"Mm, Wat?" Detecting Other-intiated Repair Requests in Dialogue

Anonymous ACL submission

Abstract

Maintaining mutual understanding is a key component in human-human conversation to avoid conversation breakdowns, in which repair, particularly Other-Initiated Repair (OIR, when one speaker signals trouble and prompts the other to resolve), plays a vital role. However, Conversational Agents (CAs) still fail to recognize user-initiated repair requests, leading to breakdowns or disengagement. This work proposes a multimodal approach to automatically detect OIR requests in Dutch dialogues by integrating linguistic and prosodic features grounded in Conversation Analysis. The results show that prosodic cues complement linguistic features and significantly improve the results of pre-trained text and audio embeddings, offering insights into how different features interact. Future directions include incorporating visual cues, exploring large language models (LLMs), and applying the model in CA systems.

1 Introduction

Conversational agents (CAs), software systems that interact with users via natural language in written or spoken form, are increasingly used in multiple domains such as commerce, healthcare, and education (Allouch et al., 2021). While maintaining smooth communication is crucial in these settings, current state-of-the-art (SOTA) CAs still struggle with handling conversational breakdowns. Unlike humans, who rely on conversational repair to resolve issues like mishearing or misunderstanding (Schegloff et al., 1977; Schegloff, 2000), CAs' repair capabilities remain limited and incomplete. Schegloff (2000) categorized repair types based on who initiates and who resolves the problem, distinguishing between self- (by the speaker who caused the issue) and **other-initiated repair (by the recipient who detects it)**, which is our focus in this work. Current CAs handle repairs in a limited fashion that mainly support agent-initiated repair (e.g., the agent asks users to repeat what they said) (Li et al., 2020; Cuadra et al., 2021; Ashktorab et al., 2019) or rely on user self-correction when they realize troubles and clarify their intent (e.g., saying "no, I mean…") (Balaraman et al., 2023). However, other-initiated self-repaired or in short other-initiated repair (OIR), where the user signals a problem and prompts the agent to clarify or correct itself, is rarely supported, while effective communication requires bidirectional (Moore et al., 2024). Supporting this, Gehle et al. (2014) found that museum guide robots failing to resolve communication issues quickly caused user disengagement, and van Arkel et al. (2020) showed that basic OIR mechanisms improve communicative success while reducing computational and interaction costs compared to relying on pragmatic reasoning.

Modeling OIR strategies on CAs that recognize user-initiated repair first requires robust automatic OIR request detection in human-human interaction. However, prior work is narrow and mostly text-based approaches, training on English corpora and relying on lexical cues (Höhn, 2017; Purver et al., 2018; Alloatti et al., 2024), which overlook prosodic markers that reliably signal repair. Prosodic cues tend to be more cross-linguistically stable than surface forms (Dingemanse and Enfield, 2015; Benjamin, 2013; Walker and Benjamin, 2017), and can provide valuable insight into the pragmatic functions of expressions like the interjection "huh". This highlights the limitations of relying solely on textual patterns for OIR request detection. Finally, understanding the OIR sequence also requires examining the local sequential environment of the surrounding turns, which we call a "dialogue micro context" (Schegloff, 2000).

These gaps motivate our main research question: **What are the multimodal indicators of OIR requests in human dialogue and how can we model them?** To address this, we analyze OIR sequences in a Dutch task-oriented corpus, focusing on text

and audio patterns where one speaker initiates an OIR request. Drawing on Conversation Analysis literature, we introduce feature sets and a computational model to detect such requests. Our contributions are in two folds: (1) a novel multimodal model for OIR request detection that integrates linguistic and prosodic features extracted automatically based on the literature, advancing beyond text- or audio-only approaches; (2) provide insights into how linguistic and prosodic features interact and contribute in OIR requests detection, grounded in Conversation Analysis, and what causes model misclassifications. The remainings of this paper is structured as follows: Section 2 reviews SOTA computational models for OIR request detection and related dialogue understanding tasks. Section 3 provides the used OIR coding schema and typology, and Section 4 details our approach, including linguistic and prosodic feature design. Section 5 presents our experiment details and results, followed by error analysis in Section 6.

2 Related Work

An early approach to automatic OIR detection was proposed by Höhn (2017), with a pattern-based chatbot handling user-initiated repair in text chats between native and non-native German speakers. Purver et al. (2018) extended this by training a supervised classifier using turn-level features in English, including lexical, syntactic, and semantic parallelism between turns. More recently, Alloatti et al. (2024) introduced a hierarchical tag-based system for annotating repair strategies in Italian task-oriented dialogue, distinguishing between utterance-specific and context-dependent functions.

Although direct research on OIR detection is still limited, advances in related dialogue understanding tasks provide promising methods for our work. Miah et al. (2024) combined pretrained audio (Wav2Vec2) and text (RoBERTa) embeddings to detect dialogue breakdowns in healthcare calls. Similarly, Huang et al. (2023) used BERT, Wav2Vec2.0, and Faster R-CNN for intent classification, introducing multimodal fusion with attention-based gating to balance modality contributions and reduce noise. Saha et al. (2020) proposed a multimodal, multi-task network jointly modeling dialogue acts and emotions using attention mechanisms. Liu et al. (2023) achieved SOTA in several tasks with a hierarchical model leveraging special tokens and turn-level attention. More recently, high-performing but more opaque and resource-intensive approaches have emerged: Chen et al. (2024) applied prompt-based learning with intent templates to enhance cross-modal alignment for intent detection, and Mohapatra et al. (2024) showed that larger LLMs outperform smaller ones on tasks like repair and anaphora resolution, albeit with higher computational cost and latency.

Despite robust performance, recent largest models remain difficult to interpret due to their black-box nature and multimodal fusion complexity (Jain et al., 2024). To address this gap, we propose a computational model for OIR request detection in Dutch that fuses pretrained text and audio embeddings with linguistic and prosodic features **grounded in Conversation Analysis**. The model also integrates a multihead attention mechanism to weigh and capture non-linear relationships across modalities, allowing our model to keep the strengths of multimodal deep learning while **offering insight from linguistic and prosodic features to intepret their interaction and impact** towards model's decision.

