our model is implemented in PyTorch and trained on an NVIDIA GTX 3090 GPU. The model undergoes 20 training epochs, with a batch size of 64. For optimization, we utilize the Adam optimizer with an initial learning rate of 0.001 and a weight decay parameter set to 0.0001. The architecture of our GRU networks comprises two layers, each containing 128 neurons, with a dropout rate of 0.3. We use binary cross-entropy with logits loss as our loss function.

4 RESULTS

We train our model on the training set and conduct a comprehensive ablation study on the validation set. And finally, we summarize our testing set performance and compare it with the counterparts.

Table 1: The evaluation of the proposed attention modules

Model	Validation Accuracy
Base-GRU	0.682
UTFM Only	0.693
MFM Only	0.711
TMAN	**0.741**

4.1 Ablation Study

We first study the contribution of the two proposed attention modules (UTFM and MFM), and the effectiveness of integrating them. Table 1 depicts the performance yielded by a common GRU-based neural netowkr model (*Base-GRU*), the completed TMAN model (*TMAN*), TMAN leaves out the MFM (*UTFM only*), and TMAN leaves out the UTFM (*MFM only*). It is worth noting that both the UTFM

and MFM can contribute to performance improvement alone, beating the *Base-GRU* model. Notably, the model performance improved significantly when a completed TMAN is employed, achieving the accuracy of 74.1%. The results demonstrate the effectiveness of the proposed attention modules, and underline the significance of leading model to attend to the salient moments (frames) and modalities (body parts).

Table 2: The impacts of different modalities

Model	Modality				Acc.
	Body Move.	Head Move. & Rota.	Eye Gaze Move.	Facial AUs Var.	
TMAN	✓	✓			0.615
	✓		✓		0.621
	✓			✓	0.613
		✓	✓		0.664
		✓		✓	0.657
			✓	✓	0.648
	✓	✓	✓		0.701
	✓		✓	✓	0.693
		✓	✓	✓	0.722
	✓	✓	✓	✓	**0.741**

In addition to the attention weights, our further experiment investigate more on the impacts of different modalities. Table 2 summarizes the performance of applying different combinations of the four modalities. It is worth noting that the model incorporate all the modalities outperforms the others revealing that the cues from head, gaze, body, and face are all useful for backchannel detection. Head and gaze modalities are more indicative than the others though.

Furthermore, we hypothesize that computing and using the variations between consecutive frames enables more effective modelling, which delivers the movement patterns to the model more directly. To validate the hypothesis, we also explore the impact of the pre-processing of the input signals in our experiment. Specifically, we evaluate impact of two pre-processing methods to get the feature variations: (1) computing the absolute difference between the consecutive frames; and (2) computing the subtraction (the raw difference value) between consecutive frames. We evaluate the TMAN model incorporating theset two pre-processing methods and compare the performance with the model that process the raw input signals.

Table 3: The impacts of data pre-processing

Input	Validation Accuracy
Raw	0.642
Absolute Diff.	0.702
Subtraction	**0.741**

The experimental results (Table 3) emphasize the significance of the pre-processing. Particularly, using subtraction to extract motion dynamics yields the highest validation accuracy. This implies that the subtraction process effectively captures critical data patterns (i.e., the variation magnitude and orientation), benefiting the modelling. These results jointly signify the effectiveness of the proposed TMAN, especially when paired with the subtraction pre-processing method with its superior ability to utilize multimodal data.

Table 4: Test results comparison

Method	Test Acc.
TMAN (Ours)	**0.686**
Amer et al.	0.664
Anony. 1	0.658
Ma et al.	0.656
Sharma et al.	0.621
Baseline '22 (Head + Pose)	0.596
Baseline '22 (All)	0.592
Baseline '22 (Trivial)	0.500

4.2 Testing Performance in the Challenge

Following the ablation study, we apply the completed TMAN with full modalities and subtraction pre-processing on the testing set for evaluation and comparsion. The results (Table 4) highlight the impressive performance of our model, with an accuracy of 68.6%. This is attributed to its capability to attend to both the salient moments and modalities, exceeding all competitors that solely focus on modality attention [1] or lack any attention [16, 23].

