

Behavioral Alignment and the Engagement for Non-Speaking Participants in Online Meetings

Trung Duc Anh Dang*
University of Copenhagen

Zhiqing Peng*
University of Copenhagen

Patrizia Paggio†
University of Copenhagen
paggio@hum.ku.dk

Abstract

Engagement detection in meetings has focused mostly on speakers, who produce explicit prosodic and acoustic cues. The engagement of non-speaking participants is harder to read because it surfaces only in nonverbal behaviour. We predict how engaged a silent listener is, on a three-level scale (Engaged, Intermittent, Disengaged), in five-second windows of remote Zoom meetings from the GEstures and Head Movements in Language (GEHM) corpus. Each window is described by three feature families: the listener’s own nonverbal behaviour, the listener’s behavioural alignment with the others present, including the active speaker, and that speaker’s prosody. Interpersonal coordination is an established correlate of rapport and involvement. We ask the narrower question of whether a compact window-level alignment measure adds predictive value over listener-only features when confident labels are scarce. Under leakage-free cross-validation, listener-speaker alignment is the most useful addition we test, and the speaker’s terminal pitch slope is an interpretable correlate of listener engagement. The gains are modest and their bootstrap intervals include zero. Unsupervised analysis finds no engagement structure. Two findings are firmer: the task is effectively binary, actively engaged versus not, bounded by the scarcity of confident disengaged labels, and the pitch-slope correlate holds despite the small sample.

1 Introduction

Online meetings have become a standard means of communication across work and education due to the increasing need for remote collaboration, and engagement detection in this setting is an active research area, since engagement is a key indicator of meeting quality, productivity, and effectiveness

(Frank et al., 2016). Research on engagement detection mainly targets the speakers, who produce explicit and quantifiable engagement signals, especially prosodic-acoustic features (Pellet-Rostaing et al., 2023). Feedback from listeners, conversely, is often conveyed through subtle and multifunctional non-verbal cues, making it challenging to interpret and classify their level of engagement (Pellet-Rostaing et al., 2023). Listener engagement is also an important factor shaping meeting dynamics and productivity, which motivates analyzing the engagement of non-speaking participants directly.

The unit of prediction is a five-second window in which a target participant is *silent* while someone else holds the floor, taken from the GEHM corpus of remote multi-party Zoom meetings (Paggio et al., 2024). For each window we predict how engaged the silent target appears, on the three-level scale used by our annotators: *Engaged*, *Intermittent*, and *Disengaged* (defined in §3.1). We adopt the behavioural definition of engagement proposed by Fredricks et al. (2004): visible participation and involvement in interaction, read from head pose, eye gaze, and facial expression (Kaur et al., 2018). The target produces no words in these windows, so engagement must be read from nonverbal behaviour and from the surrounding interaction.

Various behavioral features have been used for engagement detection and classification, including prosodic-acoustic cues, facial expressions, gestures, postures, head pose, and gaze behaviors (Oertel et al., 2011; Singh et al., 2023; Inoue et al., 2018). Recent engagement estimation models take conversational context into account, estimating a target participant’s engagement by combining the participant’s own audio-visual behaviour with cues from their conversational partners (Li et al., 2024; Müller et al., 2024). These methods depend on supervised engagement labels. They integrate the partner through learned audio-visual representations, not through an explicit measure of the be-

* Equal contribution.

† Corresponding author.

havioural relation between the two.

In social signal processing and social neuroscience, interpersonal coordination is an established correlate of rapport, affiliation, and involvement (Bernieri et al., 1991; Delaherche et al., 2012; Ohayon and Gordon, 2025); we review this work in §2. Our question is therefore not whether coordination relates to engagement but whether it can be operationalised usefully under the constraints of this setting. Does a compact, window-level measure of how similarly a silent listener behaves to the others present add predictive value over listener-only features? The setting is unfavourable: the measure comes from off-the-shelf visual representations, the meetings are remote and lack the mutual-gaze and proxemic cues of co-present interaction, and confident labels are few. We treat this as a question about incremental predictive value in an ablation. Our analysis is correlational and confined to a single window. They cannot causally separate a listener’s individual behaviour from their coordination with others; we return to this in the Limitations section.

First, we describe an exploratory clustering experiment in which we investigate whether engagement behaviour can be learnt in an unsupervised way. Then we report the results of several classification experiments in which engagement is first predicted in terms of three classes, and then based on a binary distinction, using different feature families.

Our contributions are: (1) **An explicit window-level alignment measure for silent listeners:** We quantify how much the listener’s behavioural alignment with others adds over listener-only behaviour. In our ablation, listener-speaker alignment is the strongest addition, ahead of alignment with other listeners and of speaker-only features. The margins are within fold-to-fold variance and their bootstrap intervals include zero. The speaker’s terminal pitch slope is an interpretable correlate of listener engagement. (2) **Label-scheme sensitivity analysis under label scarcity:** We test several binary reductions of the original three-class engagement scheme to examine whether the task difficulty comes from weak behavioural signal or from the label structure itself. This provides a basis for interpreting whether poor fine-grained performance reflects insufficient features or ambiguity in the engagement labels.¹

¹Code, feature-extraction pipeline, and annotation guidelines are available at <https://github.com/ducanhdt/listener-meeting-engagement>.

