CCDb+: Enhanced Annotations and Multi-Modal Benchmark for Natural Dyadic Conversations

Yang Deng
dengy29@cardiff.ac.uk
School of Computer Science and Informatics
Cardiff University
Cardiff, UK

Yu-Kun Lai*
LaiY4@cardiff.ac.uk
School of Computer Science and Informatics
Cardiff University
Cardiff, UK

Paul L. Rosin
RosinPL@cardiff.ac.uk
School of Computer Science and Informatics
Cardiff University
Cardiff, UK

Abstract

Backchannel signals play a critical role in social interaction, expressing attentiveness, agreement, and emotion in both human and human-agent conversations. However, few multi-modal databases exist in this area due to the complexity of categorisation and the high cost of precise timing, especially in naturalistic dyadic conversations. To address these challenges, we introduce CCDb+ (Cardiff Conversation Database +) an enhanced version of CCDb, with 25 newly annotated conversations and corrections to 14 previously annotated conversations, along with thorough consistency checks to ensure annotation reliability. Additionally, we propose a multi-modal process for backchannel detection as a baseline, showing that both visual and acoustic cues contribute significantly to understanding backchannel behaviour. Recognising that backchannel signals often intersect with other social cues, we introduce several detection sub-tasks—such as smile, nodding, and agreement—with baseline results for each. Finally, we demonstrate multi-modal paradigms for nuanced signals like nodding and thinking. The database and associated annotations are publicly available at https://huggingface.co/datasets/CardiffVisualComputing/CCDb.

CCS Concepts

• **Computing methodologies** → **Activity recognition and understanding**; **Supervised learning by classification**; • **Information systems** → *Multimedia information systems*; • **General and reference** → **Evaluation**.

Keywords

Multimedia; Social signals; Dyadic Conversation; Database; Backchannel; Multi-modal; Benchmark; Faces

ACM Reference Format:

Yang Deng, Yu-Kun Lai, and Paul L. Rosin. 2025. CCDb+: Enhanced Annotations and Multi-Modal Benchmark for Natural Dyadic Conversations. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755333

*Corresponding author

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755333

1 Introduction

Dyadic conversation has long been a fundamental aspect of real-world social exchanges, and especially in recent years, when these conversations have expanded beyond human-to-human communication to include engagements between humans and social interaction agents (SIA), and even between SIAs themselves. Backchannel is defined as gestural and vocal signals of active listenership in conversation [37], which significantly determine the quality of these conversations, as they express attentiveness, rapport, agreement, and emotion in interactions [29, 32, 76]. Additionally, backchannel plays an important role in turn-taking [57] and other conversational dynamics [39, 53, 74].

Early backchannel research primarily focused on vocal signals, examining non-verbal cues such as prosodic elements [74, 76]. Other studies explored verbal signals for backchanneling, highlighting specific expressions and short utterances [9, 49, 51, 63, 64, 71]. Most of these studies relied on speech-based, single-modality datasets [41, 44, 74]. However, the majority of these datasets were not made publicly available, with one notable exception: the Switchboard Dialog Act Corpus (SwDA) [13], which offers a valuable resource of telephone-style dyadic conversations for backchannel research.

Backchannel signals are inherently complex, and while early research predominantly focused on vocal signals, some studies have highlighted the importance of integrating both visual and vocal data to analyse social interactions, particularly human emotions. Emotions are primarily conveyed through facial expressions, voice, and language during everyday communication [14]. This insight has inspired backchannel research, revealing that backchanneling involves not only vocal responses but also rich visual cues such as head gestures and facial expressions [2, 5]. In particular, when interlocutors can see each other, a substantial portion of backchannel signals is communicated visually [28]. Building on this understanding, advancements in hardware and storage technologies have enabled the release of several multi-modal datasets, which incorporate both vocal and visual cues [6, 8, 15, 35, 41, 47, 48, 56–58, 65, 68]. These datasets allow for a more comprehensive analysis of social interactions, providing richer insights into backchannel behaviour by integrating multiple modalities. However, despite the increasing number of multi-modal databases, several limitations remain. Firstly, many are based on structured, story-telling interactions,

which may not fully reflect the spontaneity and complexity of natural dyadic exchanges. Secondly, a significant proportion of these datasets focus on interactions between children, limiting their relevance to adult social dynamics or broader demographic contexts. Additionally, the majority of these databases have not been made generally available, hindering the ability of researchers to replicate or build upon previous work. This lack of open access poses a significant barrier to the advancement of research in backchannel behaviour, as comparisons and validations across datasets are severely restricted.

To address these issues, we are extending CCDb to CCDb+, a multi-modal database of natural dyadic conversations that includes expanded and refined annotations for backchannel behaviours across diverse age groups. The original CCDb, previously released in video [4] and 3D sequence versions [46], provided a foundational audiovisual corpus with annotations for conversational facial expressions, head motions, speaker activity, and verbal/non-verbal utterances. CCDb+ builds on this foundation by adding 25 new conversations, refining 14 previously annotated ones, and applying a rigorous three-stage review process to enhance annotation reliability. This updated version integrates new insights on backchannel behaviour [7, 38, 55] to ensure consistency and accuracy across all annotations.

To comprehensively evaluate the potential of this database, we conducted baseline assessment experiments based on multi-modal information for backchannel and sub-signal detection tasks. Our focus on backchannel detection serves two main purposes. First, developing detection algorithms can streamline future annotation processes, allowing for automated initial detection followed by human validation, significantly saving time and effort. Second, detecting listener backchannel enables social interaction agents to generate more appropriate and positive responses in dyadic conversations, enhancing the quality of human-agent interactions.

While our current work centres on backchannel detection, CCDb has supported a range of studies over the past decade, showcasing its adaptability across various research domains. Numerous studies have leveraged CCDb for recognising visual markers of spontaneous head gestures [67] and advancing the automation of human listening behavior in both human-human and human-agent interactions [17, 22, 23, 59, 60]. Additionally, CCDb has been instrumental in turn-taking prediction in conversations [40], and in detecting and classifying social communicative events (SCE) [16].

Researchers have further enriched CCDb with additional specialised annotations to tailor it to their unique study objectives. For instance, engagement has been annotated to understand conversational involvement [30, 31], head gestures have been detailed with greater specificity [70], and smiles and laughter have been annotated across three levels of intensity [18, 19]. We encourage the research community to build upon CCDb+ in similar ways, exploring new dimensions of multi-modal interaction and, where applicable, adding custom annotations suited to their specific research topics.