3 OIR Coding Schema and Typology

We follow Dingemanse and Enfield (2015)'s coding schema, which structures OIR sequences into three components: trouble source, OIR request, and repair solution segments, in which OIR requests are categorized into three types: *open request* (the least specific, not giving clues of trouble), *restricted request* (implied trouble source location), and *restricted offer* (the most specific, proposing a candidate understanding). Following Rasenberg et al. (2022)'s OIR annotation, which aligns OIR component boundaries with Turn Construction Unit (TCU, the smallest meaningful element of speech such as a word, phrase or sentence, that can potentially complete a speaker turn) boundaries in speech annotation, we use the term segment as the unit for input data. An OIR request segment may comprise one or multiple TCUs (as in Figure 1), serving as our data input units.

There are two OIR sequence types: *minimal* (OIR request initiated immediately after the turn containing the trouble source segment), and *complex* (OIR request delayed by a few turns), as illustrated in Figures 1 and 2.

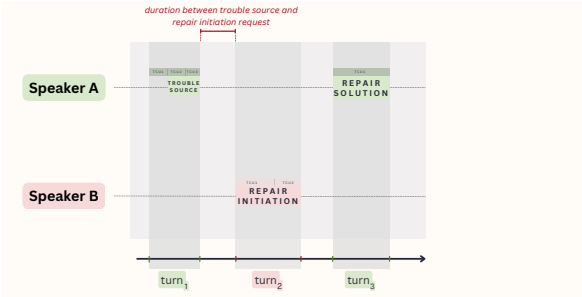

Figure 1: Visualization of a minimal OIR sequence

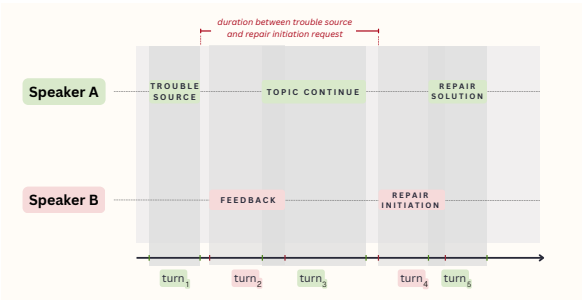

Figure 2: Visualization of a complex OIR sequence

4 Proposed Approach

4.1 Overview

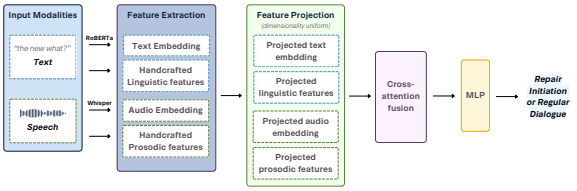

Figure 3: Overview our multimodal modal architecture

Task Formulation. We formulate the OIR request detection as a binary classification problem. Given a segment (x_i), corresponding to one or several TCUs within a speaker turn, the task is to predict whether it is an OIR request or a regular dialogue (RD) segment (*i.e.*, not belonging to an OIR sequence).

Architecture Overview. Figure 3 shows the overview of our proposed approach. For a given segment (x_i), we extract the linguistic and prosodic features respectively, then integrate text and audio embeddings extracted from pretrained model. All features are then projected to a shared dimensionality to ensure the consistency across modalities. To capture the complex interactions between text and audio embeddings with handcrafted features, a multihead attention mechanism was employed to weigh and capture non-linear relationships. Finally, the whole representation is obtained by concatenating the text embedding and the fused representation from multihead attention. We propose a multimodal approach to introduce the handcrafted linguistic and prosodic features, automatically computed based on literature review, into the pretrained models' embeddings to model the OIR request.

4.2 Pretrained Models

Language model. Our proposed approach utilizes RoBERTa (Zhuang et al., 2021), a transformer-based language model, to obtain text embedding of the current given segment. As our corpus is in Dutch, we use the pre-trained RobBERT (Delobelle et al., 2020) model, which is based on the RoBERTa architecture, pre-trained with a Dutch tokenizer, and 39 GB of training data. We use the latest release of *RobBERT-v2-base* model which pre-trained on Dutch corpus OSCAR 2023 version, which outperforms other BERT-based language models for several different Dutch language tasks.

Audio model. For audio representations, we utilize Whisper (Radford et al., 2023), an encoder-decoder Transformer-based model trained on 680,000 hours of multilingual and multitask speech data, to extract audio embeddings from our dialogue segments. Whisper model stands out for its robustness in handling diverse and complex linguistic structures, a feature that is crucial when dealing with Dutch, a language known for its intricate syntax. Besides, Whisper was trained on large datasets including Dutch and demonstrated good performance in zero shot learning, making it ideal serving as a naive baseline for task with small corpus like ours.

4.3 Dialogue Micro Context

Schegloff (2000) demonstrated that the OIR sequence is systematically associated with multiple organizational aspects of conversation, and understanding an OIR request requires examining the local sequential environment, which we call in this work the dialogue micro context (Schegloff, 1987). Therefore, for each given target segment (x_i), to capture the micro context, we iteratively concatenate the previous (x_{i-j}) and following (x_{i+j}) TCUs using special separator token of transformers (e.g. </s> for RoBERTa-based models) until reaching the maximum token limit (excluding [CLS] and [EOS]), inspired by similar

ideas in (Wu et al., 2020; Kim and Vossen, 2021). If the sequence exceeds the limit, we truncate the most recently added TCUs. The final sequence is enclosed with [CLS] and [EOS], as shown in Figure 8.

4.4 Linguistic Feature Extraction

Figure 4 outlines our linguistic feature set for the representation of the target segment, capturing local properties such as part-of-speech (POS) tagging patterns, question formats, transcribed non-verbal actions (target segment features), and features, which quantify repetition and coreference across turns to reflect backward and forward relations around the OIR request (cross-segment feature to capture micro context). The detailed description is in the Appendix B.