2 Related Work

Engagement in conversation and meetings.

Engagement and involvement have often been detected through prosodic and movement cues (Gatica-Perez et al., 2005). Oertel et al. (2011) found that prosodic cues, together with head and body movement, correlated with the level of involvement in conversation. These cues are less available for non-speaking participants, whose engagement is expressed mainly through subtler non-verbal behaviour. Recent engagement estimation studies also use cues from conversational partners. The Dialogue-Aware Transformer framework (DAT) by Li et al. (2024), for instance, uses both the participant’s behaviour and information from their conversational partner. These models learn participant–partner interactions from supervised audio-visual data, whereas we explicitly measure the listener’s behavioural alignment with other meeting participants, including the speaker, against a small manually annotated dataset.

Listener cues and label subjectivity. For silent listeners, previous work has focused mainly on feedback behaviour. Inoue et al. (2018) recognize engagement from listener behaviours such as backchannels, laughter, nodding, and gaze, while also noting that engagement judgments vary across annotators. Related work in online learning also predicts engagement from visual cues such as facial expressions, head pose, posture, and eye movement, often in settings where learners are not speaking (Dewan et al., 2019; Kaur et al., 2018). We adopt similar visual descriptors for silent listeners, while facing the same central challenge: engagement labels are partly shaped by annotator uncertainty.

Classifying the silent participant. The closest precedent for our task is Oertel et al. (2015), who study silent participants in co-present group discussions. Raters assessed both the *listener role* a silent participant occupies and their *engagement level* from 15-second video clips, and the authors then classified those roles automatically from audio-visual cues. Our scheme keeps only the second of these dimensions: the three classes grade the degree of attentional involvement a listener displays in a five-second window (§3.1), without distinguishing whether that listener is the one currently addressed. Role is therefore an uncontrolled confound for our labels, since some of what an annotator reads as intermittent involvement may be

a listener who is simply not being addressed. The settings differ as well. Their discussions are co-present; ours are remote gallery-view Zoom meetings, where mutual gaze is not recoverable and role is correspondingly harder to infer.

Interpersonal coordination as a social signal.

That interacting people coordinate their behaviour, and that the degree of coordination tracks the quality of the interaction, is established. [Bernieri et al. \(1991\)](#) showed that observer-rated movement synchrony predicts judgments of rapport. [Delaherche et al. \(2012\)](#) survey how synchrony has been operationalised across psychology, social signal processing, and clinical research; the measures differ in time scale and in what counts as similarity. The meta-analysis of [Ohayon and Gordon \(2025\)](#) finds a positive association between multimodal interpersonal synchrony and social outcomes such as affiliation and cooperation, with substantial heterogeneity across modalities and measurement choices. [Wohltjen and Wheatley \(2021\)](#) link coordination to attention directly: in dyadic conversation, moments of eye contact coincide with peaks of pupillary synchrony and are followed by its decline, so more synchrony does not simply mean more shared attention. In language technology the same idea appears as entrainment. [Levitan et al. \(2011\)](#) show that speakers entrain on acoustic-prosodic cues preceding backchannels, and that greater similarity accompanies better dialogue coordination and task success ([Levitan et al., 2012](#)). [Müller et al. \(2018\)](#) predict low rapport in small groups from non-verbal behaviour including synchronisation features. Closest to our data, [Paggio et al. \(2026\)](#) show that in the GEHM corpus speakers move their heads more in target intervals where verbal feedback is present.

Our position. We do not claim that coordination is a new indicator of engagement; the work above already supports it. What the literature does not settle is whether a compact, unlagged, window-level similarity, of the kind a small labelled sample can support, retains any of that signal for silent listeners in remote meetings. That is the question we test.

3 Methodology

We predict the engagement of a target participant in short windows in which that participant is *silent*. Within each meeting we divide every participant’s video into five-second windows with a two-second

overlap. For a given target we keep only the windows in which the target produces no speech while another participant holds the floor. Each such window is the unit of prediction. Engagement must therefore be read from the target’s nonverbal behaviour and from the surrounding interaction, since the target participant by definition produces no words in the windows under consideration. Figure 1 summarises the full pipeline.

3.1 Corpus and Labels

We use the GEHM corpus ([Paggio et al., 2024](#)), collected by one of the present authors in an earlier project: twelve remote multi-party meetings (about eight hours, English) recorded over Zoom, so each participant appears head-and-shoulders in a separate gallery-view stream with a separate audio channel, and is mostly silent while listening to whoever currently holds the floor. The engagement annotation described below was produced for this study and is new.

Each participant’s video is cut into five-second windows with a two-second overlap. We use the TextGrid speech alignments to remove windows in which the target speaks, including questions and short verbal responses, and drop windows where the target is not visible in at least 95% of frames (OpenPose detection; [Cao et al., 2021](#)) or faces substantially away from the screen, judged by average head orientation. Near-duplicate windows are clustered by CLIP (Contrastive Language–Image Pre-training; [Radford et al., 2021](#)) embedding, keeping one representative per cluster to reduce visual redundancy. Each retained window is paired with two randomly chosen co-present participants from the same timeframe, whose streams supply the reference signals for the listener $\leftrightarrow$ co-participant alignment. After filtering, 549 silent windows remain.

