ParaCalib: Semantically Calibrated Paralinguistic Modeling for Depression Detection

Anonymous ACL submission

Abstract

Paralinguistic patterns are inherently context-sensitive, with the same acoustic pattern carrying different implications across utterance content, speakers, and conversational settings. However, most existing speech-based depression detection methods model acoustic patterns as context-independent markers. To address this limitation, we introduce ParaCalib, which uses an ALM captioner to generate contextualized vocal descriptions and an LLM-based Paralinguistic State Extractor (PSE) to map them into semantically calibrated paralinguistic states. Here, semantic calibration refers to structuring context-aware vocal descriptions into comparable paralinguistic states. Experiments show that ParaCalib achieves the best Macro-F1 among evaluated methods, with scores of 71.9% on DAIC-WOZ and 90.5% on MODMA. Further analyses reveal non-monotonic and compositional depression-related speech patterns: depression-related evidence emerges from state combinations rather than isolated states, and high-attention SC-Para evidence is more internally consistent in depressed sessions. These findings support semantically calibrated rather than context-independent interpretation of vocal behavior.

1 Introduction

Depression is a prevalent mental disorder and a major contributor to disability and impaired daily functioning (World Health Organization, 2025). Early and scalable screening is therefore central to computational mental health. Speech is a promising signal because it is non-invasive, easy to collect in clinical interviews or remote interactions, and psychologically informative, reflecting both what a person says and how they say it (Cummins et al., 2015; Li et al., 2025b).

Existing speech-based depression detection methods largely treat speech as a source of acoustic markers, with the role of semantic context left implicit. Early approaches rely on low-level acoustic descriptors (LLDs), including prosodic, spectral, voice-quality, and temporal features, motivated by clinical links between depression and psychomotor slowing, longer pauses, reduced vocal energy, monotonic prosody, and altered voice quality (Ellgring and Scherer, 1996; Cummins et al., 2015; Yamamoto et al., 2020). End-to-end and self-supervised (SSL) models learn patterns from raw waveforms, time-frequency representations, or large unlabeled speech corpora (Ma et al., 2016; Chen et al., 2022b; Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022a; Wu et al., 2023; Li et al., 2025a). However, these representations are typically used as generic speech features rather than as explicit models of how semantic context changes the interpretation of paralinguistic cues.

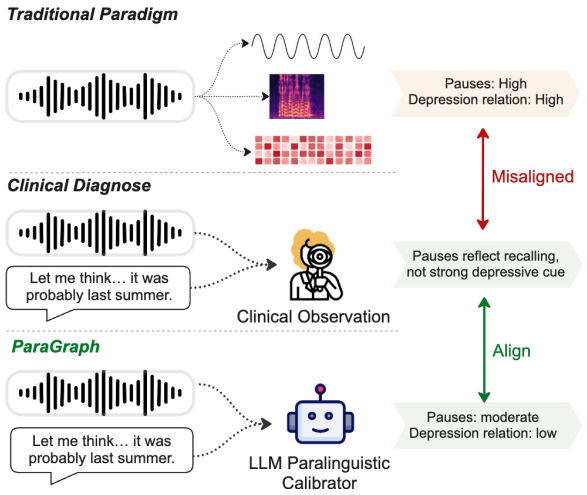

Figure 1: Illustration of semantic calibration for paralinguistic cues.

We analyze this limitation from two complementary perspectives. First, depression-related behavior is clinically heterogeneous, as diagnostic descriptions of major depressive episodes include both psychomotor agitation and retardation, and prior acoustic studies have examined differ-

ences between these depressive presentations (Association, 2022; Alpert et al., 2001). Therefore, depression-related speech should not be expected to follow a single monotonic acoustic profile, such as uniformly slower speech or uniformly more frequent pausing. Second, the clinical meaning of paralinguistic expression is not determined by acoustics alone. A long pause, flattened pitch contour, reduced intensity, or slowed speaking rate may reflect depressive psychomotor retardation in one utterance, but may also arise from hesitation, emphasis, topic shifts, syntactic planning, turn-taking, or the semantic content being expressed (Cole, 2015; Wagner and Watson, 2010). Thus, such vocal patterns should be interpreted relative to utterance-level semantic and pragmatic context rather than as absolute acoustic deviations.

Motivated by this perspective, we propose ParaCalib, a framework that represents utterance-level vocal behavior as semantically calibrated paralinguistic states. We define semantic calibration as interpreting vocal behavior relative to utterance-level semantic and pragmatic context and expressing it in a structured paralinguistic state space. Figure 1 illustrates this process. ParaCalib operationalizes this process by combining ALM-generated contextualized vocal descriptions with PSE-based mapping into predefined SC-Para dimensions. For each subject, ParaCalib treats the valid utterance-level SC-Para vectors as an instance bag and aggregates them with an attention-based MIL classifier for subject-level prediction.

Experiments on DAIC-WOZ and MODMA show that ParaCalib improves Macro-F1 over conventional acoustic features, SSL speech representations, and alternative subject-level classifiers, while providing inspectable intermediate SC-Para states. Further analyses show that the captioner, PSE structuring, and MIL aggregation each contribute to the final performance.

2 Related Work

2.1 Depression Detection from Speech

Progress in speech-based depression detection has been shaped by the AVEC challenge series (Valstar et al., 2013; Ringeval et al., 2019), which introduced standard benchmarks for multimodal depression assessment. Early systems relied on handcrafted acoustic features, including prosodic, spectral, temporal, and voice-quality descriptors (Li et al., 2025b), but such features required manual design and showed limited robustness. Beyond purely acoustic modeling, Morales et al. (Morales et al., 2018) proposed a linguistically informed fusion framework on DAIC-WOZ, where syntactic information guided acoustic feature extraction and fusion. Their results suggest that linguistic context can help interpret vocal evidence, but the approach remains a feature-level fusion framework based on handcrafted acoustic, syntactic, and visual representations, rather than an explicit model of semantically calibrated paralinguistic states. End-to-end CNN, RNN, and Transformer models reduce this dependence by learning from raw audio or mel-spectrograms (Chlasta et al., 2019; Dubagunta et al., 2019; Saidi et al., 2020; Othmani et al., 2021; Vázquez-Romero and Gallardo-Antolín, 2020; Wang et al., 2022; Salekin et al., 2018; Zhao et al., 2020; Muzammel et al., 2020). More recently, pretrained speech models such as Wav2Vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022a), and Whisper (Radford et al., 2023) have improved transferability in low-resource depression detection (Zhang et al., 2021; Toto et al., 2021; Chen et al., 2022b; Ravi et al., 2022; Wu et al., 2023; Li et al., 2025a; Zhang et al., 2025b). These methods improve acoustic representation learning, but they generally leave the contextual interpretation of paralinguistic cues implicit.