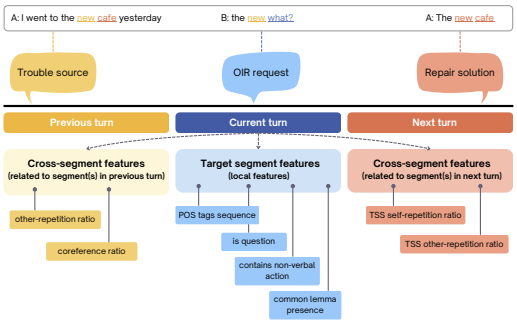

Figure 4: Visualization linguistic feature set

4.4.1 Target Segment Features

We automatically extracted the linguistic features proposed by (Ngo et al., 2024) at the intra-segment level to capture grammatical and pragmatic patterns related to the OIR request. For instance, restricted OIR requests often show a POS tag sequence pattern of interrogative pronouns followed by verbs, while OIR open requests and regular dialogue segments differ in key lemmas used of the same tag: modal auxiliary verb kunnen ("can") vs. primary auxiliary verb zijn ("to be"). Additional features include question mark usage and binary indicators for non-verbal actions (e.g., laughing, sighing) (Schegloff, 2000), are fully given in the Appendix B.

4.4.2 Cross-Segment Features

Grounded on the literature (Schegloff, 2000; Ngo et al., 2024), we define inter-segment features that capture the sequential dynamics of the OIR request, including repetitions and the use of coreferences referring to entities in prior turns containing the trouble source segment. We also compute self and other-repetition in the subsequent turn containing the repair solution segment, to capture how the trouble source speaker responds. These features reflect the global dynamics of OIR sequences.

4.5 Prosodic Features Extraction

Prosody plays a crucial role in signaling OIR requests. Previous studies in Conversation Analysis show that pitch, loudness, and contour shape can indicate whether a repair initiation is perceived as "normal" or expresses "astonishment"(Selting, 1996), and that Dutch question types differ in pitch height, final rises, and F0 register (Haan et al., 1997). Building upon these characteristics, we design a prosodic feature set that includes both local features within the target segment, such as pitch, intensity, pauses, duration, and word-level prosody, and global features across segments of the OIR sequence, such as latency between OIR sequence segments, pitch slope transitions at boundaries, and comparison to speaker-specific prosodic baselines. The features are detailed in Figure 5 and in the Appendix C.

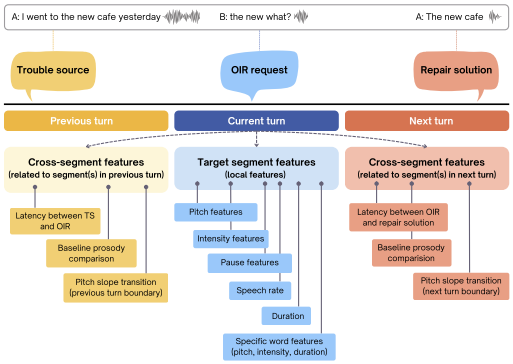

Figure 5: Visualization of prosodic feature set

4.5.1 Target Segment Features

We use Praat (Boersma, 2000) to extract prosodic features at the segment level, including: pitch features (e.g., min, max, mean, standard deviation, range, number of peaks) which are computed from voiced frames after smoothing and outlier removal, with pitch floor/ceiling set between 60–500 Hz and adapted to each speaker range (van Bezooijen, 1995; Theelen, 2017; Verhoeven and Connell, 2024); first (mean and variability of pitch slope change) and second derivatives (pitch acceleration) of pitch contour, capturing pitch dynamics. Additional features are intensity (e.g., min, max, mean,

range, standard deviation), and voice quality measures (jitter, shimmer, and harmonics-to-noise ratio). We also model pause-related features by detecting silent pauses over 200 ms and categorizing them by duration and position in the utterance, reflecting their conversational function associated with repair possibilities (van Donzel and Beinum, 1996; Hoey, 2018). Inspired by findings prosody of other-repetition in OIR request (Dingemanse et al., 2015; Walker and Benjamin, 2017), we extract pitch and intensity features for repeated words from the trouble source segment, and for the specific repair marker "wat" (what/which/any), as indicators of OIR request type and speaker perspective (Huhtamäki, 2015).

4.5.2 Cross-Segment Features

To model the speaker-specific prosodic variation (van Bezooijen, 1995; Theelen, 2017; Verhoeven and Connell, 2024), we normalize pitch and intensity using z-scores, relative percentage change, and position within the speakers' range. These features capture how far the current segment deviates from the speaker's typical behaviour across previous turns and the normalized range position of the current segment within the speaker's baseline. Inspired by work on prosodic entrainment (Levitan and Hirschberg, 2011), we also compute pitch and intensity slope transitions across segment boundaries (e.g., TS→OIR, OIR→RS), both within and across speakers, to assess prosodic alignment. We normalized slopes to semitones per second for consistency across speakers.

5 Experiments & Results

To answer the main research question mentioned in Section 1, we design the experiments to answer the following research sub-questions: *i)* **RQ1**: To what extent do audio-based features complement text-based features in identifying OIR requests? *ii)* **RQ2**: Do our proposed linguistic and prosodic features (see Figures 4 and 5) perform better than pretrained embeddings? *iii)* **RQ3**: Which prosodic and linguistic features contribute the most to OIR request detection? *iv)* **RQ4**: How does the involvement of dialogue micro context affect OIR request detection performance?

5.1 Implementation Details

Dataset. Based on (Colman and Healey, 2011)'s findings that repair occurs more frequently in task-oriented dialogues, we selected a Dutch multimodal task-oriented corpus (Rasenberg et al., 2022; Eijk et al., 2022), which contains 19 dyads collaborating on referential communication tasks in a standing face-to-face setting. Participants alternated roles to describe (Director) and identify (Matcher) geometric objects ("Fribbles") displayed on screens. The unconstrained design encouraged natural modality use and OIR sequences. Rasenberg et al. (2022) annotated OIR sequences using Dingemanse and Enfield, 2015's schema, resulting in 10 open requests, 31 restricted requests, and 252 restricted offers. We balanced the dataset with 306 randomly selected regular dialogue segments, stratified across all dyads. The high distribution of restricted offers likely originates from the task settings, where participants see all 16 candidate objects, prompting them to offer a candidate of understanding and ask for confirmation.