The major contributions of this paper are as follows.

- We release a multi-modal database of dyadic conversations, comprising 44 fully annotated interactions with video, audio and transcript.
- We propose a streamlined approach that integrates multi-modal information, combining both acoustic and visual cues, for enhanced backchannel detection. Our evaluation demonstrates that acoustic cues are particularly effective for backchannel detection, with their performance further enhanced when integrated with visual cues. The source code is publicly available at https://github.com/hsdy1125/Backchannel-Detection-CCDb.
- We propose and validate multi-modal paradigms for the detection of thinking and nodding signals.

2 Related Work

Backchannel detection is a research area of widespread interest. However, current research is constrained by two significant gaps. First, there is a notable lack of publicly available, comprehensive multi-modal datasets for backchannel studies, particularly those capturing natural dyadic interactions. Second, despite advances in detection methods, many studies still primarily rely on a single publicly available group conversation dataset, which introduces certain biases.

To address these gaps, we review existing dyadic conversation datasets with backchannel annotations, examining their strengths and limitations in representing natural backchannel behaviours. Following this, we discuss current methodological approaches to backchannel detection, highlighting the need for models that can adapt to the nuances of dyadic interactions.

2.1 Current Dyadic Conversation Databases with Backchannel Annotation

Several dyadic conversation databases have been utilised in backchannel research, especially those that integrate multi-modal data (audio, visual, and sometimes transcripts). These datasets enable a richer analysis of backchannel behaviours, which are expressed through vocal signals, facial expressions, and gestures.

In Table 1 we list recent publicly available datasets that were used in Backchannel research. However, while the availability of these datasets has grown, several limitations remain. For instance, while some databases have been released, a considerable number remain inaccessible to the broader research community, which limits their replication and further investigation [35, 41, 56]. Furthermore, many datasets are constrained by a narrow subject age range, often focusing exclusively on specific groups, such as children or young adults, reducing their applicability to broader demographics [41, 68]. Additionally, many of the available databases are based on structured scenarios, such as storytelling or interviews, which may not fully capture the spontaneity and complexity of natural, unstructured dyadic interactions [8, 47, 58]. This scenario-based limitation may reduce the ecological validity of the findings derived from these datasets, as they do not reflect the natural flow of conversation typical of real-life exchanges. Moreover, the annotation of backchannel in some datasets is based on a limited range of cues, such as noddings or vocal signals, without incorporating multi-modal signals like facial expressions or other head gestures. This narrow annotation basis may lead to an incomplete representation of backchannel behaviours, which are inherently diverse and context-dependent [8, 21, 41, 56, 58, 65].

Furthermore, despite the growing emphasis on multi-modal data, some databases still focus primarily on unimodal datasets, which mainly consist of audio data. Many existing single-modality databases, including the Maekawa Corpus [44] and others [13, 41, 74], continue to be utilised, even though they primarily focus on audio cues. This reliance on single-modality datasets limits the potential for a more comprehensive understanding of backchannel behaviours, which are inherently multi-modal in nature.

Notably, CCDb+ includes a wide range of annotations, allowing for in-depth exploration of backchannel types and their contextual significance.

2.2 State-of-the-art Backchannel Study Methods

In backchannel research, two primary tasks have emerged: prediction, which anticipates backchannel timing and type during conversation, and detection, which identifies backchannel directly in conversation data. While prediction was initially more popular, driven by its applications in developing responsive conversational agents, detection has gained increasing attention in recent years as it provides a more nuanced understanding of backchannel behaviour, especially in natural, multi-modal settings. Early backchannel research focused on prediction, using unimodal models that leveraged features like low pitch regions to signal backchannel timing, laying the foundation for subsequent studies [74]. Later models incorporated handcrafted prosodic and part-of-speech features [9, 33, 34, 51], and as multi-modal datasets became available, head gestures and eye gaze were integrated to enhance prediction beyond just acoustic features [20, 43, 48]. The rise of deep learning brought further advances, with Long Short-Term Memory (LSTM) and Transformers making significant strides in processing continuous time-series data for backchannel prediction [1, 26, 35, 64, 72]. In contrast, the backchannel detection task, first proposed in [52], has only recently garnered attention, particularly with the push from the MultiMediate'22 Challenge [50]. Detection differs from prediction in that it focuses on identifying backchannel within a broader conversational flow, posing unique challenges in terms of accuracy and contextual adaptability. Recent state-of-the-art methods in backchannel detection employ multi-modal transformers [3, 73], with some studies also exploring the use of Gated Recurrent Units (GRUs) and attention modules [72] or graph neural networks (GNNs) [66]. However, these approaches largely rely on visual features provided by the MultiMediate'22 Challenge, which revealed that adding voice features to models that already utilised visual information often led to worse performance, with voice-only models even performing no better than random guessing. In contrast, our baseline results show that using either visual or acoustic information alone outperforms random guessing, and combining both modalities yields better results than using either modality individually. This difference in findings may be due to the distinct feedback mechanisms in group conversations compared to dyadic interactions [12].

3 CCDb+ Introduction

CCDb+ contains 44 fully annotated dyadic conversations, covering interlocutors' activities such as conversational facial expressions, head gestures, verbal and non-verbal utterances, as well as agreements and disagreements. Of these, 38 conversations were initially collected before 2013. However, only 8 of these were annotated at that time [4]. Later, an additional 6 conversations were recorded with four volunteers, and these recordings include not only video but also corresponding 3D sequences [46]. To facilitate future research on conversational dominance [42], we provide details on participant backgrounds, including their affiliation and experience, which can impact conversational dynamics. Table 2 categorises participants as staff, students, children, or actors, to aid comparative studies on interlocutor dominance. Note that the children in conversations were paired with their parents.

In this work, we annotated 30 of the remaining conversations and revised the previous annotations to ensure consistency with the updated guidelines. This comprehensive update aligns all 44 conversations with our refined annotation scheme, further enhancing the usability and accuracy of CCDb+ for studying dyadic interactions. Table 3 shows the basic facts of CCDb+.

3.1 Collection

3.1.1 Recording Equipment. CCDb+ was collected in two phases, using consistent equipment and setup to capture natural, unscripted dyadic conversations. In both phases, participants were seated opposite each other and recorded with video and audio at 44.1 KHz using lapel microphones. To ensure precise synchronisation of audio and video, a handheld buzzer and LED device were employed.