2.2 Audio-Language Models for Affective Speech Understanding

Recent Audio-Language Models (ALMs), such as SALMONN (Tang et al., 2024), Qwen-Audio (Chu et al., 2023), Qwen2-Audio (Chu et al., 2024), and GAMA (Ghosh et al., 2024), align audio encoders with pretrained LLMs and have shown increasing potential for affective and mental-state speech analysis. Many prior studies have used ALMs or LLMs for closed-set speech emotion recognition, free-form speech emotion captioning, or explanation-style prediction refinement grounded in transcripts and acoustic cues (Zhang et al., 2025a; Xu et al., 2024b). Related ordinal alignment work introduces contrastive ranking objectives, but defines ordinality over valence-arousal spaces rather than depression-related paralinguistic characteristics (Chandra et al., 2025).

Despite these advances, a gap remains between semantically rich vocal descriptions and structured representations for downstream depression detection. ALM-based audio captioners can de-

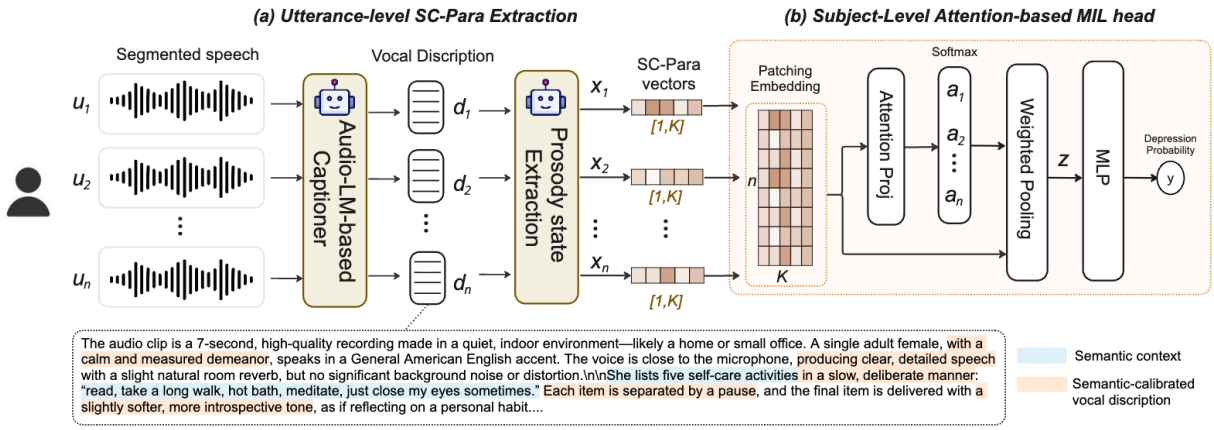

Figure 2: Overview of ParaCalib. A frozen ALM captioner converts utterance-level speech into contextualized vocal descriptions, and PSE structures those descriptions into utterance-level SC-Para vectors. The attention-based MIL classifier uses an attention projection only to estimate utterance-level instance weights, forms the subject-level representation as a weighted sum of the original seven-dimensional SC-Para vectors, and feeds this pooled representation to an MLP for depression detection.

scribe vocal behavior in natural language, but their free-form outputs are difficult to compare across utterances and subjects. Classifier-based methods often collapse affective information into categorical labels, while depression-specific speech LLMs (Zhang et al., 2024; Zhao et al., 2025) leave intermediate paralinguistic representations largely opaque. These limitations motivate compact, comparable SC-Para states that preserve the contextual sensitivity of ALM-generated descriptions while making caption-derived paralinguistic evidence more structured and inspectable for subject-level modeling.

3 Methodology

3.1 Overview

As shown in Figure 2, ParaCalib consists of three stages: ALM-based vocal captioning, PSE-based SC-Para extraction, and attention-based MIL aggregation.

3.2 Utterance-level ALM Captioning

For each utterance-level audio clip u_i, we use a frozen Qwen3-Omni-30B-A3B ALM captioner[1] to generate a contextualized vocal description d_i. The ALM captioning inference settings are provided in Appendix D. As shown in Appendix I.5 Table 13, the caption provides a context-aware natural-language description of vocal behavior for subsequent SC-Para structuring.

[1] https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Captioner

3.3 SC-Para Extraction

While ALM-generated captions capture rich vocal behaviors, their free-form descriptions are not directly comparable across utterances or subjects. PSE therefore converts each vocal description d_i into a fixed, clinically motivated SC-Para schema, following recent work on LLM-based structured information extraction and annotation (Xu et al., 2024a; Dagdelen et al., 2024; Ziems et al., 2024). This schema abstracts depression-relevant paralinguistic evidence into a compact set of interpretable states, rather than preserving all acoustic variation in the signal. Because PSE operates on contextualized captions rather than raw audio, the extracted states instantiate the calibrated state representation defined above. The PSE inference settings are detailed in Appendix E. To assess the reliability of these intermediate representations, we conduct both a repeated-decoding stability check for PSE extraction in Appendices G and a small-scale human consistency check for the generated captions and the five constituent SC-Para states in Appendix H.

The SC-Para schema as shown in Appendix I.4 Table 11 consists of seven dimensions, including five constituent paralinguistic states and two holistic scores. The five constituent states characterize pitch variability, speaking rate, rhythm regularity, pause pattern, and vocal energy. These dimensions are selected to reflect paralinguistic phenomena commonly discussed in clinical and affective speech studies (Cummins et al., 2015; Koops et al., 2023). They are encoded as discrete paralinguistic

states and then mapped to numerical values. The two holistic scores are PSE-produced utterance-level intermediate variables, not gold utterance-level depression labels. The evidence score is produced by PSE on a 0–100 scale and normalized to $[0, 1]$ before classification, while the confidence score is represented on $[0, 1]$.

Let $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(K)}\}$ denote the predefined SC-Para dimensions, where $K = 7$. For each utterance u_i, PSE maps d_i into an utterance-level SC-Para vector:

$$x_i = \text{PSE}(d_i) = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(K)}]. \quad (1)$$

During preprocessing, we discard utterances if any categorical paralinguistic state is `unknown`. To assess the impact of this filtering step, we report utterance-level and per-subject discard rates by split and subject-level label group in Appendix F. The dimension definitions, discrete state space, numerical score definitions, PSE prompt, and inference settings are provided in Appendix I.