The active speaker for a window is the co-participant whose transcript overlaps it the most, identified for about 64% of windows; windows with no overlapping speech have no identified active speaker, and the speaker family is then imputed with the training-fold median (§3.3).

Engagement definitions. Annotators judged, for each five-second window, how engaged the silent participant appeared *in that window*, using the behavioural notion of engagement already mentioned in §1: visible participation and involvement, read off head pose, gaze direction, facial expression, and head and hand movement ([Fredricks et al., 2004](#);

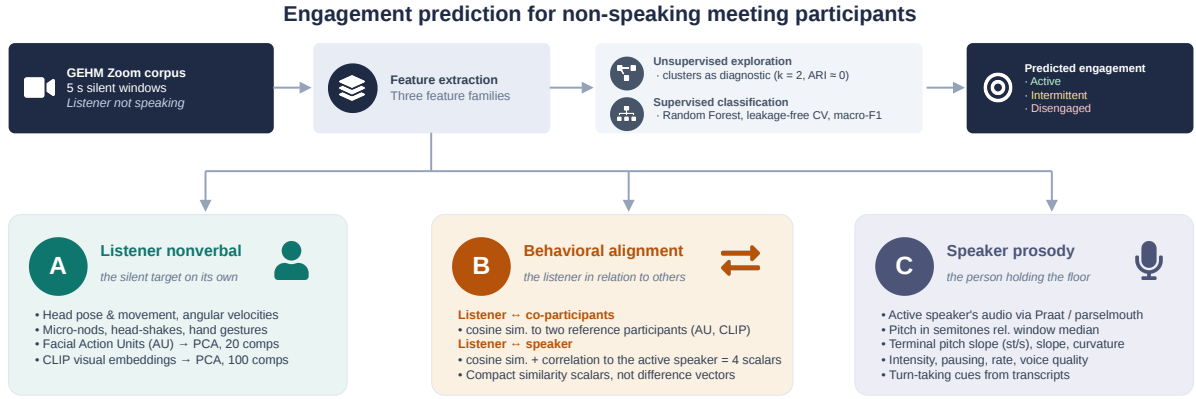


Figure 1: Overview of the approach. Each five-second window in which the target participant is silent is described by three feature families: the listener’s own nonverbal behaviour (A), the behavioural alignment between the listener and others (B), and the active speaker’s prosody (C). The same features feed an unsupervised exploration, where clustering is used only as a diagnostic, and the supervised classification that predicts the listener’s engagement.

Kaur et al., 2018). The guideline defined the three levels as follows. *Engaged* (E): the target appears to be attending to the ongoing talk throughout the window, typically oriented towards the screen, with responsive behaviour such as nodding, gaze tracking the active speaker’s tile, or facial reactions that follow the content. *Disengaged* (D): the target shows no sign of attending to the ongoing talk, for instance sustained gaze away from the screen, sustained absence of responsive movement, or visible involvement in an unrelated activity. *Intermittent* (I) was intended as the label for windows that could not confidently be placed in either of these two classes: the target shows mixed or weak cues, for example occasional distraction while still broadly following the speaker, so that the annotator cannot say with confidence whether the target is engaged or not. Annotators marked a window *unclear* when they could not make any judgment at all. The middle class is thus by design a region of annotator uncertainty rather than a distinct behavioural state, which bears on how the three-class results should be read (§4.4).

Annotation and agreement. Four non-expert annotators labelled 107 selected windows under this guideline. They saw only the target’s video stream for the window, without audio and without the other participants’ streams, so the judgment rested on the target’s visible behaviour rather than on what was said or on how others reacted. One annotator’s labels diverged from the other three (mean pairwise Cohen’s $\kappa = 0.17$ (Cohen, 1960), far below the agreement among the other three) and were excluded. The exclusion was decided from

the agreement matrix alone, before performing any experiment. The three retained annotators reach Fleiss’ $\kappa = 0.42$ (Fleiss, 1971), moderate agreement. Engagement judgments from thin slices of silent video vary across annotators (Inoue et al., 2018), and most of our disagreement falls on the middle *Intermittent* class rather than on the *Engaged/Disengaged* contrast. Windows without a majority label, or marked *unclear* by most annotators, were removed, leaving 102 windows (46 unanimous, the rest by majority vote). Removing windows that lacked two reference participants gives the final 89. The supervised experiments use those 89 windows (44 Engaged, 28 Intermittent, 17 Disengaged); the unsupervised exploration (§4.1) uses all 549 silent windows.

Restricting the evaluation to windows with a clear majority label is a trade. It removes the windows where the label is least trustworthy, at the cost of evaluating on a smaller and easier sample, so the absolute scores we report are an upper bound on performance over all silent windows. The resulting 89-window sample is small and class-imbalanced, which is why we report against a majority-class floor and treat all gains as directional. All unsupervised steps, standardisation, PCA, and the speaker-feature imputation statistics, however, are fit on all 549 silent windows, not only the 89 labelled ones. We treat the three-class formulation as primary because it is what the annotators used. The moderate agreement on the middle class is why we also report a family of label reductions (§3.3, §4.4) to verify whether the difficulty lies in the behaviour or in the label scheme.