Training Details. We fine-tuned our models using 10-fold cross-validation, in which the optimal learning rate was 2e-5. We employed AdamW optimizer with a weight decay of 0.01 and a learning rate scheduler with 10% warmup steps. Training ran for up to 20 epochs with 3-epoch early stopping patience, and batch size 16.

Evaluation Metrics. We evaluated model performance using binary classification metrics including precision, recall, and macro F1-score.

5.2 Experiment Scenarios & Results Analysis

Model	Modal & Features	Precision	Recall	F1-score
Text_{Emb}	U & T	72.0 ± 4.0	87.6 ± 7.5	78.9 ± 4.7
$\text{Audio}_{\text{Emb}}$	U & A	72.6 ± 9.7	76.3 ± 13.1	70.6 ± 8.1
$\text{Multi}_{\text{Emb}}$	M & T+A	79.1 ± 5.4	82.2 ± 3.8	82.1 ± 0.9
$\text{Text}_{\text{Ling}}$	U & L	82.2 ± 3.6	80.4 ± 6.1	80.4 ± 3.8
$\text{Audio}_{\text{Pros}}$	U & P	81.7 ± 4.2	77.4 ± 5.4	77.3 ± 2.7
$\text{Multi}_{\text{LingPros}}$	M & L+P	81.7 ± 7.6	82.2 ± 1.5	81.8 ± 3.4
$\text{Multi}_{\text{Ours}}$	M & T+A+L+P	$\mathbf{93.2 \pm 2.8}$	$\mathbf{96.1 \pm 2.6}$	$\mathbf{94.6 \pm 2.3}$

U: Unimodal, **M**: Multimodal, **T**: Text, **A**: Audio, **P**: Prosodic features, **L**: Linguistic features

Table 1: Overall results across modalities for OIR request detection. The table groups models by research question: **RQ1** compares unimodal vs. multimodal combinations of audio and text; **RQ2** compares handcrafted features with pretrained embeddings.

RQ1: Audio vs. Text Complementarity. To address RQ1, we compare the performance of unimodal against multimodal models, including: *i)* Single $\textbf{Text}_{\textbf{Emb}}$ or $\textbf{Audio}_{\textbf{Emb}}$ vs. $\textbf{Multi}_{\textbf{Emb}}$; *ii)* Single $\textbf{Text}_{\textbf{Ling}}$ or $\textbf{Audio}_{\textbf{Pros}}$ *vs.* $\textbf{Multi}_{\textbf{LingPros}}$. We want here to see if adding the audio-based

features, either by pretrained embeddings or by using handcrafted prosodic features, will improve the performance of the text-based models. The multimodal models include **Multi$_{\text{Emb}}$**, which fuses pretrained text and audio embeddings, and **Multi$_{\text{LingPros}}$**, which combines handcrafted linguistic and prosodic features, using cross-attention fusion as illustrated in Figure 3.

From Table 1, we observe that multimodal models consistently outperform unimodal ones across all metrics. For both pretrained embeddings and handcrafted features, text-based models outperform audio-based ones individually. However, incorporating audio improves performance in both settings. Specifically, in the pretrained setting, the multimodal model **Multi$_{\text{Emb}}$** achieves an F1-score of 82.1, improving over **Text$_{\text{Emb}}$** by 3.2 percentage points (pp) and over **Audio$_{\text{Emb}}$** by 11.5 pp. Similarly, in the handcrafted feature setting, combining linguistic and prosodic features **Multi$_{\text{LingPros}}$** yields an F1 of 81.8, outperforming Text$_{\text{Ling}}$ by 1.4 pp and **Audio$_{\text{Pros}}$** by 4.5 pp. Interestingly, the unimodal handcrafted models **Text$_{\text{Ling}}$**, **Audio$_{\text{Pros}}$** show higher precision than recall, whereas **Multi$_{\text{LingPros}}$** shows slightly higher recall, suggesting a tendency to favor detection over omission. This is potentially beneficial in interactive systems where missing an OIR request could be more disruptive than a false alarm. For embedding-based models, recall exceeds precision in all cases, but the multimodal model shows a notable gain in precision, indicating a better trade-off between identifying true OIR requests and minimizing false positives.

RQ2: Handcrafted Features vs. Pretrained Embeddings. To address RQ2, we compare the performance of models using handcrafted features against the models using embeddings from pretrained models. We thus compare: *i)* Text representations: text embeddings (**Text$_{\text{Emb}}$**) *vs.* handcrafted linguistic features (**Text$_{\text{Ling}}$**); *ii)* Audio representations: audio embeddings (**Audio$_{\text{Emb}}$**) *vs.* handcrafted prosodic features (**Audio$_{\text{Pros}}$**); *iii)* Combined approaches: multimodal models using pretrained embeddings (**Multi$_{\text{Emb}}$**) *vs.* using handcrafted linguistic and prosodic features (**Multi$_{\text{LingPros}}$**) and *vs.* our proposed approach leveraging both of them **Multi$_{\text{Ours}}$**.

We want here to see if the sets of handcrafted features grounded by literature in Conversation Analysis performed better or complement the pretrained models' embeddings. Results in Table 1 demonstrate that handcrafted feature models are comparable to embedding-based approaches. In unimodal settings, Text$_{\text{Ling}}$ achieves higher precision (+10 pp) with comparable F1-score (+1.5 pp) to **Text$_{\text{Emb}}$**, despite lower recall (-7.2 pp). Likewise, **Audio$_{\text{Pros}}$** outperforms **Audio$_{\text{Emb}}$** across all metrics (precision +9.1 pp, recall +1.1 pp, F1-score +6.7 pp). For multimodal approaches, **Multi$_{\text{Emb}}$** and **Multi$_{\text{LingPros}}$** perform almost identically (F1-score difference of just 0.3 pp), with handcrafted features providing better balanced precision-recall trade-offs. Furthermore, our proposed Multi$_{\text{Ours}}$ model, combining pretrained embeddings with handcrafted features from both modalities, substantially outperforms all other approaches, raising F1-score by 12.5 pp, precision by 14.1 pp, and recall by 13.9 pp, which suggests that handcrafted features effectively complement pretrained embedding models, with more balanced trade-off between False Positives (regular dialogues misclassified as OIR requests) and False Negatives (OIR requests misclassified as regular dialogue).