3.4 Attention-based MIL Classification

For a subject b, let $S_b = \{x_{b,1}, x_{b,2}, \ldots, x_{b,n_b}\}$ denote the bag of valid utterance-level SC-Para vectors after PSE structuring, where n_b may vary across subjects. The attention-based MIL classifier assigns an importance score to each instance using an additive attention projection with a learnable query vector:

$$e_{b,i} = v^\top \tanh(W_{\text{att}} x_{b,i} + b_{\text{att}}), \quad (2)$$

$$\alpha_{b,i} = \frac{\exp(e_{b,i})}{\sum_{j=1}^{n_b} \exp(e_{b,j})}. \quad (3)$$

The utterance-level SC-Para vectors are then pooled into a subject-level representation:

$$z_b = \sum_{i=1}^{n_b} \alpha_{b,i} x_{b,i}. \quad (4)$$

The subject-level representation z_b is passed to a two-hidden-layer MLP, which produces the scalar logit s_b for binary depression classification. In this design, the attention projection estimates utterance-level instance weights, while the classifier operates on the pooled seven-dimensional SC-Para representation.

3.5 Optimization Objective

We optimize a binary depression classification objective, converting the subject-level logit s_b to $\hat{y}_b = \sigma(s_b)$ and using weighted binary cross-entropy to address class imbalance:

$$\mathcal{L} = -\frac{1}{B} \sum_{b=1}^{B} \left[w y_b \log \hat{y}_b + (1 - y_b) \log(1 - \hat{y}_b) \right], \quad (5)$$

where B is the batch size, y_b is the binary subject-level label, and $w = N_{\text{neg}}/N_{\text{pos}}$ weights the positive class.

4 Data and Evaluation

We evaluate ParaCalib on DAIC-WOZ (Gratch et al., 2014) and MODMA (Cai et al., 2022). DAIC-WOZ contains 189 English-speaking participants, and we follow the official train/dev/test split. MODMA includes 52 Mandarin-speaking participants. We use only the interview task, which contains 18 utterances per subject. Since MODMA does not provide an official split, we use subject-level 5-fold cross-validation. We report subject-level Macro-F1 as the primary metric, with sensitivity, specificity, and ROC-AUC as supplementary metrics, averaged over five random seeds. Further dataset, preprocessing, label correction and implementation details are provided in Appendices A and B.

5 Results and Analysis

5.1 Depression Detection Performance

Table 1 summarizes the subject-level results on DAIC-WOZ and MODMA. ParaCalib achieves the best Macro-F1, sensitivity, and AUC on both datasets. On DAIC-WOZ, it improves Macro-F1 by 11.4 pp over the strongest baseline, HAREN-CTC. Although HAREN-CTC obtains higher specificity, ParaCalib achieves substantially higher Macro-F1 and a better overall balance. On MODMA, ParaCalib improves Macro-F1 by 26.1 pp and sensitivity by 24.5 pp over the strongest baselines. It also improves AUC by 22.4 pp, while matching the best baseline specificity of 90.0%. These results suggest that SC-Para provides effective and generalizable paralinguistic representations for subject-level depression detection.

5.2 Representation Analysis

To examine the contribution of semantic calibration, we compare SC-Para with LLD-based, SSL-based, transcript-based, and caption-based representations in Table 2.

Dataset	Methods	Input	Macro-F1 (↑)	Sensitivity (↑)	Specificity (↑)	AUC (↑)
DAIC-WOZ	DepAudioNet (Ma et al., 2016)*	MFbanks	48.2 (.04)	34.6 (.25)	66.7 (.19)	51.0 (.05)
	NUSD (Wang et al., 2023)*	Raw speech	51.4 (.05)	30.0 (.14)	74.5 (.10)	57.5 (.03)
	Wu et al., 2023 (Wu et al., 2023)*	WavLM	53.5 (.05)	32.8 (.24)	78.2 (.18)	54.7 (.08)
	SpeechFormer (Chen et al., 2022b)*	Wav2Vec 2.0	59.8 (.02)	42.9 (.12)	77.3 (.12)	54.0 (.04)
	HAREN-CTC (Li et al., 2025a)*	WavLM	60.5 (.02)	37.1 (.12)	**83.6** (.07)	55.6 (.01)
	ParaCalib	SC-Para	**71.9** (.05)	**77.1** (.15)	73.3 (.07)	**83.3** (.03)
MODMA	DepAudioNet (Ma et al., 2016)*	MFbanks	55.7 (.03)	56.2 (.03)	56.0 (.04)	58.6 (.03)
	Wu et al., 2023 (Wu et al., 2023)*	WavLM	63.5 (.10)	40.0 (.14)	**90.0** (.09)	74.0 (.15)
	HAREN-CTC (Li et al., 2025a)*	WavLM	64.4 (.03)	66.5 (.02)	63.5 (.04)	68.0 (.03)
	ParaCalib	SC-Para	**90.5** (.11)	**91.0** (.12)	**90.0** (.15)	**96.4** (.07)

Table 1: Main subject-level depression detection results on DAIC-WOZ and MODMA. * indicates reproduced baselines. For open-source baselines, we use the released code with the reported best configurations. For Wu et al. (Wu et al., 2023), whose implementation is not publicly available, we reproduce the pretrained WavLM setting without data augmentation following the paper. Mean values are reported in %. SD is reported on the 0–1 scale.

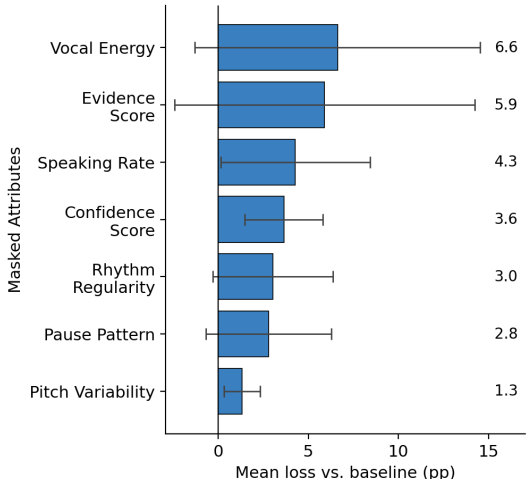

Figure 3: Mean Macro-F1 loss after masking each attribute, relative to the full model. Error bars indicate SD across runs.

The LLD-based results show that raw acoustic descriptors have limited discriminative power for subject-level depression detection, likely reflecting their sensitivity to speaker, channel, and phonetic variability. Discretizing LLDs improves Macro-F1 from 43.9% to 49.0% on DAIC-WOZ and from 53.4% to 67.8% on MODMA, suggesting that state abstraction can suppress local numerical variability and capture more stable vocal patterns. However, these states are still derived from uncalibrated acoustic measurements and remain substantially below the strongest caption-based representations.