Since data for the two participants in each conversation (C1 and C2) were collected independently using identical setups but with distinct camera labels, we defined a camera classification task to assess potential data collection biases. Using a Transformer-based model [69], we achieved a test accuracy of 0.8906, indicating notable camera-related differences. Further analysis revealed three main sources of bias: (1) some participants were seated slightly differently, resulting in minor spatial inconsistencies across views; (2) variations in camera focal lengths, affecting body proportions in the frames; and (3) minor contrast differences, particularly in C2.

To reduce camera-induced bias, we conducted separate classification experiments using facial landmarks and histograms as inputs, achieving accuracies of 0.7668 and 0.8808, respectively. We then applied a series of preprocessing steps—including recentring, recropping, histogram equalisation, and landmark re-extraction – which modestly reduced the classification performance (to 0.7168 for landmarks and 0.7960 for histograms), suggesting a partial mitigation of bias. However, these refinements relied on the dlib library, which limited the precision of normalisation. We suggest that future researchers explore more robust preprocessing strategies tailored to their specific applications to better address such biases.

3.1.2 Recording Method. Phase 1 included 30 conversations among 16 speakers, aged 8 to 56. Although topics were suggested based on a pre-session questionnaire, the participants were not required to follow any script, allowing for spontaneous, dynamic exchanges. In Phase 2, six additional conversations were recorded with four volunteers aged 20 to 50, recruited from the public. Three participants had acting experience; however, they were instructed to converse

Table 1: Comparison of multi-modal databases of dyadic conversations for backchannel studies. "Content" denotes the modalities present: A (audio), V (visual), and T (transcriptions). "Pub." indicates whether backchannel annotations are publicly available

Dataset	Year	Content	Scenario	Age Ranges	Size	Pub.	Language	Annotation of backchannel types
Spontal [15]	2010	A V T	Spontaneous conversation	–	40mins	No	Swedish	Utterances, Head Position, Breathing, Coughing, Laughter
HCRC [48]	2010	A V	Storytelling	–	1h40mins	No	English	Utterances, Gaze, Nodding
CCDb [4, 46]	2013	A V	Spontaneous conversation	8-55	1h39mins	Yes	English	Utterance, Nodding, Shaking, Tilt, Laughter, Facial expressions, Thinking and Confusion
NoXi [8]	2017	A V	Spontaneous conversation	21-45	25h18mins	Yes	7 languages	Utterances
P2PSTORY [68]	2018	A V	Storytelling	5-6	1h15mins	Yes	English	Utterance, Nodding, Smile, Brow Raise, Lean Toward, Gaze
Vyaktitv [35]	2021	A V	Spontaneous conversation	19-24	14h	No	Hindi	Utterance, Nodding, Shaking, Laughter, Eyebrow
Slovak dialogue corpus [56]	2022	A V	Interview	–	8h7mins	No	Slovak	Utterance
Two-party dialogues [58]	2023	A V	Storytelling	–	–	No	Japanese	Utterance
VideoCall [6]	2023	A V	Online Playing Games	6-12	–	No	France	Posture, Gaze, Nodding, Shaking, Laughter, Eyebrow
KoreanInterview [65]	2023	–	Interview	20-30	5h6mins	No	Korean	Nodding
AutisticChildren [41]	2023	–	–	4-15	–	No	English	Utterances
NoXi+J [21]	2024	A V T	Spontaneous conversation	18-45	41h11mins	Yes	5 languages	Utterance, Nodding
CCDb+	2025	A V T	Spontaneous conversation	8-55	3h44mins	Yes	English	Utterance, Nodding, Shaking, Tilt, Laughter, Facial expressions, Thinking and Confusion

naturally on everyday topics like hobbies, films, and travel, mirroring the unscripted approach of Phase 1. Further collection details are documented in prior studies [4, 46].

Participants engaged in multiple conversations with different partners, enabling cross-participant interactions that are valuable for studying the stability of facial expressions and emotional responses across different social contexts. Additionally, interactions were left unscripted to capture authentic, spontaneous conversation dynamics. To further validate the natural flow of these dialogues, we calculated the lexical density [36] for each conversation. The mean lexical density for CCDb was 0.46, and CCDb+ averaged 0.53, aligning with typical spoken conversation values [24]. This balance reflects the appropriate level of spontaneity and conversational depth for dyadic interactions, supporting the dataset's aim to capture realistic social behaviours.

Table 2: Participants' Roles in CCDb+ Conversations

Role	Participants
Staff	P1, P3, P4, P6, P7, P8, P10, P13, P14, P15, P18
Student	P2, P5, P9, P12, P16, P19
Child	P11, P17
External Actor	P20, P21, P22

Table 3: Basic facts of CCDb+

No.of distinct participants	22
No.of conversations	44
Length of conversations	224 mins
Length of labelled set	188 mins
Length per conversation	5 mins
Age range of participants	8-55 years old
Male-to-female ratio	15:7
Frame Rate	60 FPS

3.2 Annotation

All manual annotations were carried out in ELAN [75]. Annotations included Backchannel, Frontchannel, Agree, Disagree, Head gestures, Facial expressions, Verbal utterance, Non-verbal utterance, and 11 additional commonly used conversation labels, totalling 18 labels; see some examples in Figure 1.

3.2.1 Annotation Protocol. We annotated 30 conversations with the help of 15 proficient PhD students. They were given guidelines, examples, and explanations on ambiguous labels. During annotation, expert advice was sought for clarification on some questions. Each annotation underwent a three-stage review process involving discussions and modifications to ensure accuracy. Despite potential semantic ambiguities, labels were retained if two out of three annotators agreed. The three-stage review process improves annotation accuracy by resolving ambiguities through discussion and reducing individual biases. It ensures consistency by retaining labels with majority agreement. In addition, all transcripts were independently verified by two native English speakers to ensure linguistic accuracy.

3.2.2 Annotation Definition. We harmonised the differences between the two sets of annotation labels across our previous two papers [4, 46], establishing this framework as the foundation for this work. Drawing from recent, more comprehensive discussions in the research community on backchannel behaviour, particularly the discussions in [7, 25, 54, 55], with an emphasis on identifying clear indicators for categorising interaction behaviours, we refined the annotation labels to improve clarity and consistency. The definition of each refined label is made available alongside the dataset upon access. For backchannel specifically, [38] served as a key reference, guiding our multi-stage process. This process involves playing recordings at half-speed to accurately identify backchannel through lexical and nodding cues, then classifying them by type (e.g., single or repeated words) and function (e.g., continuers, convergence tokens) based on O'Keeffe and Adolphs' functional distinctions. The scheme incorporates nodding cues as outlined in previous studies, while we further extend this by including considerations for facial expressions.