RQ3: Handcrafted Feature Importance Analysis. Although the linguistic and prosodic features could not solely outperform pretrained text and audio embeddings, they are useful in interpreting the model's behaviours, especially to see if they are aligned with the Conversation Analysis findings. To answer RQ3, we used SHAP (SHapley Additive exPlanations) analysis to analyse the contribution and behaviours of linguistic and prosodic features towards the model's decision. Figure 6 illustrates the top 20 features by SHAP value, which measures how much each single feature pushed the model's prediction compared to the average prediction. The pausing behaviours (positions and durations), intensity measures (max, mean, and relative change), and harmonic-to-noise ratio (HNR) appear particularly important among prosodic features. For linguistic features, the grammatical structure linking to coreference used, some POS tags, and various word type ratios rank highly. The most important features include the number of long and medium pauses, the relative position of the longest pause, and the verb-followed-by-coref structure, all scoring near 1.0 on the importance scale, which aligned with the works in (Hoey, 2018; Ngo et al., 2024) about pauses in OIR requests and the structure of OIR request, respectively.

Figure 7 displays the synergy (Ittner et al., 2021)

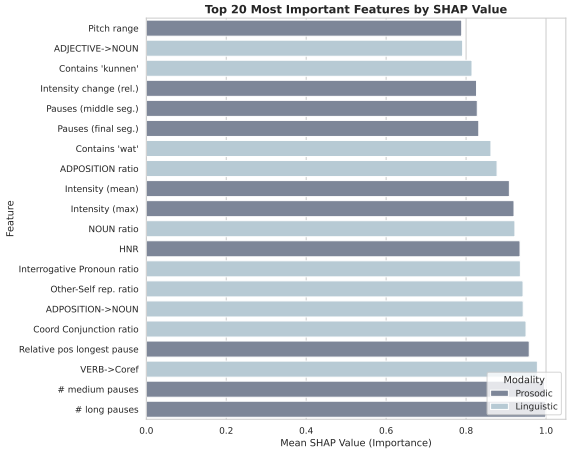

Figure 6: Visualization feature importance

between linguistic and prosodic features, which are computed based on the SHAP interaction values. It reflects how complementary a pair of linguistic and prosodic features is in improving model performance, in which high synergy means that combining both features adds more value than what each of them contributes individually. These features do not always need to co-vary, but their combination brings useful information for the model. Coordinating conjunction ratio (CCONJ ratio) shows the strongest synergy (0.26) with harmonics-to-noise ratio (HNR), while other speaker self-repetition ratio has strong synergy (0.23) with maximum intensity. This suggests that certain grammatical patterns work closely with specific voice qualities, particularly how conjunctions interact with voice clarity and how self-repetition correlates with voice intensity. The results indicate that conversation involves a complex interplay between what we say (linguistic elements) and how we say it (prosodic elements), which is aligned with the Conversation Analysis work.

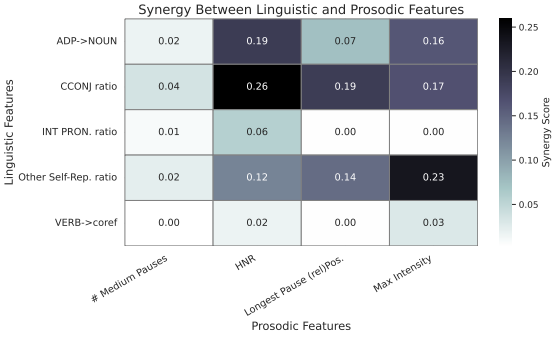

Figure 7: Visualization feature interaction heatmap

RQ4: Dialogue Micro Context Analysis. To address RQ4, we experimented 4 scenarios of concatenating micro context, including: (1) **Past$_{\text{Context}}$** - concatenated current input segment with the TCUs in the prior turns and cross-segment handcrafted features (past-related, Figure 4, 5); (2) **Future$_{\text{Context}}$** - concatenated current input segment with the TCUs in the subsequent turns and handcrafted cross-segment features (future-related, Figure 4, 5); (3) **Current$_{\text{Context}}$** - no context concatenation and used only current input segment features (Figure 4, 5); (4) **Multi$_{\text{Ours}}$** - the full context scenario, where we concatenate current input segment with both the prior and subsequent TCUs and use full handcrafted feature set. For (1) and (4), we experimented with window_length of 2 and max (the micro context are concatennated as much as possible until it reach maximum token limit) based on results from corpus analysis; for (2) only max was used, as repair solutions typically occur immediately within maximum 2 turns in this corpus. Table 2 highlights the impact of different micro-context configurations, in which integrating surrounding TCUs from prior, and subsequent segments combining with the whole handcrafted feature set leads to the best overall performance, as also stated in Table 1. Notably, our this full context setting with smaller window_length=2 achieves the highest results across all metrics, while introducing to the maximum allowed token limits degrades the performance, with a drop of approximately 6.3 pp of F1-score, 9 pp of precision, and 4.1 pp of recall. It suggests that while surrounding context of input segment is helpful, overly long concatenation may introduce noise and irrelevant information to model. Besides, integrating past or current segments yields moderate performance, with F1-scores ranging from approximately 80.2% to 83.6%, while future context alone results in the lowest scores, indicating that the upcoming dialogue can offer informative cues but less relevant than the prior and current input segments, which aligned with the nature of OIR sequence.