The transcript-based baselines examine whether lexical information, or using transcripts to guide acoustic feature extraction, can explain the gains of semantic calibration. On DAIC-WOZ, the transcript-only baseline achieves a strong Macro-F1 of 63.1%, indicating that lexical content is informative. However, the model combining text features derived from transcripts with WavLM representations performs substantially worse than the transcript-only baseline, and LLD-PSE with transcripts achieves high specificity but low sensitivity. These results suggest that lexical context is useful but insufficient, and that ParaCalib's gains do not simply come from adding transcripts to acoustic features.

The SSL-based results show that higher-capacity acoustic representations do not necessarily yield clinically aligned paralinguistic features. Although Wav2Vec 2.0, HuBERT, and WavLM encode rich speech information, their performance remains below the strongest caption-based methods on both datasets. This suggests that depression-relevant cues may remain entangled with speaker and channel factors in small clinical datasets.

Within the caption-based group, Transcript-masked Caption-PSE masks explicit transcript-derived lexical content while retaining ALM-generated descriptions of vocal behavior. Its performance indicates that, even after explicit transcript content is suppressed, the retained vocal-behavior descriptions still provide useful paralinguistic evidence for subject-level depression classification. The remaining gap to SC-Para suggests that the full calibrated state representation captures additional context-dependent vocal evidence.

Overall, SC-Para achieves the best Macro-F1, sensitivity, and AUC on both datasets, with particularly large gains over LLD-based, SSL-based, and transcript-based alternatives. Compared with direct Caption Embeddings, SC-Para further improves performance by mapping free-form ALM

Dataset	Representation type	Representation	Macro-F1 (↑)	Sensitivity (↑)	Specificity (↑)	AUC (↑)
DAIC-WOZ	LLD-based	LLDs	43.9 (.09)	35.7 (.29)	62.4 (.35)	48.1 (.04)
		Discretized LLDs	49.0 (.06)	40.0 (.21)	61.2 (.14)	50.4 (.07)
	SSL-based	Wav2Vec 2.0	41.8 (.10)	40.0 (.26)	53.3 (.33)	42.3 (.05)
		HuBERT	46.3 (.06)	37.1 (.13)	57.0 (.07)	40.1 (.03)
		WavLM	41.1 (.07)	28.6 (.35)	66.7 (.37)	46.6 (.06)
	Transcript-based	Transcript	63.1 (.13)	52.9 (.30)	79.4 (.13)	73.5 (.03)
		LLD-PSE w/ transcript	54.5 (.08)	28.6 (.19)	**84.2** (.15)	64.2 (.07)
		Transcript + WavLM	46.1 (.11)	20.0 (.22)	81.2 (.29)	49.7 (.12)
	Caption-based	Transcript-masked Caption-PSE	64.6 (.04)	51.4 (.09)	78.2 (.08)	77.9 (.02)
		Caption Embedding	66.6 (.14)	68.6 (.39)	75.8 (.14)	77.1 (.12)
		SC-Para	**71.9** (.05)	**77.1** (.15)	73.3 (.07)	**83.3** (.03)
MODMA	LLD-based	LLDs	53.4 (.12)	35.4 (.14)	79.1 (.12)	57.8 (.13)
		Discretized LLDs	67.8 (.19)	56.4 (.26)	82.0 (.12)	78.7 (.19)
	SSL-based	Wav2Vec 2.0	55.6 (.08)	62.6 (.24)	59.2 (.17)	63.7 (.07)
		HuBERT	61.6 (.25)	62.8 (.19)	75.1 (.18)	71.8 (.22)
		WavLM	68.8 (.20)	63.2 (.26)	79.1 (.16)	76.5 (.20)
	Caption-based	Transcript-masked Caption-PSE	60.6 (.20)	58.2 (.15)	71.9 (.17)	69.2 (.13)
		Caption Embedding	85.7 (.10)	86.4 (.09)	85.3 (.11)	94.3 (.07)
		SC-Para	**90.5** (.11)	**91.0** (.12)	**90.0** (.15)	**96.4** (.07)

Table 2: Representation analysis of ParaCalib across various representation types. Results are reported as mean% (SD). SD is reported on the 0–1 scale. Shaded cells indicate the best performance within each representation type, while bold values denote the overall best result per dataset and metric. Transcript-based baselines are omitted for MODMA due to the lack of available transcripts. Full extraction configurations and dimension specifications for all evaluated representations are detailed in Appendix J.

descriptions into a structured paralinguistic state space. This suggests that ParaCalib's key advantage lies not only in using natural-language descriptions, but also in converting them into comparable SC-Para states across utterances and subjects.

5.3 Subject-Level Aggregation

Table 3 shows that SC-Para remains competitive across classifiers, including MLP (Rumelhart et al., 1986), LSTM (Graves, 2012), Bi-LSTM (Schuster and Paliwal, 1997), and Transformer models with and without positional encoders (Vaswani et al., 2017), as detailed in Appendix B. This indicates that the representation itself carries discriminative paralinguistic evidence. Attentive MIL performs best on both datasets, suggesting that depression-related cues are unevenly distributed across utterances and can benefit from instance-level weighting.

5.4 Interpreting SC-Para Evidence

Attribute Contributions To assess the contribution of each SC-Para dimension to subject-level prediction, we mask each dimension individually and measure the resulting Macro-F1 loss relative to the full model. As shown in Figure 3, the largest average losses occur for vocal energy, the PSE-derived paralinguistic evidence score, and speaking rate. Figure 5 further shows that these dimensions contribute through different metric profiles. Vocal energy and the evidence score mainly affect sensitivity, suggesting that they help identify depressed-labeled subjects, whereas speaking rate contributes more to specificity, indicating its role in identifying non-depressed subjects. The extraction confidence score shows a more balanced effect across metrics, suggesting that extraction reliability provides a stable complement to the constituent paralinguistic states.

The remaining constituent states provide more specialized evidence. Rhythm regularity mainly contributes to specificity, pause pattern contributes more to sensitivity, and pitch variability has the smallest overall effect. These patterns suggest that SC-Para dimensions provide complementary information, with holistic evidence, vocal energy, and pause patterns supporting sensitivity, while speaking rate and rhythm regularity support specificity.