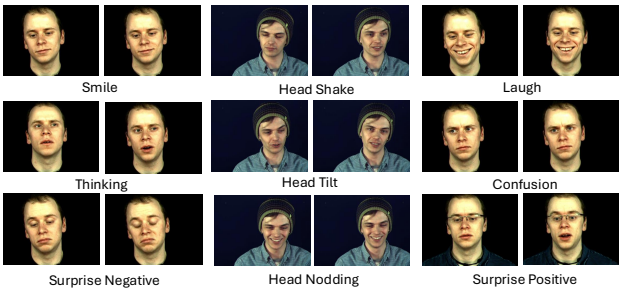

Figure 1: Examples of recorded facial expressions and head gestures

3.2.3 Inter-Annotation Agreement. We selected a subset of the CCDb+ (5% of the data) and re-annotated it independently to calculate event scores for each category to show annotation reliability. Table 4 illustrates Cohen's Kappa [11] values for various labels.

In our annotation, most labels achieved good reliability, with Cohen's Kappa values exceeding 0.75, demonstrating substantial agreement among annotators. However, certain labels, specifically Backchannel, Surprised-Positive, and Surprised-Negative, exhibited lower reliability, with Cohen's Kappa values falling below 0.4. For the Surprised expressions, low reliability can largely be attributed to the rarity of occurrence in natural conversation, especially for Surprised-Negative, which did not appear in our selected subset, resulting in a Cohen's Kappa value of 0. Surprised-Positive also had a low Kappa value due to its average duration being less than 1 second.and its subtle facial muscle movements, as evidenced by the inter-annotator confusion matrix, a substantial number of instances marked as Surprised-Positive by Annotator 1 were not marked by Annotator 2, highlighting the difficulty in consistently identifying this expression.

Table 4: Annotation Data with Total Duration, Counts, and Cohen's Kappa. Labels with Cohen's Kappa below 0.4 are in bold, and those with Cohen's Kappa above 0.75 are underlined, and the table is sorted by Cohen's Kappa values from high to low.

Label	Duration(s)	Counts	Cohen's Kappa
Non-Verbal	594.57	928	0.8698
Laugh	949.03	505	0.8476
Verbal	10780.15	3987	0.8385
Utterance	11355.15	4389	0.8242
Disagree	164.31	138	0.7835
Frontchannel	11391.3	2776	0.7605
Agree	910.39	979	0.7526
Thinking	2487.37	1481	0.6672
Happy	5464.88	1356	0.6603
Smile	5314.83	1344	0.6485
Head Nodding	1554.83	1146	0.6247
Confusion	483.43	260	0.5665
Head Shake	709.23	543	0.5457
Head Tilt	861.47	625	0.4902
Backchannel	3312.14	1898	0.3798
Surprised-Positive	321.88	236	0.2109
Surprised-Negative	24.98	19	0.0000

Additionally, the Backchannel label exhibits a relatively lower Cohen's Kappa score of 0.3798, indicating weaker inter-annotator agreement. This score was obtained by comparing the final consensus labels with an independent annotator's annotations, who occasionally misclassified frontchannel utterances or overlooked non-verbal cues. After identifying and resolving these disagreements through a secondary review with an additional checker, the score improved significantly to 0.6667. As shown in the inter-annotator confusion matrix, a substantial portion of disagreements arise from the confusion between Backchannel and Frontchannel, especially in the presence of utterances. This ambiguity likely stems from the current definition of backchannel as listener responses that do not convey new information but signal attention, understanding, or agreement. However, in practice, the boundary can be subtle. For instance, in response to a question like "Have you brought my shoes?", a listener's "yeah" may vary in prosody: a loud, emphatic "yeah" is typically interpreted as a frontchannel response, whereas a softer, neutral "yeah" may still function as a backchannel [62]. These nuanced variations complicate annotation and reduce label consistency. Nevertheless, we remain confident in

Table 5: Co-occurrence Counts and Durations of Backchannel, with Duration Proportions relative to the Total Duration.

Annotation	Count	Duration (s)	Duration Proportion
Surprised-Negative	12	10.8	0.24%
Head shake	110	103.42	2.27%
Surprised-Positive	118	114.77	2.52%
Confusion	79	131.68	2.89%
Head Tilt	155	143.06	3.15%
Thinking	248	232.36	5.11%
Non-Verbal	389	261.59	5.75%
Laugh	274	500.59	11.00%
Head Nodding	584	702.34	15.44%
Verbal	1039	785.9	17.28%
Smile	717	1497.83	32.90%

Table 6: Number of Occurrences of Backchannel in Different Time Periods (Contex Windows)

Time Period (s)	Number of Occurrences
0-1	491
1-2	488
2-3	490
3-4	473
4-5	504
5-6	483
6-7	478
7-8	483
8-9	497
9-10	517

the overall annotation reliability, as the final version was reviewed and approved by three independent annotators.

4 Database Baseline Evaluation

In this section, we outline the backchannel detection task, examining key backchannel cues such as Utterances, Facial expressions, and Head gestures. Given the multi-faceted nature of backchannel signals, detecting each type separately allows us to assess their individual contributions to social interaction. Our dataset reveals a rich distribution across these cues, with Facial expressions accounting for 49.9% of backchannel instances, Head gestures 23.8%, and Utterances (verbal and non-verbal) comprising 26.3% (see more details in Table 5).

To gain a comprehensive understanding of backchannel, we conduct detection tasks for each category, thereby not only assessing overall backchannel detection but also providing insights into how well models capture the nuances of each type of signal independently. This enables a more precise evaluation of model performance, especially given that backchannel may vary in form depending on interaction context.

4.1 Tasks Description

The MultiMediate dataset has influenced many recent studies in backchannel detection due to its rich multi-modal data from group conversations. It uses a 10-second context window to detect backchannel appearing specifically at the end of each window. This setup is advantageous for creating sliding-window models but introduces a significant limitation: backchannel in their data is heavily concentrated in the final seconds of each window, rather than being distributed evenly across the entire span. This approach risks underrepresenting naturally occurring backchannel cues that may appear more uniformly within conversational exchanges, thus limiting its applicability in detecting backchannel throughout various conversation phases.