5 CONCLUSION

In this study, we proposed a novel model Tempo- ral Multimodal Attention Network (TMAN) for backchannel detection. We conducted comprehensive experiments to extensively understand the contribution of different modules, modalities, and pre-processing methods. The experimental results show the effectiveness of the proposed method, especially with the capability in determining and guiding model to focus on the salient moments and modalities. In the future, we will keep improving the model such as exploring different network backbones, and pre-processing methods. Furthermore, we will evaluate our method on different datasets with diverse scenarios, such as remote video conferences.

ACKNOWLEDGMENTS

This work was supported, in part, by the Hong Kong Polytechnic University under Grant P0039489, and the Hong Kong Research Grant Council under Grant 15600219.

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. *arXiv preprint arXiv:2306.01656* (2023).

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[4] Elisabetta Bevacqua, Sathish Pammi, Sylwia Julia Hyniewska, Marc Schröder, and Catherine Pelachaud. 2010. Multimodal backchannels for embodied conversational agents. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*. Springer, 194–200.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[6] Starkey Duncan and Donald W Fiske. 2015. *Face-to-face interaction: Research, methods, and theory.* Routledge.

[7] Eugene Yujun Fu, Grace Ngai, Hong Va Leong, Stephen CF Chan, and Daniel TL Shek. 2023. Using attention-based neural networks for predicting student learning outcomes in service-learning. *Education and Information Technologies* (2023), 1–27.

[8] Eugene Yujun Fu and Michael W Ngai. 2021. Using motion histories for eye contact detection in multiperson group conversations. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4873–4877.

[9] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. 2005. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *Ninth European Conference on Speech Communication and Technology*.

[10] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Hybrid attention based multimodal network for spoken language classification. In *Proceedings of the Conference. association for Computational Linguistics. meeting*, Vol. 2018. NIH Public Access, 2379.

[11] Dichao Hu. 2020. An introductory survey on attention mechanisms in NLP problems. In *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2*. Springer, 432–448.

[12] Peerumporn Jiranantanagorn, Haifeng Shen, Robert Goodwin, and Kung-Keat Teoh. 2015. Classense: a mobile digital backchannel system for monitoring class morale. *International Journal of Learning and Teaching* 1, 2 (2015), 161–167.

[13] Robert M Krauss, Connie M Garlock, Peter D Bricker, and Lee E McMahon. 1977. The role of audible and visible back-channel responses in interpersonal communication. *Journal of personality and social psychology* 35, 7 (1977), 523.

[14] Jing Liu, Mitja Nikolaus, Kübra Bodur, and Abdellah Fourtassi. 2022. Predicting backchannel signaling in child-caregiver multimodal conversations. In *Companion publication of the 2022 international conference on multimodal interaction*. 196–200.

[15] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2023. MultiMediate '23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions. In *Proceedings of the 31st ACM International Conference on Multimedia*. https://doi.org/10.1145/3581783.3613851

[16] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7109–7114.

[17] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *23rd International Conference on Intelligent User Interfaces*. 153–164.

[18] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems* 34 (2021), 14200–14213.

[19] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing* 452 (2021), 48–62.

[20] Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, Jeez! or uh-huh? A listener-aware Backchannel predictor on ASR transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8064–8068.

[21] Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*. Springer, 146–158.

[22] Sonya Rajan, Scotty D Craig, Barry Gholson, Natalie K Person, Arthur C Graesser, and Tutoring Research Group. 2001. AutoTutor: Incorporating back-channel feedback and other human-like conversational behaviors into an intelligent tutoring system. *International Journal of Speech Technology* 4 (2001), 117–126.

[23] Garima Sharma, Kalin Stefanov, Abhinav Dhall, and Jianfei Cai. 2022. Graph-based group modelling for backchannel detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7190–7194.

[24] Apeksha Shewalkar, Deepika Nyavanandi, and Simone A Ludwig. 2019. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research* 9, 4 (2019), 235–245.

[25] Wai Cheong Tam, Eugene Yujun Fu, Jiajia Li, Richard Peacock, Paul Reneke, Grace Ngai, Hong Va Leong, Thomas Cleary, and Michael Xuelin Huang. 2023. Real-time flashover prediction model for multi-compartment building structures using attention based recurrent neural networks. *Expert Systems with Applications* 223 (2023), 119899.

[26] Yao Wan, Jingdong Shu, Yulei Sui, Guandong Xu, Zhou Zhao, Jian Wu, and Philip Yu. 2019. Multi-modal attention network learning for semantic source code retrieval. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 13–25.

[27] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems* 35 (2022), 38571–38584.