Using only the PSE-derived paralinguistic evidence score yields a Macro-F1 of 63.2%, substantially lower than the 71.9% obtained by the full SC-Para vector. Conversely, removing this score still yields 67.0%, indicating that ParaCalib does not rely solely on this holistic score. Instead, the con-

Dataset	Classifier	Macro-F1 (↑)	Sensitivity (↑)	Specificity (↑)	AUC (↑)
DAIC-WOZ	Bi-LSTM	49.6 (.03)	40.0 (.13)	61.2 (.10)	56.8 (.04)
	LSTM	42.2 (.05)	18.6 (.15)	70.9 (.25)	45.7 (.04)
	Transformer w/ Pos. Enc.	40.2 (.10)	22.9 (.43)	**79.4** (.44)	55.8 (.04)
	Transformer w/o Pos. Enc.	60.6 (.06)	**85.7** (.09)	50.9 (.11)	73.2 (.04)
	MLP	65.5 (.05)	58.6 (.09)	74.5 (.03)	73.7 (.04)
	Attentive MIL	**71.9** (.05)	77.1 (.15)	73.3 (.07)	**83.3** (.03)
MODMA	LSTM	79.3 (.08)	74.0 (.09)	85.3 (.14)	94.2 (.04)
	Transformer w/ Pos. Enc.	77.6 (.09)	71.4 (.09)	86.4 (.17)	94.4 (.07)
	Transformer w/o Pos. Enc.	81.3 (.08)	77.0 (.07)	86.0 (.15)	95.7 (.06)
	Bi-LSTM	82.8 (.08)	80.6 (.07)	85.5 (.12)	94.2 (.08)
	MLP	84.7 (.13)	90.0 (.10)	81.9 (.19)	90.5 (.11)
	Attentive MIL	**90.5** (.11)	**91.0** (.12)	**90.0** (.15)	**96.4** (.07)

Table 3: Classifier architecture comparison using SC-Para representations. All classifiers use the same utterance-level SC-Para vectors as input, with classifier-specific subject-level aggregation strategies detailed in Appendix B. Results are reported as mean% (SD). SD is reported on the 0–1 scale. Bold indicates the best result within each dataset and metric.

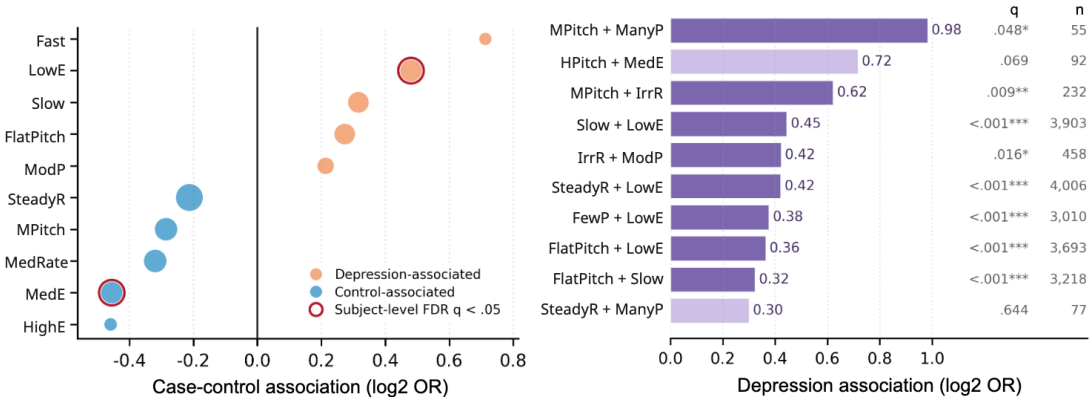

Figure 4: Dataset-level associations between SC-Para states and subject-level depression labels. Left: marginal utterance-level $\log_2$ odds ratios, where each utterance inherits its subject-level label and bubble size indicates utterance count. Right: top state pairs associated with depressed-labeled subjects, ranked by $\log_2$ odds ratio, with q and n denoting FDR-corrected significance and pair count. Abbreviations: H/M/FlatPitch = high/moderate/flat pitch variability; MedRate = medium speaking rate; Low/Med/HighE = low/medium/high vocal energy; ModP = moderate pauses; Steady/IrrR = steady/irregular rhythm.

stituent paralinguistic states, extraction confidence, and subject-level aggregation provide complementary information for depression detection.

State-Level and Pairwise Associations To complement the seven-dimensional masking analysis above, we further examine dataset-level associations between subject-level depression labels and the five categorical constituent SC-Para states. We exclude the two continuous holistic scores from this analysis to avoid treating summary scores as discrete vocal states. The marginal analysis shows that these associations are selective rather than monotonic. Low energy is the clearest individual state associated with depressed-labeled subjects, while fast and slow speaking-rate states, flat pitch, and moderate pauses also show positive associations. At the same time, several states, including medium energy, medium speaking rate, and moderate pitch, are associated with non-depressed subjects. This pattern cautions against interpreting isolated SC-Para states, such as slow speech or frequent pausing, as standalone depression markers.

The pairwise analysis further supports the compositional nature of paralinguistic evidence. Several well-supported state pairs associated with depressed-labeled subjects involve low energy combined with slow speech, steady rhythm, few pauses, or flat pitch. Other high-effect pairs, such as moderate pitch with many pauses, have smaller support and should be interpreted together with their corrected significance and counts. This analysis provides post-hoc dataset-level evidence of group differences in extracted paralinguistic states and should not be interpreted as an utterance-level

clinical label or diagnostic rule.

6 Discussion

The results suggest that paralinguistic evidence for depression detection is better treated as context-dependent and compositional, rather than as a set of fixed acoustic markers. This finding is consistent with the heterogeneous nature of depression-related speech behavior, since similar acoustic patterns may carry different implications depending on utterance content, interaction context, and co-occurring vocal characteristics. By converting ALM-generated vocal descriptions into SC-Para states, ParaCalib provides a compact representation that improves over conventional low-level descriptors and SSL embeddings.

This formulation shifts interpretability from post-hoc explanations of opaque acoustic models to inspectable intermediate states. SC-Para explicitly represents pitch variability, speaking rate, rhythm, pauses, vocal energy, depression-related paralinguistic evidence, and extraction confidence before subject-level aggregation, making it possible to inspect the utterance-level evidence behind a prediction. The pairwise association analysis and the attention-weighted consistency analysis in Appendix K further support this interpretation by showing that depression-related vocal evidence is better captured through structured state combinations than isolated acoustic indicators.