3.2 Feature Families

We represent every window with three families of features, organised so that the contribution of the *interaction* can be isolated from that of the listener alone.

Family A: listener nonverbal behaviour. The first family characterises the target in isolation, following the head-pose and head-movement descriptors used for engagement in mostly silent settings such as online learning (Gupta et al., 2016; Dewan et al., 2019). From per-frame body and face keypoints (Cao et al., 2021) and estimated head orientation, we derive head-pose statistics, angular velocities, and counts of micro-nods and head-shakes, together with hand-gesture cues detected from wrist and face keypoints (Lugaresi et al., 2019). We add per-frame facial Action Units (AU) (Baltrušaitis et al., 2018) and CLIP visual embeddings (Radford et al., 2021) of the listener’s frames. The Action Units and CLIP embeddings are each reduced by principal component analysis to the number of components that preserves nearly all of their variance: 20 components for the Action Units and 100 for the CLIP embeddings, retaining roughly 100% and 97% of the variance, respectively. This family is the target-only baseline against which the interactional families are measured.

Family B: behavioral alignment. The second family measures how closely the listener’s behaviour resembles others in the meeting, motivated by the long-standing finding that interpersonal entrainment and movement synchrony reflect rapport and affiliation (Levitan et al., 2011; Bernieri et al., 1991; Müller et al., 2018; Ohayon and Gordon, 2025). Among the many ways synchrony has been operationalised (Delaherche et al., 2012), ours is at the simple end: an unlagged, within-window similarity between representations of two people’s visible behaviour, a choice driven by sample size rather than theory (see Limitations). We compute alignment along two distinct relations. *Listener* $\leftrightarrow$ *co-participants* alignment is the cosine similarity between the target and two randomly chosen co-present listeners, together with the similarity between those two, on the AU and CLIP channels, capturing whether the listener behaves in step with the others and stays oriented toward the meeting rather than looking away. *Listener* $\leftrightarrow$ *speaker* alignment is computed directly between the target and the single concurrently active speaker in the

same window, on the visual (CLIP) and Action-Unit channels, via cosine similarity and correlation. We encode alignment as a small set of similarity *scalars* rather than as a high-dimensional element-wise difference vector, which keeps this family compact enough to be usable given the relatively small sample size we are working with.

Family C: speaker prosody. The third family characterises the person currently holding the floor, on the premise that the prosody of ongoing speech is part of the interactional context that shapes how engaged listeners appear: speakers provide prosodic cues, terminal pitch movement among them, that invite listener feedback (Gravano and Hirschberg, 2011). The active speaker for a window is the co-participant whose speech overlaps it the most. From that speaker’s separated audio, we extract acoustic prosody features with Praat (Boersma and Weenink, 2021) and the Parselmouth library (Jadoul et al., 2018), expressed in semitones relative to the window median so that it is comparable across speakers: mean pitch and pitch variability, overall and terminal pitch slope, intensity, pausing, speaking rate, and voice quality. From the time-aligned transcripts, we also retain conversational turn-taking cues: how many participants are speaking in the window, how long the current speaker has already held the floor, how recently the listener last spoke, how much speech from others overlaps, and the window’s position in the meeting. Because the speaker generally addresses the group rather than the specific silent listener, we read these cues as concurrent interactional *context* that co-varies with the listener’s state, not as a signal directed at the listener.

3.3 Experimental Setup and Evaluation

Unsupervised exploration. Given the fact that reliable annotated labels of engagement only cover a small subset of the material, we first ask whether engagement can be recovered *without* them. This follows a semi-supervised intuition: listeners attending to the same speaker should behave similarly, so engagement states might surface as natural groups in the behavioural feature space, which a handful of labelled examples per group could then name. We cluster the windows with three standard algorithms, KMeans, a Gaussian Mixture Model (GMM), and Ward agglomerative clustering, on two feature configurations, the listener-only baseline and the baseline augmented with alignment,

and select the number of groups by silhouette width. We then compare the clusters with the gold labels through the Adjusted Rand Index. We treat clustering as a diagnostic of whether engagement is a salient axis of behavioural variation, not as a replacement for classification (§4.1).