In contrast, we prioritise an even temporal distribution of backchannel. As shown in Table 6, CCDb+ backchannel is evenly distributed throughout the segments. Each conversation video is sequentially divided into 10-second segments to perform binary classification based on a specified minimum duration threshold (0.1s) for each backchannel type. This approach allows for contextually relevant backchannel detection across the entire segment rather than confining it to a particular timeframe. Similarly, label definitions for other tasks are structured with this logic, ensuring balanced temporal representation and reducing the impact of isolated occurrences.

4.2 Baseline Models

Audio-visual representation We extracted a comprehensive set of features from each 10-second video segment for all detection tasks to capture both acoustic and visual cues.

For acoustic features, we used Spafe [45] to analyse the signal's properties across multiple domains. Our feature selection included Cepstral domain features, such as MFCCs and LFCCs, as well as spectral features like PNCCs, PSRCCs, and GFCCs, resulting in a diverse 91-dimensional representation across a 7-feature space. This combination captures both fine-grained spectral details and broader cepstral characteristics, offering a detailed acoustic profile that can enhance backchannel detection.

For visual (face) features, we leveraged OpenFace 2.0 [77], an advanced open-source toolkit that excels in facial behaviour analysis. OpenFace 2.0 is extensively used in fields like facial expression analysis, human-computer interaction, and affective computing due to its improved accuracy, computational efficiency, and adaptability to various research contexts. We extracted 714 visual features, covering head pose, facial landmarks, eye gaze, and action units, thereby providing a rich representation of facial and gestural dynamics critical for understanding multi-modal backchannels.

Baselines To provide a solid baseline for evaluating new models, we tested both a linear Support Vector Machine (SVM) [27] and a simple transformer-based architecture [69]. Recent studies in multi-modal fusion for backchannel detection have shown that a one-stream transformer architecture, where visual and acoustic features are concatenated before entering the transformer layer, delivers high performance [3]. Following these insights, we adopted this straightforward one-stream architecture to ensure clarity in baseline performance. This approach involves concatenating the full 10-second visual and acoustic feature vectors and inputting them into the transformer model to detect backchannel. This process, illustrated in Figure 2, serves as a reliable framework for comparison and advances in multi-modal backchannel detection.

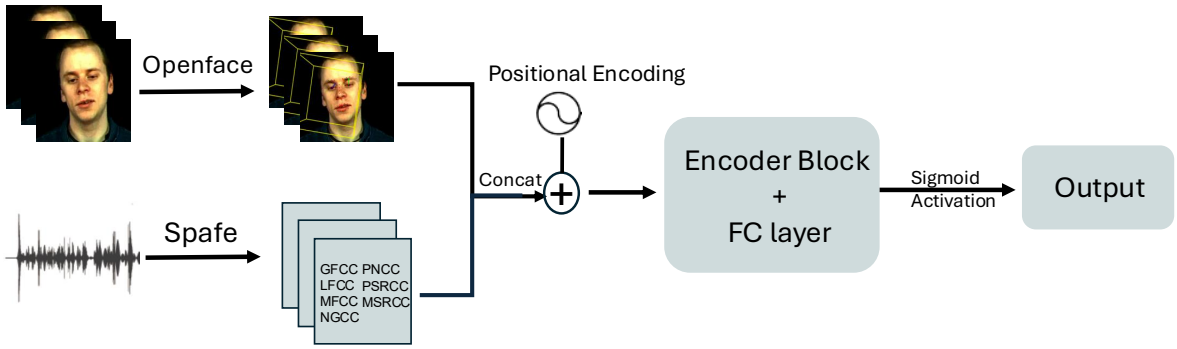

Figure 2: Overview of the baseline detection pipeline

4.3 Evaluation Metrics and Stopping Strategy

Due to class imbalance observed across labels, we consistently report both accuracy and binary F1-score for all tasks. However, our early stopping and model selection strategies depend on the degree of imbalance. Specifically, for relatively balanced tasks (positive ratio >15%), we select the epoch with the lowest validation loss, using accuracy to break ties. For highly imbalanced tasks (positive ratio $\leq$ 15%), validation binary F1-score guides model selection, as metrics like recall often become unstable (e.g., dropping to zero) during early epochs. The 15% threshold is empirically determined solely for guiding model selection, not for evaluation purposes. Binary F1 is preferred over recall-based metrics as it effectively captures both false positives and false negatives in imbalanced scenarios.

4.4 Implementation details

For tasks with a positive class ratio below 25%, we apply SMOTE [10] separately to the visual, acoustic, and combined features, ensuring a balanced training set, which is then processed in minibatches of size 64. The SVM model uses a linear kernel with a regularization parameter $C = 0.1$, while the transformer-based model is implemented in PyTorch [61] with a custom Transformer Encoder featuring 10 attention heads and a feedforward dimension of 1000. This model is optimised using Adam (learning rate 5×10^{-5}, weight decay 0.0005), with binary cross-entropy loss and a learning rate adjusted by a linear scheduler with warm-up. Training is conducted on an NVIDIA RTX 4090 GPU. Due to the pre-extraction and reusability of all visual and acoustic features across tasks, training is conducted with high efficiency. Each task completes training in approximately 3 minutes for the transformer model and around 1 minute for the SVM model.

5 Results and Discussion

To mitigate potential biases from specific data splits and ensure more robust conclusions, we employ cross-validation. The 44 conversations are divided into 11 groups, with each group sequentially serving as the test set. Within the remaining 10 groups, we randomly select one as the validation set, while the remaining 9 are used for training. This process is repeated for each fold, and we compute the average F1 score and accuracy across all folds to produce a final, reliable evaluation of model performance. Note that all transformer-based results are reported as the mean ± standard deviation across three random seeds, enhancing robustness by accounting for variability due to random initialization. This approach provides a clearer view of the model's consistency and generalisability.

Backchannel Detection task: To align with previous studies using the MultiMediate dataset, we conducted experiments using input durations of 1, 3, and 10 seconds to evaluate performance across different temporal windows for backchannel detection. As shown in Table 7, our models consistently outperform the random guessing baseline, confirming that both visual and acoustic features provide valuable information.

Obviously, the 10 seconds input window produces the best results overall, with substantial gains over random guessing across all feature types. Notably, the Transformer model with combined features achieves an F1 score of 0.7530 and the highest accuracy of 0.7170, highlighting the value of extended acoustic data.