However, SC-Para should be understood as computational evidence rather than a set of clinical symptoms or diagnostic rules. Individual characteristics, such as low energy, slowed speech, or flattened pitch, should not be interpreted in isolation. Instead, ParaCalib aggregates multiple paralinguistic signals across a dialogue, providing structured evidence for screening-oriented analysis without directly implying diagnosis. This distinction is especially important in clinical settings, where model outputs should be interpreted as decision-support signals rather than diagnostic conclusions.

Practically, ParaCalib decouples representation extraction from subject-level classification. Once SC-Para states are extracted, prediction requires only seven-dimensional inputs and a lightweight attentive MIL classifier, yielding a modular and auditable pipeline. Deployment, however, still requires addressing the cost, latency, privacy, and possible distillation or caching of the ALM/LLM components.

More broadly, the same formulation may be useful beyond depression detection, especially for tasks where vocal behavior must be interpreted relative to content and interaction context, such as emotion recognition, speaker-state modeling, and clinical speech assessment (Schuller and Batliner, 2013; Cummins et al., 2015; König et al., 2015; Braga et al., 2019).

7 Limitations

First, the seven-dimensional schema is intentionally compact. This improves comparability and interpretability, but may not cover all paralinguistic phenomena relevant to depression or other tasks. Future work could extend the schema with task-specific dimensions while preserving the same semantic calibration framework.

Second, our experiments do not provide a fully controlled counterfactual test of semantic calibration, where identical acoustic realizations are paired with different semantic contexts. Controlled synthetic data or data augmentation could help test how semantic context affects SC-Para assignment under fixed or systematically varied acoustic conditions.

Finally, demographic factors should be examined more explicitly in future work (Yang et al., 2024). Some work has specifically analyzed gender bias on DAIC-WOZ (Bailey and Plumbley, 2021; Galanina et al., 2025; Kim et al., 2025), showing that dataset imbalance can affect depression-detection performance estimates. Future studies could therefore evaluate ParaCalib across demographic subgroups and examine whether SC-Para distributions or prediction errors vary with gender, age, or other population factors.

8 Conclusion

This paper introduced ParaCalib, a framework that represents dialogue-level speech through semantically calibrated paralinguistic states for depression detection. Across DAIC-WOZ and MODMA, SC-Para consistently improves subject-level depression screening while preserving inspectable utterance-level paralinguistic states. These results support treating depression-related vocal behavior as contextual and compositional evidence rather than as fixed acoustic markers.

References

Murray Alpert, Enrique R Pouget, and Raul R Silva. 2001. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1):59–69.

American Psychiatric Association. 2022. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association Publishing.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Andrew Bailey and Mark D Plumbley. 2021. Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 596–600. IEEE.

Diogo Braga, Ana M Madureira, Luis Coelho, and Reuel Ajith. 2019. Automatic detection of parkinson's disease based on acoustic analysis of speech. *Engineering Applications of Artificial Intelligence*, 77:148–158.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, and 1 others. 2022. A multimodal open dataset for mental-disorder analysis. *Scientific data*, 9(1):178.

Shreeram Suresh Chandra, Lucas Goncalves, Junchen Lu, Carlos Busso, and Berrak Sisman. 2025. Emotionrankclap: Bridging natural language speaking styles and ordinal speech emotion via rank-n-contrast. *arXiv preprint arXiv:2505.23732*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022a. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2022b. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. *arXiv preprint arXiv:2203.03812*.

Karol Chlasta, Krzysztof Wołk, and Izabela Krejtz. 2019. Automated speech-based screening of depression using deep convolutional neural networks. *Procedia Computer Science*, 164:618–628.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Jennifer Cole. 2015. Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1-2):1–31.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418.

S Pavankumar Dubagunta, Bogdan Vlasenko, and Mathew Magimai Doss. 2019. Learning voice source related information for depression detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529. IEEE.

Heiner Ellgring and Klaus R. Scherer. 1996. Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20(2):83–110.

Marina Galanina, Anna Rekiel, Anna Bączyk, and Bozena Kostek. 2025. Depression analysis and detection using machine learning: Incorporating gender differences in a comparative study. *IEEE Access*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and 1 others. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik.

Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

June-Woo Kim, Haram Yoon, Wonkyo Oh, Dawoon Jung, Sung-Hoon Yoon, Dae-Jin Kim, Dong-Ho Lee, Sang-Yeol Lee, and Chan-Mo Yang. 2025. Domain adversarial training for mitigating gender bias in speech-based mental health detection. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7. IEEE.

Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H Robert, and 1 others. 2015. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124.

Sanne Koops, Sanne G Brederoo, Janna N De Boer, Femke G Nadema, Alban E Voppel, and Iris E Sommer. 2023. Speech as a biomarker for depression. *CNS & Neurological Disorders-Drug Targets-CNS & Neurological Disorders)*, 22(2):152–160.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Yuxin Li, Eng Siong Chng, and Cuntai Guan. 2025a. Hierarchical self-supervised representation learning for depression detection from speech. *arXiv preprint arXiv:2510.08593*.

Yuxin Li, Sinchana Kumbale, Yanru Chen, Tanmay Surana, Eng Siong Chng, and Cuntai Guan. 2025b. Automated depression detection from text and audio: A systematic review. *IEEE Journal of Biomedical and Health Informatics*.

Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 35–42.

Michelle Morales, Stefan Scherer, and Rivka Levitan. 2018. A linguistically-informed fusion approach for multimodal depression detection. In *proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 13–24.

Muhammad Muzammel, Hanan Salam, Yann Hoffmann, Mohamed Chetouani, and Alice Othmani. 2020. Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis. *Machine Learning with Applications*, 2:100005.

Alice Othmani, Daoud Kadoch, Kamil Bentounes, Emna Rejaibi, Romain Alfred, and Abdenour Hadid. 2021. Towards robust deep neural networks for affect and depression recognition from speech. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 5–19. Springer.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. 2022. A step towards preserving speakers' identity while detecting depression via speaker disentanglement. In *Interspeech*, volume 2022, page 3338.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, and 1 others. 2019. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Afef Saidi, Slim Ben Othman, and Slim Ben Saoud. 2020. Hybrid cnn-svm classifier for efficient depression detection system. In *2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, pages 229–234. IEEE.

Asif Salekin, Jeremy W Eberle, Jeffrey J Glenn, Bethany A Teachman, and John A Stankovic. 2018. A weakly supervised learning framework for detecting social anxiety and depression. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(2):1–26.

Björn Schuller and Anton Batliner. 2013. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and

Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pages 16607–16629.

Ermal Toto, ML Tlachac, and Elke A Rundensteiner. 2021. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4145–4154.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Adrián Vázquez-Romero and Ascensión Gallardo-Antolín. 2020. Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, 22(6):688.