Supervised classification. The supervised pipeline measures the contribution of each feature family with stratified five-fold cross-validation. All standardisation, PCA, and imputation are fit on the training folds only, so the protocol is leakage-free. In particular, speaker features that are missing when no speaker is identified are assigned the value of the training-fold median. Because the classes are imbalanced (44/28/17), we report macro- F_1 as the primary metric, with accuracy and a majority-class floor (always predicting the most frequent class) as references. The primary estimator is a Random Forest of 300 trees with balanced class weights, which handles the high-dimensional visual features well; we also report Logistic Regression, with balanced class weights and up to 2000 iterations, to confirm the direction of effects. We quantify the sampling uncertainty of each comparison with a paired bootstrap: 5000 resamples stratified by the true label, sampling windows with replacement within each class so every replicate keeps all classes and the original class balance, from which we report a 95% percentile interval for every macro- F_1 margin. The bootstrap holds the fold models fixed and reuses their out-of-fold predictions, so it estimates the variability due to which windows are labelled rather than model refitting; it is paired, meaning the same resampled windows score every configuration, so a margin’s interval is not inflated by between-configuration differences, and the one-sided p is the resampled probability that a margin is at most zero. All models use scikit-learn (Pedregosa et al., 2011) under a fixed random seed (42), including the clustering algorithms used in the unsupervised exploration. To separate engagement signal from annotation noise, the same protocol is applied to several binary reductions of the original three-class label scheme as a sensitivity analysis.

4 Results and Discussion

4.1 Initial Exploration: Does Engagement Cluster?

Our starting point, following the original formulation of this task, was semi-supervised (§3.3): if listeners attending to the same speaker exhibit similar behavioural patterns, different engagement states should form distinguishable clusters in the feature space, which can then be interpreted using a small number of labelled examples. This is particularly useful in settings where high-confidence labels are scarce. We therefore cluster the windows before training any classifier and ask whether the groups coincide with the three engagement classes; the labels are used only afterwards, to check the correspondence.

Two findings emerge. First, the feature space prefers *two* groups rather than three: silhouette width peaks at $k = 2$ (KMeans 0.098, against 0.030 at the annotated $k = 3$), cross-checked with the GMM’s Bayesian and Akaike Information Criteria (BIC/AIC) and a Ward dendrogram. Second, the groups do not correspond to engagement: forcing $k = 3$, the Adjusted Rand Index between clusters and gold labels is near zero (-0.01 to -0.03), and adding alignment features does not change this.

This second result holds across the other two clusterers, so it is not an artefact of one algorithm: Gaussian mixture is degenerate on this sample, collapsing the labelled windows into a single component with a silhouette curve that coincides with KMeans at every k , and Ward agglomerative clustering agrees on the substance, with weak structure, a marginal $k = 3$ at silhouette 0.086, and no label recovery. Instead, the clusters separate different behavioural styles of listening, such as stillness versus visible movement, which cut across engagement levels. Clustering thus recovers real structure, but that structure is behavioural *style*, not engagement, so the semi-supervised plan would mostly propagate style. We read this as motivating rather than negative: engagement is not the dominant axis of behavioural variation, which is why the rest of the paper predicts it with supervised models. The preference for two groups also anticipates §4.4, where the distinction the data supports is shown to be binary rather than the annotators’ three classes.

4.2 The Contribution of Alignment Features

Among the families we test, *alignment* features, which describe the listener only in relation to oth-

ers, give the largest improvement over a listener-only model. Table 1 reports the supervised factorial ablation: each feature family is stacked on the listener-only baseline under the leakage-free protocol described in §3.3. The baseline already clears the majority-class macro- F_1 baseline (macro- F_1 0.341 vs 0.220). The best addition is listener $\leftrightarrow$ speaker alignment, which raises macro- F_1 to 0.378, the highest in the table, and lifts accuracy to 0.505. It exceeds listener $\leftrightarrow$ co-participant alignment (0.326), the speaker’s own prosody (0.306), and the listener-only baseline itself. The ordering matches what the synchrony literature (§2) predicts: coordination with the person holding the floor carries more social signal than coordination with co-present non-speakers. It is an ordering of point estimates whose margins, quantified below, sit within fold-to-fold variance.

What gain there is comes from measuring the relation, not from giving the model more features. Listener $\leftrightarrow$ speaker alignment produces the largest improvement while adding only four dimensions to the baseline (166 $\rightarrow$ 170), namely the cosine similarity and correlation between the listener and the speaker on the visual (CLIP) and Action-Unit channels, whereas folding in the speaker’s raw prosody adds thirty-seven dimensions (166 $\rightarrow$ 203) and *lowers* macro- F_1 to 0.306. The blocks do not stack, either: the compact speaker-alignment on its own beats every richer combination (co-participant alignment + prosody 0.364, all three 0.337), because at $n = 89$ the extra dimensions cost more than they add. A four-number summary of how two people co-behave is the single best configuration we find, and that parsimony is what makes the relation usable at this sample size.

Accuracy changes little because the task is imbalanced. A classifier that always predicts the majority class scores accuracy 0.494 but macro- F_1 only 0.220; our balanced-weight models trade raw agreement for minority-class recall, so the listener-only baseline scores below that accuracy floor (0.459) even as its macro- F_1 rises, and adding speaker alignment reaches just 0.505. On a 44/28/17 split, gains in minority-class recall rather than raw agreement are exactly what we want, and the reason we read macro- F_1 as primary.

Across the five cross-validation folds, macro- F_1 already varies by about 0.09 simply from how the data happens to be split, so the differences we describe here are directional rather than significant: most fall within that natural fold-to-fold range.