Interestingly, in some cases, combining acoustic and visual features yields slightly lower results than using acoustic features alone, especially for the Transformer model. This discrepancy may stem from the redundancy or potential noise introduced by merging modalities, where less relevant or conflicting information between visual and acoustic channels could interfere with model optimization.

Other Detection tasks: In the detection results of various tasks, the results reveal differences between balanced and imbalanced classes. For tasks with relatively balanced classes, we selected Thinking and Nodding as single examples for analysis. In the Nodding task, the transformer model performed well with visual features, achieving the highest F1 score of 0.5652, indicating the effectiveness of visual features in detecting this type of task. Additionally, the combination of visual and audiovisual features also showed significant improvement, demonstrating that multi-modal feature fusion can effectively capture rich social behaviour cues. In the Thinking task, the combination of audiovisual features with the transformer model achieved a good performance (F1 score 0.5557), which is a slight improvement over single features. This further suggests that for cognitive expression recognition tasks, multi-modal data

Table 7: Backchannel detection results based on RandomGuess, SVM, and Transformer. The best results for each model using different feature types are highlighted in bold.

Input Length	Feature Type	Test Accuracy			Test F1 Score		
		RandomGuess	SVM	Transformer	RandomGuess	SVM	Transformer
1 second	Acoustic	0.5	0.5410	**0.5774 ± 0.0235**	0.5056	0.5525	**0.5974 ± 0.0332**
	Visual		0.5746	0.5608 ± 0.0146		**0.5898**	0.5831 ± 0.0275
	Combined		**0.5781**	0.5764 ± 0.0167		0.5816	0.5910 ± 0.0094
3 seconds	Acoustic	0.5	0.5622	0.6013 ± 0.0283	0.5056	0.5661	**0.6165 ± 0.0334**
	Visual		0.5741	0.5903 ± 0.0225		0.5792	0.6061 ± 0.0203
	Combined		**0.5936**	**0.6048 ± 0.0075**		**0.5949**	0.6113 ± 0.0144
10 seconds	Acoustic	0.5	0.6287	0.6721 ± 0.0128	0.5056	**0.6422**	0.6845 ± 0.0141
	Visual		0.6013	0.6520 ± 0.0150		0.6012	0.6623 ± 0.0008
	Combined		**0.6458**	**0.7170 ± 0.0107**		0.6389	**0.7530 ± 0.0156**

can provide more cues for the model to capture subtle behavioural expressions.

For tasks with extreme class imbalance, such as Disagree and Confusion, the overall F1 scores were low and close to random guessing levels, indicating that the model struggled to effectively capture the features of the minority class. After applying the SMOTE technique for oversampling, the F1 scores for some tasks increased slightly; however, in certain cases, SMOTE introduced too much noise, leading to a decline in model performance, particularly affecting the transformer's results (e.g., for the Disagree task with visual features, F1 score is 0.0426±0.0603). This phenomenon indicates that while data augmentation techniques can alleviate some data imbalance issues, they may introduce more noise in cases of severe imbalance, necessitating further optimization and exploration of other techniques in future work to address this problem.

Overall, these experimental results suggest that multi-modal features and transformer architectures provide some improvement for detection in relatively balanced tasks, while traditional oversampling techniques may introduce noise when dealing with severely imbalanced data. Future work could further explore effective methods for these tasks, such as incorporating more diverse data augmentation or applying techniques specifically designed for imbalanced learning.

6 Ablation Study

We evaluate the performance of each acoustic feature of our method via an ablation study on the performance of SVM. This ablation study investigates the impact of adding various acoustic features to visual features on model performance in multiple tasks. The results indicate that incorporating acoustic features alongside visual-only baselines leads to varying degrees of improvement across tasks. In particular, when combining all acoustic features with visual information, the test accuracy for each task improves noticeably: Backchannel reaches 64.58%, Nodding 63.02%, and Thinking 62.82%. This suggests that, while individual acoustic features may provide modest benefits on their own, the combined set of features provides a more substantial improvement.

For example, in the Nodding task, the inclusion of all acoustic features raises test accuracy significantly over the visual-only setup (from 60.34% to 63.02%), indicating that a comprehensive acoustic profile enhances the model's ability to detect nodding behaviour. Similarly, Thinking benefits from additional acoustic cues, achieving an accuracy of 62.82% with the combined features versus lower individual accuracies with specific acoustic features. For Backchannel, the combined features lead to the highest improvement, reaching 64.58%, compared to only modest gains when using individual acoustic features like 'MSRCC' (60.78%).

These findings suggest that, although some tasks may benefit from specific acoustic features, the inclusion of a full acoustic set with visual features yields the most consistent and significant performance boosts across tasks.

7 Conclusion

This study advances backchannel detection by integrating multimodal data—vocalisations, facial expressions, and head gestures—to better capture the nuances of social interaction. Using CCDb+, a natural dyadic conversation corpus with audio-visual modalities, we benchmarked different backchannel types with SVM and a basic Transformer, establishing performance baselines. Results show that a one-stream multi-modal fusion improves detection accuracy and underscores the benefit of combining visual and vocal cues for richer behavioural insight. However, annotation quality remains a challenge, especially for Backchannel, Agree, and Disagree labels, due to individual expressiveness and context-dependent interpretations. Temporal misalignment between dialogue scripts and brief visual signals further limited the use of scripts as input features. We also noted slight positional bias introduced during data collection. Future work should improve annotation protocols for context-sensitive cases and explore alignment strategies for incorporating dialogue scripts. This study lays the groundwork for more robust multi-modal models of social interaction.

Acknowledgments

We would like to thank all participants involved in the CCDb+ recordings, as well as the annotators for their valuable contributions. We are also grateful to the professionals who supported various stages of the data collection and annotation process. Their efforts were essential to the development of this dataset. The work was partially supported by the Engineering and Physical Sciences Research Council (No. EP/Y028805/1).

References

[1] Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. 2021. Towards immediate backchannel generation using attention-based early prediction model. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7408–7412.

[2] Jens Allwood and Loredana Cerrato. 2003. A study of gestural feedback expressions. In *First Nordic Symposium on Multimodal Communication*. 7–22.

[3] 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.

[4] Andrew J Aubrey, David Marshall, Paul L Rosin, Jason Vendeventer, Douglas W Cunningham, and Christian Wallraven. 2013. Cardiff conversation database (CCDb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 277–282.