Michael Wagner and Duane G. Watson. 2010. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7-9):905–945.

Dong Wang, Yanhui Ding, Qing Zhao, Peilin Yang, Shuping Tan, and Ya Li. 2022. Ecapa-tdnn based depression detection from clinical speech. In *Interspeech*, pages 3333–3337.

Jinhan Wang, Vijay Ravi, and Abeer Alwan. 2023. Non-uniform speaker disentanglement for depression detection from raw speech signals. In *Interspeech*, volume 2023, page 2343.

World Health Organization. 2025. Depressive disorder (depression). WHO Fact Sheet, accessed 2026.

Wen Wu, Chao Zhang, and Philip C Woodland. 2023. Self-supervised representations in speech-based depression detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024a. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024b. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19323–19331.

Masahiro Yamamoto, Akihiro Takamiya, Kazutaka Sawada, Masaru Yoshimura, Masahiro Kitazawa, Kuang Liang, Takashi Fujita, Masaru Mimura, and Taishiro Kishimoto. 2020. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PLOS ONE*, 15(9):e0238726.

Michael Yang, Abd-Allah El-Attar, and Theodora Chaspari. 2024. Deconstructing demographic bias in speech-based machine learning models for digital health. *Frontiers in Digital Health*, 6:1351637.

Muqiao Yang, Joseph Konan, David Bick, Anurag Kumar, Shinji Watanabe, and Bhiksha Raj. 2022. Improving speech enhancement through fine-grained speech characteristics. *arXiv preprint arXiv:2207.00237*.

Pingyue Zhang, Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2021. Depa: Self-supervised audio embedding for depression detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 135–143.

Wenyu Zhang, Yingxu He, Geyu Lin, Zhuohan Liu, Shuo Sun, Bin Wang, Xunlong Zou, Jeremy HM Wong, Qiongqiong Wang, Hardik Bhupendra Sailor, and 1 others. 2025a. Beyond classification: Towards speech emotion reasoning with multitask audiollms. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1132–1148.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024. When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 146–158.

Xiangyu Zhang, Hexin Liu, Qiquan Zhang, Beena Ahmed, and Julien Epps. 2025b. Speecht-rag: Reliable depression detection in llms with retrieval-augmented generation using speech timing information. *arXiv preprint arXiv:2502.10950*.

Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C Oliveira, Robert Desimone, Dominic Dwyer, and Zongyuan Ge. 2025. It hears, it sees too: Multimodal llm for depression detection by integrating visual understanding into audio language models. *arXiv preprint arXiv:2511.19877*.

Ziping Zhao, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Haishuai Wang, and Björn Schuller. 2020. Hierarchical attention transfer networks for depression assessment from speech. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech*

and signal processing (ICASSP), pages 7159–7163. IEEE.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Dataset Details and Preprocessing

We correct two known annotation inconsistencies in the labels. For DAIC-WOZ, we relabel subject 409 as depressed, following the preprocessing protocol of Bailey et al. (Bailey and Plumbley, 2021) [2]. For MODMA, although the original dataset provides psychiatrist-based diagnostic labels, we use PHQ-based binary labels for consistency with the depression-severity labeling protocol used in this study. Accordingly, we relabel subject 2010037 as a non-depression based on its PHQ score, consistent with Li et al. (Li et al., 2025a).

Demographic and clinical statistics for both datasets are summarized in Table 4.

For DAIC-WOZ, raw audio is segmented into participant utterances using transcript timestamps, and transcripts are available for transcript-based ablations. To account for transcript–audio misalignment in several recordings, we apply subject-specific timestamp offsets during preprocessing, shifting the timestamps by 34.000 s for subject 318, 3.355 s for subject 321, 6.070 s for subject 341, and 16.540 s for subject 362. MODMA is distributed with sentence-level segmentation.

B Training Details

B.1 Classifier Configuration

All compared classifiers are trained under the same optimization, batching, device, and early-stopping settings, unless otherwise specified.

Multilayer Perceptron (MLP) (Rumelhart et al., 1986): The MLP baseline first averages the variable-length set of utterance-level SC-Para vectors for each subject into a single seven-dimensional subject-level vector. This vector is then fed into a two-hidden-layer MLP classifier with hidden dimensions of 64 and 32, matching the classifier architecture used in Table 5. This model has 2,753 trainable parameters (∼2.7K).

LSTM (Graves, 2012): The LSTM baseline treats the utterance-level SC-Para vectors as a sequence ordered by utterance index. It uses a 2-layer LSTM with hidden size 64, and the final hidden state is passed to a two-hidden-layer MLP classifier head for subject-level classification. It has 58,369 trainable parameters (∼58K).

Bi-LSTM (Schuster and Paliwal, 1997): The Bi-LSTM baseline uses the same ordered utterance sequence as input. It applies a 2-layer bidirectional LSTM with hidden size 64 in each direction, concatenates the final forward and backward hidden states, and passes the resulting subject-level representation to a two-hidden-layer MLP classifier head. It has 147,201 trainable parameters (∼147K).

Transformer with positional encoding (Vaswani et al., 2017): The Transformer baseline uses 2 encoder layers, hidden dimension 64, 4 attention heads, a feed-forward dimension of 256, and sinusoidal positional encoding. The utterance-level SC-Para vectors are encoded into 64-dimensional utterance representations and mean-pooled to form the subject-level representation, which is then passed to a two-hidden-layer MLP classifier head. It has 106,497 trainable parameters (∼106K).

Transformer without positional encoding (Vaswani et al., 2017): This variant follows the same input encoding, mean-pooling aggregation, and classifier-head design as the Transformer baseline above, but removes the sinusoidal positional encoding. It has 106,497 trainable parameters (∼106K).

Attention-based MIL: The attention-based MIL model uses additive attention with a learnable query to estimate utterance-level instance weights. It then computes a weighted sum of the original seven-dimensional SC-Para vectors and feeds the resulting subject-level representation to the two-hidden-layer MLP classifier. It has 3,329 trainable parameters (∼3.3K). Additional details are provided in Section 3 and Table 5.