Context	Window length	Precision	Recall	F1-score
(1) Past$_{\text{Context}}$	2	86.0 ± 3.0	78.4 ± 5.4	82.0 ± 4.1
(1) Past$_{\text{Context}}$	max	86.6 ± 5.2	81.0 ± 6.1	83.5 ± 4.3
(2) Current$_{\text{Context}}$	-	84.6 ± 3.8	82.9 ± 6.0	83.6 ± 4.4
(3) Future$_{\text{Context}}$	max	84.00 ± 1.53	78.20 ± 5.78	80.18 ± 2.52
(4) Multi$_{\text{Ours}}$	2	$\mathbf{93.2 \pm 2.8}$	$\mathbf{96.1 \pm 2.6}$	$\mathbf{94.6 \pm 2.3}$
(4) Multi$_{\text{Ours}}$	max	87.7 ± 3.5	89.1 ± 5.3	88.3 ± 3.7

Table 2: Dialogue micro context concatenation results

6 Error Analysis

In order to better interprete the model's performance, we analyze the False Negative (Fn) instances, which are the OIR requests that were misclassified as regular dialogue segments, to see whether there are common patterns in these instances that our models find it hard to predict, illustrated in Table 3. We compare the FN instances across our proposed multimodal model with the 2 unimodal models using linguistic and prosodic features. The results show that our model fails on only around 3.8% of the test set, while the FN errors accounted for around 15% and 24% on $\text{Text}_{\text{Ling}}$ and $\text{Audio}_{\text{Pros}}$, respectively. For the $\text{Text}_{\text{Ling}}$ model, it seems to struggle in detecting samples with vague references, especially in restricted offer type, even when there are OIR syntactic forms like *question mark* presented. In terms of $\text{Audio}_{\text{Pros}}$, even though the important prosodic cues were presented, the model seems to over-rely on pause structure and pitch contour. Short declaratives with flat intonation were often misclassified, suggesting the impact of lacking syntactic form information in this model. Finally, our proposed multimodal failed with mostly short phrases and subtle prosodic signals, which are not strongly marked as an OIR request. Considering the error across 3 types of OIR requests, it seems that only $\text{Audio}_{\text{Pros}}$ struggled with varied types of OIR requests; the other 2 models misclassified on restricted offer and open request instances only. However, as this corpus is imbalanced between the 3 types of OIR requests, with the majority of restricted offers, it could be the reason.

Modal	%error	Samples	Patterns	OIR Request Type
$\text{Text}_{\text{Ling}}$	15%	*(or a) triangle*	Vague, elliptical reference	Restricted Offer
		yes uh yes on the right side right? or ascending yes	Disfluencies, vague interrogative	Restricted Offer
		yes the one with the protrusion	Referential expression, lacks direct marker	Restricted Offer
$\text{Audio}_{\text{Pros}}$	24%	*with a sunshade*	Short declarative, flat prosody	Restricted Offer
		uh but the platform sits that cuts the	Flat intonation, mostly short pauses in the beginning	Restricted Offer
		Is it vertical?	Question intonation, has few short pauses	Restricted Offer
		uh and is his arm uh round but also a bit with angles?	High pitch, question intonation, pauses in beginning and middle, but complex OIR	Restricted Offer
		but what did you say at the beginning?	Rising intonation, wide pitch range, high pitch	Restricted Request
$\text{Multi}_{\text{Ours}}$	3.8%	*with a sunshade*	Short, declarative structure	Restricted Offer
		oh who so	Short declarative, high but flat pitch, no rising intonation	Restricted Offer
		sorry again?	Clear OIR but subtle prosodic signal	Open Request

Table 3: Samples of False Negative (FN) instances from unimodal (text/audio) and multimodal models, with qualitative patterns.

7 Conclusion & Future Works

This work presents a new approach to model and detect OIR requests in human-human conversation that takes advantage of automatically extracted linguistic and prosodic features and is based on studies of OIR sequences in Conversation Analysis. Our results show that the participation of these hand-crafted features significantly improves the performance of pretrained embeddings. We also show that the audio modality complements and improves the performance of the textual modality, either by using pre-trained embeddings or handcrafted linguistic and prosodic features.

In addition, our feature analysis revealed not only the impact of each feature in the linguistic and prosodic feature set individually, but also their complementary contribution. While key prosodic indicators include pause-related features, intensity, and harmonic-to-noise ratio (HNR), influential linguistic features involve grammatical patterns, specific POS tags, and lemma ratios. Furthermore, synergy analysis shows that linguistic and prosodic features do not act independently, such as coordinating conjunction usage, which shows strong synergy with HNR, and the trouble source speaker's self-repetition contributes notably to the presence of maximum intensity. These patterns highlight the nature of the OIR sequence, where how something is said modulates what is being said.

Additionally, our results highlight the crucial role of dialogue micro-context in OIR request detection. Models with integration of both prior and subsequent turns significantly outperform those relying only on the current target segment. However, concatenating too much micro-context could lead to noise and irrelevant information, which affects the model's performance. This result supports the literature in Conversation Analysis that OIR sequences are inherently context-sensitive, and their interpretation often depends on the surrounding interactional structure.

Finally, error analysis revealed that while the text-based model failed with vague references and disfluencies, the audio-based model was prone to misclassifying flat or subtle prosodic cues, which raised the need for a multimodal model. The proposed multimodal model mitigates these weaknesses, but it still struggles with short, minimally marked OIR requests that lack both strong syntactic and prosodic cues.

Building on these insights, future work will explore the integration of visual features to more accurately model the embodied aspects of OIR sequences, as well as the development of multilingual and cross-context corpora to assess the robustness and generalizability of the detection approach.

Limitations

Dataset Limitations and Generalizability. Due to the limited multimodal OIR-labeled corpora, our study utilized the only available multimodal OIR-labeled corpus, which is specific to Dutch language and referential object matching tasks. This specificity could limit the generalizability of our model across different OIR categories, languages, and conversation settings. Future works should test the model on more diverse datasets to validate its robustness and establish broader applicability.

Adaptability in Real-time Processing. Despite the computational efficiency of our approach using handcrafted features compared to Large Language Models, several limitations remain for real-time adaptation. The feature extraction of some linguistic and prosodic features, such as coreference chains, requires additional computation with pre-trained models, potentially introducing latency. Future work should explore real-time feature extraction pipelines and incremental processing architectures, while evaluating potential trade-offs between model complexity and real-time performance to make the system practical for CA systems.