Configuration	Macro- F_1	Acc.	Dim
Majority baseline	0.220	0.494	—
Listener only (baseline)	0.341	0.459	166
+ Co-participant align.	0.326	0.482	172
+ Speaker align.	0.378	0.505	170
+ Speaker prosody	0.306	0.483	203
+ Co-participant align. + prosody	0.364	0.517	209
+ All families	0.337	0.494	213

Table 1: Supervised factorial ablation on the three-class target, stacking interactional feature families on the listener-only baseline (the nonverbal family, A). Co-participant alignment and speaker alignment are the two relations of the alignment family (B); speaker prosody is family C. Cross-validated Random Forest macro- F_1 and accuracy (standard deviation across folds $\approx$ 0.09 throughout); *Dim* is the feature dimensionality.

A paired bootstrap over the held-out predictions (§3.3) makes this concrete: the 95% interval for the listener $\leftrightarrow$ speaker gain over the baseline runs from -0.06 to $+0.09$ and includes zero (one-sided $p \approx 0.26$), and every other family’s interval includes zero as well. The strength of the claim is the consistency of the pattern across metrics and configurations, not any single delta. Impurity importance is biased toward the high-dimensional CLIP block, about 0.63 of the total, simply because it contributes the most columns, and permutation importance is near zero for every feature at $n = 89$, so no single feature is reliably pivotal. Among the interpretable structural families, listener head-motion dynamics, yaw velocity and its variability, carry the most weight, consistent with orientation and attention being the readable part of a silent listener’s engagement. We therefore read the ablation, not the importance scores, as the evidence for what helps.

4.3 Speaker Prosody Partly Predicts Listener Engagement

The speaker side of the interaction also carries information about the listener’s state, concentrated narrowly rather than spread across measures. Of the many prosodic measures we extract, almost none help on their own: as a standalone family, speaker prosody is a weak predictor (macro- F_1 0.306), and voice quality and turn-taking context add nothing on top of pitch. Within pitch, nearly all of the engagement-relevant signal resides in a single measure, the *terminal slope*: the rate at which the speaker’s pitch rises or falls over the final half-second of voiced speech. We measure pitch in semitones relative to the window’s median,

which makes it comparable across speakers (§3.2), so the terminal slope is a value in semitones per second (st/s). Averaged within each engagement class, this slope rises for windows where the listener is *Engaged* (+1.5 st/s) and *Intermittent* (+2.6 st/s), but is flat or falling where the listener is *Disengaged* (−3.6 st/s): engaged listeners co-occur with rising speaker intonation, disengaged ones with falling intonation. Because the transcripts carry no punctuation, this question-like intonation is recoverable only from the acoustics. We read it as a speaker-side correlate that partly predicts the listener’s state: rising intonation is characteristic of questions, which typically invite a reaction, so a listener being addressed with a question is more likely to respond and appear engaged. Because the speaker addresses the group rather than the specific listener (§3.2), the cue indexes local interactional context rather than a signal aimed at the listener.

4.4 Label Scarcity Makes the Task Effectively Binary

The three-class results raise a prior question: is engagement simply hard to read from behaviour, or is the three-class *labelling* the obstacle? The middle *Intermittent* class is the prime suspect: it is where annotators disagree the most (§3.1), and the unsupervised exploration recovers only two natural groups (§4.1). If that ambiguous class is what holds performance down, merging it away should make the task markedly easier; if the features are simply too weak, no regrouping will improve the predictions. We therefore collapse the three labels into a binary contrast in three different ways, re-run the classification experiments, and compare each scheme against its own majority-class macro- F_1 floor (Table 2). The binary reductions differ only in their treatment of the *Intermittent* class. *Engaged vs. Non-active* groups *Intermittent* with *Disengaged*, while *Active vs. Disengaged* groups *Intermittent* with *Engaged* to form the *Active* class. *Engaged vs. Disengaged* excludes *Intermittent* from the evaluation and retains only high-confidence samples labelled as *Engaged* or *Disengaged* ($E \mid D$; $n = 61$).

The finding is clean: engagement is learnt more effectively as a binary *Engaged vs. Non-active* distinction. *Engaged vs. Non-active* reaches macro- F_1 0.634, +0.30 over its floor, the only wide, stable margin in the table. Every grouping that instead hinges on the *Disengaged* class collapses toward chance, whether the middle class

is merged with *Engaged* (*Active vs. Disengaged*, 0.500 vs 0.447) or dropped (*Engaged vs. Disengaged*, 0.450 vs 0.419); the full three-class task clears its floor only modestly (0.378 vs 0.220). The paired bootstrap agrees: only *Engaged vs. Non-active* has a margin over its floor whose 95% interval excludes zero ($[+0.02, +0.28]$, one-sided $p = 0.015$), whereas the intervals for the *Disengaged*-dependent schemes reach down to zero.

Label scheme	Grouping	Macro- F_1	Floor	Acc.
Three-class	$E \mid I \mid D$	0.378	0.220	0.505
Engaged vs. Non-act.	$E \mid I + D$	0.634	0.336	0.641
Active vs. Diseng.	$E + I \mid D$	0.500	0.447	0.820
Engaged vs. Diseng.	$E \mid D$	0.450	0.419	0.722

Table 2: Label-scheme sensitivity (listener-only + speaker-alignment features). Base classes: E = Engaged (44), I = Intermittent (28), D = Disengaged (17). *Grouping*: “|” splits different groups. *Floor* is the majority-class macro- F_1 baseline. Only *Engaged vs. Non-active* clears its floor by a wide margin; schemes requiring the *Disengaged* class sit at their floor.