[5] German Barquero, Johnny Núnez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. 2022. Didn't see that coming: a survey on non-verbal social human behavior forecasting. In *Understanding Social Behavior in Dyadic and Small Group Interactions*. PMLR, 139–178.

[6] Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. 2023. Using video calls to study children's conversational development: the case of backchannel signaling. *Frontiers in Computer Science* 5 (2023), 1088752.

[7] Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation. In *12th Edition of its Language Resources and Evaluation Conference (LREC 2020)*. European Language Resources Association (ELRA), 549–558.

[8] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 350–359.

[9] Nicola Cathcart, Jean Carletta, and Ewan Klein. 2003. A shallow model of backchannel continuers in spoken dialogue. In *Conference on European Chapter of the Association for Computational Linguistics*. 51–58.

[10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[11] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[12] Gus Cooney, Adam M Mastroianni, Nicole Abi-Esber, and Alison Wood Brooks. 2020. The many minds problem: disclosure in dyadic versus group conversation. *Current Opinion in Psychology* 31 (2020), 22–27.

[13] Noah Coccaro Rachel Martin Marie Meteer Klaus Ries Elizabeth Shriberg Andreas Stolcke Paul Taylor Carol Van Ess-Dykema Daniel Jurafsky, Rebecca Bates. 1998. Johns Hopkins LVCSR Workshop-97 Switchboard Discourse Language Modeling Project Final Report. (1998).

[14] Liyanage C De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications*, Vol. 1. IEEE, 397–401.

[15] Jens Edlund, Jonas Beskow, Kjell Elenius, Kahl Hellmer, Sofia Strömbergsson, and David House. 2010. Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture.. In *LREC*. 2992–2995.

[16] Kevin El Haddad, Hüseyin Cakmak, Marwan Doumit, Gueorgui Pironkov, and Ugur Ayvaz. 2017. Social Communicative Events in Human Computer Interactions. In *Proceedings of the 11th International Summer Workshop on Multimodal Interfaces-eNTERFACE'16*.

[17] Kevin El Haddad, Hüseyin Çakmak, Emer Gilmartin, Stéphane Dupont, and Thierry Dutoit. 2016. Towards a listening agent: a system generating audiovisual laughs and smiles to show interest. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 248–255.

[18] Kevin El Haddad and Thierry Dutoit. 2020. Cross-Corpora Study of Smiles and Laughter Mimicry in Dyadic Interactions. In *Laughter and Other Non-Verbal Vocalisations Workshop: Proceedings (2020)*.

[19] Kevin El Haddad, Yara Rizk, Louise Heron, Nadine Hajj, Yong Zhao, Jaebok Kim, Trung Ngô Trọng, Minha Lee, Marwan Doumit, Payton Lin, et al. 2018. End-to-end listening agent for audiovisual emotional and naturalistic interactions. *Journal of Science and Technology of the Arts* 10, 2 (2018), 49–61.

[20] S Fujie. 2004. A conversation robot using head gesture recognition as paralinguistic Information. *Proc. International Workshop on Robot and Human Interactive Communication RO-MAN* (2004).

[21] Marius Funk, Shogo Okada, and Elisabeth André. 2024. Multilingual dyadic interaction corpus NoXi+J: Toward understanding asian-european non-verbal cultural characteristics and their influences on engagement. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 224–233.

[22] Sarah Gillet. 2024. *Computational Approaches to Interaction-Shaping Robotics*. Ph. D. Dissertation. KTH Royal Institute of Technology.

[23] Sarah Gillet, Daniel Marta, Mohammed Akif, and Iolanda Leite. 2024. Shielding for Socially Appropriate Robot Listening Behaviors. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2279–2286.

[24] Emer Gilmartin, Christian Saam, Brendan Spillane, Maria O'Reilly, Ketong Su, Arturo Calvo Devesa, Loredana Cerrato, Killian Levacher, Nick Campbell, and Vincent Wade. 2018. The ADELE corpus of dyadic social text conversations: Dialog act annotation with ISO 24617-2. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[25] Agustín Gravano, Julia Hirschberg, and Štefan Beňuš. 2011. Affirmative cue words in task-oriented dialogue. *Computational Linguistics* 38, 1 (2011), 1–39.

[26] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. *Listener* 162 (2018), 364.

[27] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications* 13, 4 (1998), 18–28.

[28] Anna Hjalmarsson and Catharine Oertel. 2012. Gaze direction as a back-channel inviting cue in dialogue. In *IVA 2012 workshop on realtime conversational virtual agents*, Vol. 9.

[29] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. Virtual Rapport 2.0. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. Springer, 68–79.

[30] Yuyun Huang, Emer Gilmartin, and Nick Campbell. 2016. Conversational Engagement Recognition Using Auditory and Visual Cues.. In *Interspeech*. 590–594.

[31] Yuyun Huang, Emer Gilmartin, and Nick Campbell. 2017. Speaker Dependency Analysis, Audiovisual Fusion Cues and a Multimodal BLSTM for Conversational Engagement Recognition.. In *INTERSPEECH*. 3359–3363.

[32] Benjamin Inden, Zofia Malisz, Petra Wagner, and Ipke Wachsmuth. 2013. Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 181–188.

[33] Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Real-time and continuous turn-taking prediction using voice activity projection. *arXiv preprint arXiv:2401.04868* (2024).

[34] Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. 2024. Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection. *arXiv preprint arXiv:2410.15929* (2024).

[35] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring semi-supervised learning for predicting listener backchannels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[36] Victoria Johansson. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics* 53 (2008), 61–79.

[37] Dawn Knight. 2009. *A multi-modal corpus approach to the analysis of backchanneling behaviour*. Ph. D. Dissertation. University of Nottingham Nottingham, United Kingdom.

[38] Dawn Knight, Anne O'Keeffe, Geraldine Mark, Christopher Fitzgerald, Justin McNamara, Svenja Adolphs, Benjamin Cowan, Tania Fahey Palma, Fiona Farr, and Sandrine Peraldi. 2024. Indicating engagement in online workplace meetings: The role of backchannelling head nods. *International Journal of Corpus Linguistics* (2024).

[39] Birgit Knudsen, Ava Creemers, and Antje S Meyer. 2020. Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology* 11 (2020), 593671.

[40] Peter A Krause and Alan H Kawamoto. 2021. Predicting one's turn with both body and mind: Anticipatory speech postures during dyadic conversation. *Frontiers in Psychology* 12 (2021), 684248.