References

Francesca Alloatti, Francesca Grasso, Roger Ferrod, Giovanni Siragusa, Luigi Di Caro, and Federica Cena. 2024. A tag-based methodology for the detection of user repair strategies in task-oriented conversational agents. *Computer Speech & Language*, 86:101603.

Merav Allouch, A. Azaria, and Rina Azoulay-Schwartz. 2021. Conversational agents: Goals, technologies, vision and challenges. *Sensors (Basel, Switzerland)*, 21.

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Vevake Balaraman, Arash Eshghi, Ioannis Konstas, and Ioannis Papaioannou. 2023. No that's not what I meant: Handling third position repair in conversational question answering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 562–571, Prague, Czechia. Association for Computational Linguistics.

Trevor Michael Benjamin. 2013. *Signaling trouble: on the linguistic design of other-initiation of repair in English conversation*. Ph.D. thesis. Relation: http://www.rug.nl/ Rights: University of Groningen.

Paul Boersma. 2000. A system for doing phonetics by computer. 5.

Yuzhao Chen, Wenhua Zhu, Weilun Yu, Hongfei Xue, Hao Fu, Jiali Lin, and Dazhi Jiang. 2024. Prompt learning for multimodal intent recognition with modal alignment perception. *Cogn. Comput.*, 16:3417–3428.

Marcus Colman and Patrick G. T. Healey. 2011. The distribution of repair in dialogue. *Cognitive Science*, 33.

Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My bad! repairing intelligent voice assistant errors improves interaction. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Mark Dingemanse and N. J. Enfield. 2015. Other-initiated repair across languages: Towards a typology of conversational structures.

Mark Dingemanse, Seán G Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S Gisladottir, Kobin H Kendrick, Stephen C Levinson, Elizabeth Manrique, and 1 others. 2015. Universal principles in the repair of communication problems. *PloS one*, 10(9):e0136100.

Lotte Eijk, Marlou Rasenberg, Flavia Arnese, Mark Blokpoel, Mark Dingemanse, Christian F. Doeller, Mirjam Ernestus, Judith Holler, Branka Milivojevic, Asli Özyürek, Wim Pouw, Iris van Rooij, Herbert Schriefers, Ivan Toni, James Trujillo, and Sara Bögels. 2022. The cabb dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, 264.

Raphaela Gehle, Karola Pitsch, and Sebastian Benjamin Wrede. 2014. Signaling trouble in robot-to-group interaction.emerging visitor dynamics with a museum guide robot. *Proceedings of the second international conference on Human-agent interaction*.

Judith Haan, Vincent Van Heuven, Jos Pacilly, and R.L. Bezooijen. 1997. An anatomy of dutch question intonation. *J. Coerts & H. de Hoop (eds.), Linguistics in the Netherlands 1997, 97 - 108 (1997)*, 14.

Elliott Hoey. 2018. How speakers continue with talk after a lapse in conversation. *Research on Language and Social Interaction*, 51.

Sviatlana Höhn. 2017. A data-driven model of explanations for a chatbot that helps to practice conversation in a foreign language. In *Proceedings of the 18th*

Annual SIGdial Meeting on Discourse and Dialogue, pages 395–405, Saarbrücken, Germany. Association for Computational Linguistics.

Xuejian Huang, Tinghuai Ma, Li Jia, Yuanjian Zhang, Huan Rong, and Najla Alnabhan. 2023. An effective multimodal representation and fusion method for multimodal intent recognition. *Neurocomputing*, 548:126373.

Martina Huhtamäki. 2015. The interactional function of prosody in repair initiation: Pitch height and timing of va 'what' in helsinki swedish. *Journal of Pragmatics*, 90:48–66.

Jan Ittner, Lukasz Bolikowski, Konstantin Hemker, and Ricardo Kennedy. 2021. Feature synergy, redundancy, and independence in global model explanations using shap vector decomposition. *ArXiv*, abs/2107.12436.

D. Jain, Anil Rahate, Gargi Joshi, Rahee Walambe, and K. Kotecha. 2024. Employing co-learning to evaluate the explainability of multimodal sentiment analysis. *IEEE Transactions on Computational Social Systems*, 11:4673–4680.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *Preprint*, arXiv:2108.12009.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. pages 3081–3084.

Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. 2020. Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 1094–1107, New York, NY, USA. Association for Computing Machinery.

Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. 2023. Hierarchical dialogue understanding with special tokens and turn-level attention. *ArXiv*, abs/2305.00262.

Md Messal Monem Miah, Ulie Schnaithmann, Arushi Raghuvanshi, and Youngseo Son. 2024. Multimodal contextual dialogue breakdown detection for conversational ai models. *ArXiv*, abs/2404.08156.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA. Association for Computational Linguistics.

Robert J. Moore, Sungeun An, and Olivia H. Marrese. 2024. Understanding is a two-way street: User-initiated repair on agent responses and hearing in conversational interfaces. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Anh Ngo, Dirk Heylen, Nicolas Rollet, Catherine Pelachaud, and Chloé Clavel. 2024. Exploration of human repair initiation in task-oriented dialogue: A linguistic feature-based approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 603–609, Kyoto, Japan. Association for Computational Linguistics.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in Cognitive Science*, 10(2):425–451.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Marlou Rasenberg, Wim Pouw, Asli Özyürek, and Mark Dingemanse. 2022. The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, 12.

Tulika Saha, Aditya Patra, S. Saha, and P. Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. pages 4361–4372.

Emanuel A. Schegloff. 1987. Between micro and macro: contexts and other connections. In Richard Munch Jeffrey C. Alexander, Bernhard Giesen and Neil J. Smelser, editors, *The Micro-Macro Link*, page 207–234. University of California Press, Berkeley.

Emanuel A. Schegloff. 2000. When 'others' initiate repair. *Applied Linguistics*, 21:205–243.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53:361.