Three readings converge: the construct (*Intermittent* was defined as the label for windows an annotator could not confidently call Engaged or Disengaged, §3.1), the unsupervised $k = 2$ structure (§4.1), and learnability (only the *Engaged vs. Non-active* cut yields a large, stable gain). The task is effectively binary here, bounded by the scarcity of confident *Disengaged* labels (17 of 89), not by the feature set.

5 Conclusion

We measured the engagement of silent meeting participants through behavioral *alignment*, the relation between the listener and the people they attend to, rather than the listener alone. On the GEHM corpus under leakage-free cross-validation, the listener $\leftrightarrow$ speaker relation gave the largest improvement over a listener-only model, in the direction predicted by the literature. The margin is within fold variance, so it is a directional result rather than an established effect, and the design cannot attribute it to coordination rather than to shared exposure to the same speaker. The firmer findings are structural: with few and ambiguous labels the three-class task is effectively binary, bounded by the scarcity of confident disengaged labels (17 of 89) rather than by the feature set, and the speaker’s terminal pitch slope remains an interpretable correlate of listener engagement.

Limitations

The clearest limitation is statistical: our supervised analysis rests on only 89 labelled windows, and every difference we report is comparable to or smaller than how much macro- F_1 already varies across cross-validation folds simply from how the data is split (roughly 0.09; §4), so no single result in Table 1 is statistically significant on its own. A paired bootstrap over the held-out predictions (§3.3) puts the 95% interval for the listener $\leftrightarrow$ speaker gain over the baseline at $[-0.06, +0.09]$, including zero. The listener $\leftrightarrow$ speaker alignment headline is a hypothesis supported by a consistent pattern, not an established effect size; replication on a larger labelled set is the obvious next step.

A second limitation is what the design can say about alignment at all. The alignment features are correlational and computed within a single five-second window, between two people participating in the same meeting. We therefore cannot separate the listener’s individual behaviour from their coordination with others in a causal sense. A high listener $\leftrightarrow$ speaker similarity is consistent with three accounts our features cannot distinguish: the listener is attending to and coordinating with the speaker; both are reacting independently to a third participant or a shared event; or the speaker’s behaviour drives the listener’s. Since alignment is measured from synchronic behaviours, we also cannot recover the lead-lag asymmetry that would begin to separate these accounts and that much of the synchrony literature relies on (Delaherche et al., 2012). Ablation gains should be read as incremental predictive value under a specific feature combination, not as evidence that coordination itself contributes engagement signal over individual behaviour. Separating the accounts would need lagged or directional measures, and ideally temporally resolved evidence of the kind used in dyadic work on shared attention (Wohltjen and Wheatley, 2021).

The labels are a third constraint: four non-expert annotators reached only modest agreement (Fleiss’ $\kappa = 0.42$), mostly on the middle *Intermittent* class, which was defined as a region of annotator uncertainty and behaves as one (§3.1, §4.4). The binding scarcity, however, is the 17 confident *Disengaged* windows; with so few examples we cannot tell whether disengagement is genuinely hard to read from behaviour or simply under-sampled.

The listener $\leftrightarrow$ speaker relation, our central re-

sult, is also measured under partial coverage: the active speaker is identified for only about 64% of windows, with the rest imputed to the training-fold median (§3.3). This keeps the protocol leakage-free but dilutes the measured contribution of the alignment and prosody families; a more robust speaker attribution, for instance from diarised audio, might sharpen the signal we observe.

Every feature family is also computed strictly within the target window, with no representation of the listener’s own behaviour in earlier windows beyond the small amount of turn-taking history already in the speaker features (§3.2). Because engagement plausibly persists or drifts across turns rather than resetting every five seconds, a short history of the listener’s own recent behaviour might recover signal that a single window cannot.

Finally, the findings are tied to one setting: GEHM consists of remote, gallery-view Zoom meetings in English, where the mutual-gaze and proxemic cues of co-present interaction are absent or distorted, so the alignment and prosody patterns we report may not transfer to in-person meetings, other languages, or different recording conditions. We also treat the speaker’s prosody as concurrent context rather than a signal directed at the listener, since the speaker generally addresses the group; the terminal pitch slope is therefore an interpretable correlate (§4), not a validated causal cue.