[41] Grace O Lawley, Peter A Heeman, and Steven Bedrick. 2023. Computational Analysis of Backchannel Usage and Overlap Length in Autistic Children. In *Proceedings of the First Workshop on Connecting Multiple Disciplines to AI Techniques in Interaction-centric Autism Research and Diagnosis (ICARD 2023)*. 17–23.

[42] Helena M Leet-Pellegrini. 1980. Conversational dominance as a function of gender and expertise. In *Language*. 97–104.

[43] RM Maatman, Jonathan Gratch, and Stacy Marsella. 2005. Natural behavior of a listening agent. In *Intelligent Virtual Agents: 5th International Working Conference, IVA 2005, Kos, Greece, September 12-14, 2005. Proceedings 5*. Springer, 25–36.

[44] Kikuo Maekawa. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

[45] Ayoub Malek. 2023. Spafe: Simplified python audio features extraction. *Journal of Open Source Software* 8, 81 (2023), 4739.

[46] Andrew David Marshall, Paul L Rosin, Jason Vandeventer, and Andrew Aubrey. 2015. 4D Cardiff Conversation Database (4D CCDb): A 4D database of natural, dyadic conversations. *Auditory-Visual Speech Processing,{AVSP} 2015* (2015), 157–162.

[47] Gary McKeown, William Curran, Johannes Wagner, Florian Lingenfelser, and Elisabeth André. 2015. The Belfast storytelling database: A spontaneous social interaction database with laughter focused annotation. In *2015 International*

Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 166–172.

[48] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous agents and multi-agent systems* 20 (2010), 70–84.

[49] Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. 2015. Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques. In *Human-Computer Interaction: Interaction Technologies: 17th International Conference, Proceedings, Part II 17*. Springer, 329–340.

[50] 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.

[51] Ryota Nishimura, Norihide Kitaoka, and Seiichi Nakagawa. 2007. A spoken dialog system for chat-like conversations considering response timing. In *Text, Speech and Dialogue: 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007. Proceedings 10*. Springer, 599–606.

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

[53] Neal R Norrick. 2012. Listening practices in English conversation: The responses responses elicit. *Journal of Pragmatics* 44, 5 (2012), 566–576.

[54] Manfred Nusseck, Douglas W Cunningham, Christian Wallraven, and Heinrich H Bülthoff. 2008. The contribution of different facial regions to the recognition of conversational expressions. *Journal of vision* 8, 8 (2008), 1–1.

[55] Anne O'Keeffe, Dawn Knight, Geraldine Mark, Christopher Fitzgerald, Justin McNamara, Svenja Adolphs, Benjamin Cowan, Tania Fahey Palma, Fiona Farr, and Sandrine Peraldi. 2024. "We've lost you Ian": Multi-modal corpus innovations in capturing, processing and analysing professional online spoken interactions. *Research in Corpus Linguistics* 12, 2 (2024).

[56] Stanislav Ondáš, Eva Kiktová, and Matúš Pleva. 2022. Slovak dialogue corpus with backchannel annotation. In *2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 1–4.

[57] Stanislav Ondáš, Eva Kiktová, Matúš Pleva, and Jozef Juhár. 2023. Analysis of Backchannel Inviting Cues in Dyadic Speech Communication. *Electronics* 12, 17 (2023), 3705.

[58] Toshiki Onishi, Naoki Azuma, Shunichi Kinoshita, Ryo Ishii, Atsushi Fukayama, Takao Nakamura, and Akihiro Miyata. 2023. Prediction of Various Backchannel Utterances Based on Multimodal Information. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 1–4.

[59] Maria Teresa Parreira, Sarah Gillet, and Iolanda Leite. 2023. Robot Duck Debugging: Can Attentive Listening Improve Problem Solving? *arXiv preprint arXiv:2301.06511* (2023).

[60] Maria Teresa Parreira, Sarah Gillet, Katie Winkle, and Iolanda Leite. 2023. How did we miss this? a case study on unintended biases in robot social behavior. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 11–20.

[61] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32 (2019).

[62] Vojtěch Pipek. 2007. *On backchannels in English conversation*. Ph. D. Dissertation. Masarykova univerzita, Pedagogická fakulta.

[63] Robin Ruede. 2017. *Backchannel prediction for conversational speech using recurrent neural networks*. Ph. D. Dissertation. Informatics Institute.

[64] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, right, uh-huh: a deep learning backchannel predictor. In *Advanced social interaction with agents: 8th international workshop on spoken dialog systems*. Springer, 247–258.

[65] Sukyung Seok, Tae-Hee Jeon, Yu-Jung Chae, ChangHwan Kim, and Yoonseob Lim. 2023. Explorative Study on the Non-verbal Backchannel Prediction Model for Human-Robot Interaction. In *International Conference on Social Robotics*. 264–275.

[66] 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.

[67] Mohit Sharma, Dragan Ahmetovic, László A Jeni, and Kris M Kitani. 2018. Recognizing visual signatures of spontaneous head gestures. In *2018 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 400–408.

[68] Nikhita Singh, Jin Joo Lee, Ishaan Grover, and Cynthia Breazeal. 2018. P2PSTORY: dataset of children as storytellers and listeners in peer-to-peer interactions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

[70] Pierre Vuillecard, Arya Farkhondeh, Michael Villamizar, and Jean-Marc Odobez. 2024. CCDb-HG: Novel Annotations and Gaze-Aware Representations for Head Gesture Recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–9.

[71] Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-taking and backchannel prediction with acoustic and large language model fusion. *arXiv preprint arXiv:2401.14717* (2024).

[72] Kangzhong Wang, MK Michael Cheung, Youqian Zhang, Chunxi Yang, Peter Q Chen, Eugene Yujun Fu, and Grace Ngai. 2023. Unveiling Subtle Cues: Backchannel Detection Using Temporal Multimodal Attention Networks. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9586–9590.

[73] Kangzhong Wang, Xinwei Zhai, MK Cheung, Eugene Yujun Fu, Peter Qi Chen, Grace Ngai, and Hong Va Leong. [n. d.]. TMAN: a Temporal Multimodal Attention Network for Backchannel Detection. http://dx.doi.org/10.2139/ssrn.5086867.

[74] 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.

[75] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*. 1556–1559.

[76] 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.

[77] Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit tadas baltrušaitis. In *IEEE International Conference on Automatic Face and Gesture Recognition*.