Margret Selting. 1996. *Prosody as an activity-type distinctive cue in conversation: the case of so-called 'astonished' questions in repair initiation*, page 231–270. Studies in Interactional Sociolinguistics. Cambridge University Press.

Mathilde Theelen. 2017. Fundamental frequency differences including language effects. *Junctions: Graduate Journal of the Humanities*, 2:9.

Jacqueline van Arkel, Marieke Woensdregt, Mark Dingemanse, and Mark Blokpoel. 2020. A simple repair mechanism can alleviate computational demands of pragmatic reasoning: simulations and complexity analysis. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 177–194, Online. Association for Computational Linguistics.

Reneé van Bezooijen. 1995. Sociocultural aspects of pitch differences between japanese and dutch women. *Language and Speech*, 38(3):253–265. PMID: 8816084.

Monique van Donzel and Florien Beinum. 1996. Pausing strategies in discourse in dutch. pages 1029 – 1032 vol.2.

Jo Verhoeven and Bruce Connell. 2024. Intrinsic vowel pitch in hamont dutch: Evidence for if0 reduction in the lower pitch range. *Journal of the International Phonetic Association*, 54(1):108–125.

Traci Walker and Trevor Benjamin. 2017. Phonetic and sequential differences of other-repetitions in repair initiation. *Research on Language and Social Interaction*, 50(4):330–347.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Dialogue Micro Context

B Detailed Linguistic Features

C Detailed Prosodic Features

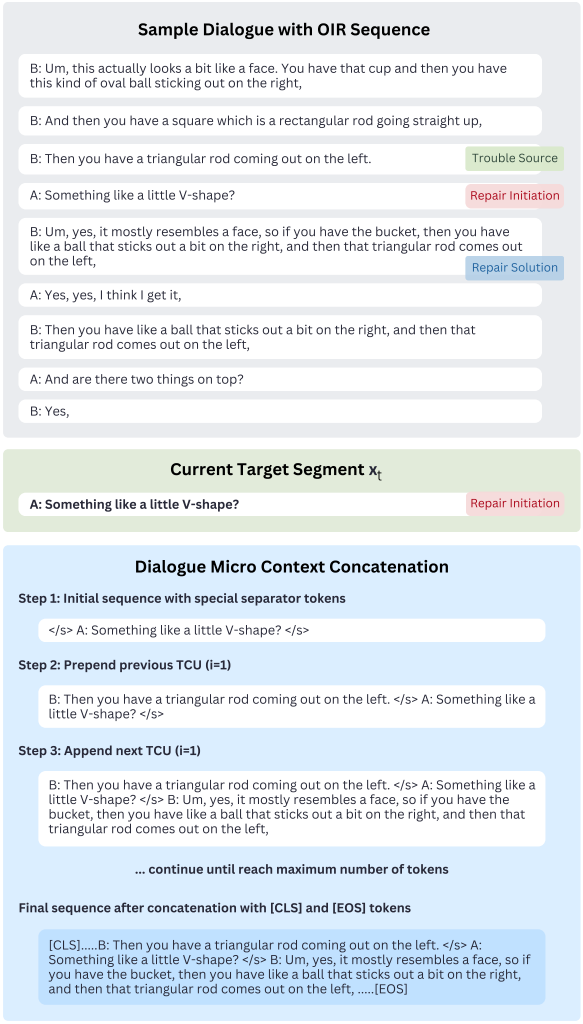

Figure 8: Dialogue micro context concatenation approach. *Micro context* refers to the immediate conversational environment, including the prior turns and the subsequent turns of the current target turn in dialogue (Schegloff, 1987).

Level	Feature Group	Feature Type(s)	Description
Segment-level	POS tags sequence	POS tag bigrams, POS tag ratios	Binary features for frequent POS tag bigrams (e.g., PRON_Prs→VERB, VERB→COREF); POS tags frequency ratios computed per utterance.
	Lemma	contains_lemma (e.g., nog, kunnen)	Binary indicators for presence of high-frequency lemmas relevant to different type of OIRs.
	Question form	ends_with_question_mark	Binary feature indicating whether the utterance ends with a question mark.
	Non-verbal action	contains_laugh, contains_sigh, etc.	Binary features for transcribed non-verbal actions like #laugh#, #sigh#, etc.
Cross-segment level (prior turns related)	Repetition from previous turn	other_repetition_ratio	Ratio of tokens in the current utterance that are repeated from the other speaker's previous turn relative to total utterance length.
	Coreference from previous turn	coref_used_ratio	Ratio of coreference phrases (e.g., pronouns or noun phrases referring to previous turn) relative to total utterance length.
Cross-segment level (subsequent turns related)	Repair solution TSS self-repetition	other_speaker_self_rep_ratio	Ratio of self-repetition in the turn following the OIR.
	Repair solution TSS other-repetition	other_speaker_other_rep_ratio	Ratio of other-repetition in the turn following the OIR

Table 4: Summary of linguistic feature set used for modeling OIR request.

Level	Feature Group	Feature Type	Description
Segment-level	Pitch features	pitch_min, pitch_max, mean, std, range, num_peaks	Extracted from voiced frames; outliers removed; peaks from smoothed contour (Savitzky-Golay).
	Pitch dynamics	slope, acceleration, inflection_rate	Captures pitch variation and shape within utterance.
	Intensity features	min, max, mean, std, range	Computed from nonzero intensity frames; reflects loudness.
	Voice quality	jitter, shimmer, hnr	Reflects vocal fold irregularity and breathiness.
	Pause features	num, durations, short/med/long, positional counts, rel_longest	Pause detection using adaptive thresholds; categorized by duration and position.
	Speech timing	rate, duration	Utterance length and estimated speech rate (e.g., syllables/sec).
Cross-segment	Transition features	end_slope, start_slope, transition	Pitch slope difference across turn boundaries (prev→cur, cur→next); in semitones/sec.
	Baseline comparison	z_score, rel_change, range_pos	Comparison to speaker's pitch/intensity baseline for markedness detection.
	Latency	TS→OIR, OIR→RS	Silence duration between trouble source and repair turns.

Table 5: Summary of prosodic feature set used for modeling OIR request.