References

- Tadas Baltrušaitis, 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)*, pages 59–66. IEEE.
- Frank J. Bernieri, J. Steven Reznick, and Robert Rosenthal. 1991. The rhythm of rapport: Interpersonal synchrony and social perception. *Journal of Personality and Social Psychology*, 60(2):243–253.
- Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 <http://www.praat.org/>.
- Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. *OpenPose: Realtime multi-person 2D pose estimation using part affinity fields*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

- Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. 2012. [Interpersonal synchrony: A survey of evaluation methods across disciplines](#). *IEEE Transactions on Affective Computing*, 3(3):349–365.
- M. Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. 2019. [Engagement detection in online learning: A review](#). *Smart Learning Environments*, 6(1):1–20.
- Joseph L. Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76(5):378–382.
- Maria Frank, Ghassem Tofghi, Haisong Gu, and Renate Fruchter. 2016. [Engagement detection in meetings](#). *Preprint*, arXiv:1608.08711.
- Jennifer A. Fredricks, Phyllis C. Blumenfeld, and Alison H. Paris. 2004. [School engagement: Potential of the concept, state of the evidence](#). *Review of Educational Research*, 74(1):59–109.
- Daniel Gatica-Perez, Iain McCowan, Dong Zhang, and Samy Bengio. 2005. [Detecting group interest-level in meetings](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 489–492. IEEE.
- Agustín Gravano and Julia Hirschberg. 2011. [Turn-taking cues in task-oriented dialogue](#). *Computer Speech & Language*, 25(3):601–634.
- Abhay Gupta, Arjun D’Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. [DAiSEE: Towards user engagement recognition in the wild](#). arXiv preprint arXiv:1609.01885. *Preprint*, arXiv:1609.01885.
- Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. 2018. [Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue](#). *APSIPA Transactions on Signal and Information Processing*, 7:1–16.
- Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. [Introducing Parselmouth: A Python interface to Praat](#). *Journal of Phonetics*, 71:1–15.
- Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. 2018. [Prediction and localization of student engagement in the wild](#). In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE.
- Rivka Levitan, Agustín Gravano, and Julia Hirschberg. 2011. [Entrainment in speech preceding backchannels](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 113–117, Portland, Oregon, USA. Association for Computational Linguistics.
- Rivka Levitan, Agustín Gravano, Laura Willson, Štefan Beňuš, Julia Hirschberg, and Ani Nenkova. 2012. [Acoustic-prosodic entrainment and social behavior](#). In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19, Montréal, Canada. Association for Computational Linguistics.
- Jia Li, Yangchen Yu, Yin Chen, Yu Zhang, Peng Jia, Yunbo Xu, Ziqiang Li, Meng Wang, and Richang Hong. 2024. [DAT: Dialogue-aware transformer with modality-group fusion for human engagement estimation](#). In *Proceedings of the 32nd ACM International Conference on Multimedia (MM ’24)*. Association for Computing Machinery.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, and 1 others. 2019. [MediaPipe: A framework for building perception pipelines](#). *arXiv preprint arXiv:1906.08172*.
- Philipp Müller, Michal Balazsia, Tobias Baur, Michael Dietz, Alexander Heimerl, Anna Penzkofer, Dominik Schiller, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2024. [MultiMediate’24: Multi-domain engagement estimation](#). In *Proceedings of the 32nd ACM International Conference on Multimedia (MM ’24)*, pages 11377–11382. Association for Computing Machinery.
- Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. [Detecting low rapport during natural interactions in small groups from non-verbal behaviour](#). In *Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI ’18)*, pages 153–164. Association for Computing Machinery.
- Catharine Oertel, Cécile De Looze, Stefan Scherer, Andreas Windmann, Petra Wagner, and Nick Campbell. 2011. [Towards the automatic detection of involvement in conversation](#). In *Analysis of Verbal and Non-verbal Communication and Enactment. The Processing Issues: COST 2102 International Conference*, pages 163–170. Springer.
- Catharine Oertel, Kenneth A. Funes Mora, Joakim Gustafson, and Jean-Marc Odobez. 2015. [Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions](#). In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI ’15)*, pages 107–114, New York, NY, USA. Association for Computing Machinery.
- Shay Ohayon and Ilanit Gordon. 2025. [Multimodal interpersonal synchrony: Systematic review and meta-analysis](#). *Behavioural Brain Research*, 480:115369.
- Patrizia Paggio, Manex Agirrezabal, Giulia Di Cristina, Bart Jongejan, and Costanza Navarretta. 2026. [Multimodal entrainment and feedback in online group](#)

- meetings. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9139–9149, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Patrizia Paggio, Manex Agirrezabal, Costanza Navarretta, and Leo Vitasovic. 2024. [Multimodal behaviour in an online environment: The GEHM zoom corpus collection](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11890–11900, Torino, Italia. ELRA and ICCL.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Armand Pellet-Rostaing, Ronan Bertrand, Adrien Boudin, Stéphane Rauzy, and Philippe Blache. 2023. [A multimodal approach for modeling engagement in conversation](#). *Frontiers in Computer Science*, 5.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.
- Monisha Singh, Ximi Hoque, Donghuo Zeng, Yanan Wang, Kazushi Ikeda, and Abhinav Dhall. 2023. [Do i have your attention: A large scale engagement prediction dataset and baselines](#). In *Proceedings of the 25th International Conference on Multimodal Interaction, ICMI 2023*, page 174–182. Association for Computing Machinery.
- Sophie Wohltjen and Thalia Wheatley. 2021. [Eye contact marks the rise and fall of shared attention in conversation](#). *Proceedings of the National Academy of Sciences*, 118(37):e2106645118.