B.2 Implementation and Training Configuration

All experiments were implemented in PyTorch and executed on a workstation equipped with a single NVIDIA RTX 6000 GPU. The implementation and training configuration are shown in Table 5

Although we acknowledge several recent closed-source or unreleased LLM-based depression detection systems, direct comparison is not feasible because their reported results rely on different data

[2] https://github.com/adbailey1/daic$_{w}oz_{p}rocess$

Dataset	Gender	Group	N	PHQ severity range				PHQ mean (SD)
				0–4	5–9	10–19	20+	
DAIC-WOZ	Female	Non-depression	56	38	18	0	0	3.40 (3.09)
	Female	Depression	31	0	0	25	6	14.50 (4.19)
	Male	Non-depression	76	48	28	0	0	6.80 (5.92)
	Male	Depression	26	0	0	25	1	14.00 (3.22)
MODMA	Female	Non-depression	9	8	1	0	0	2.11 (2.20)
	Female	Depression	7	0	0	3	4	18.14 (5.15)
	Male	Non-depression	21	18	3	0	0	2.86 (2.17)
	Male	Depression	15	0	0	10	5	18.87 (2.92)

Table 4: Participant statistics for DAIC-WOZ and MODMA. Counts are stratified by dataset, gender, group, and PHQ severity range. DAIC-WOZ uses the PHQ-8 score (Kroenke et al., 2009), while MODMA uses the PHQ-9 score (Kroenke et al., 2001).

Category	Hyperparameter	Value
Input	Unit	Utterance-level SC-Para vector
Input	Dimension	7
Input	Grouping	Valid vectors from one subject
Attention	Projection	Linear, $7 \rightarrow 64$
Attention	Nonlinearity	Tanh
Attention	Query	Learned 64-d vector
Attention	Weights	Softmax over utterances
Aggregation	Readout	Weighted sum of original 7-d SC-Para vectors
Classifier	Architecture	Two-hidden-layer MLP
Classifier	Hidden dimensions	64, 32
Classifier	Normalization	BatchNorm after first hidden layer
Classifier	Activation / dropout	ReLU / 0.3
Classifier	Output	Scalar logit with sigmoid probability
Optimization	Optimizer	AdamW
Optimization	Learning rate / weight decay	5×10^{-4} / 1×10^{-4}
Optimization	Positive class weight	$N_{\text{neg}}/N_{\text{pos}}$
Optimization	Gradient clipping	1.0
Optimization	LR scheduler	ReduceLROnPlateau, factor 0.5, patience 3
Training	Maximum epochs / batch size	50 / 4
Training	Early stopping patience	10
Evaluation	Decision threshold	0.5

Table 5: Implementation and training configuration for ParaCalib

splits, modalities, or development-set evaluation protocols, and their implementations are not publicly available.

C Additional Analysis for Attribute Contributions

D ALM Captioning Inference Settings

We use Qwen3-Omni-30B-A3B-Captioner[3] as the frozen ALM captioner and serve it locally with an OpenAI-compatible vLLM[4] backend in `bfloat16` precision. Generation uses temperature $= 0$, top-$p = 1.0$, and a maximum generation length of 16,384 tokens.

[3]https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Captioner

[4]https://github.com/vllm-project/vllm

E PSE Inference Settings

We use DeepSeek-R1-Distill-Qwen-32B [5] as the PSE model and serve it locally with an OpenAI-compatible vLLM [6] backend in `bfloat16` precision. The maximum context length is set to 32,768 tokens. Generation uses temperature $= 0$ and top-$p = 1.0$.

F Unknown-State Filtering

Because PSE assigns `unknown` when the caption provides insufficient or ambiguous paralinguistic evidence, we exclude utterances with any `unknown` categorical SC-Para state from down-

[5]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B

[6]https://github.com/vllm-project/vllm

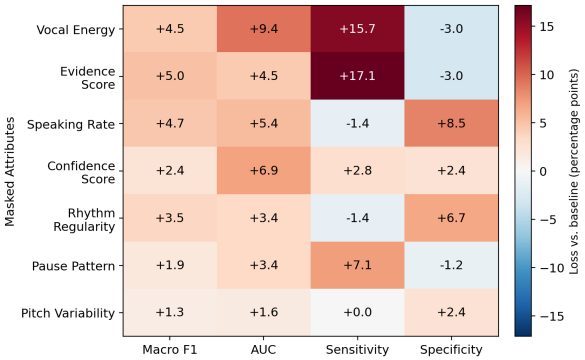

Figure 5: Metric-wise loss after masking each state. All values are reported in pp. Positive values indicate performance degradation and negative values indicate improvement.

stream subject-level pooling. To examine whether this filtering step removes a substantial or class-skewed portion of the data, we summarize discard rates by dataset, split, and PHQ label group in Table 12.

Overall discard rates are low across both datasets. In DAIC-WOZ, the utterance-level discard rate ranges from 4.7% to 6.2% across splits. In MODMA, the corresponding rate ranges from 0.6% to 2.5%. The class-specific utterance and per-subject discard rates are also similar between non-depression (ND) and depression groups, suggesting that unknown-state filtering does not substantially alter the effective class composition of the data.

G PSE Stability

To assess the reliability of PSE extraction, we conducted a repeated-decoding check on 100 randomly sampled ALM-generated captions. The PSE was applied three times using the same decoding configuration as in the main experiments. As shown in Table 6, the five categorical SC-Para states exhibit high exact agreement, ranging from 98% to 99%. The two holistic numerical dimensions also remain highly stable under the strict criterion of exact-value agreement, with 94% agreement for the depression-related paralinguistic evidence score and 95% agreement for extraction confidence. These results suggest that PSE extraction is largely stable.

H Human Validation

To assess the reliability of the intermediate representations used by ParaCalib, we conducted a small-scale human validation on 20 randomly sampled utterances. Two annotators independently evaluated two aspects of each utterance: (i) whether the ALM-generated caption was consistent with the audible content and vocal characteristics of the audio, and (ii) whether the five PSE-derived constituent paralinguistic states were consistent with the caption and perceived vocal evidence. We did not ask annotators to validate the holistic depression-related paralinguistic evidence score or extraction confidence, since these two dimensions require clinical judgment or model-level calibration beyond perceptual consistency annotation. We therefore treat this validation as a quality check for the caption and constituent-state extraction steps, rather than a full validation of all SC-Para dimensions.

The results show that the generated captions and the five PSE-derived constituent states are generally consistent with perceived vocal evidence. This supports the perceptual reliability of the caption-to-PSE extraction process within the scope of the five constituent paralinguistic dimensions, while leaving validation of the holistic depression-related paralinguistic evidence score and extraction confidence to future work.

Annotation protocol Two annotators were recruited from our research group with speech or NLP background. They were asked to evaluate 20 randomly sampled utterances. For each utterance, annotators listened to the audio, read the ALM-generated caption, and judged whether the caption was consistent with the audible content and vocal characteristics. They then reviewed the five PSE-derived constituent SC-Para states and judged whether these states were consistent with the caption and perceived vocal evidence. Annotators were instructed not to infer speaker identity, life situation, or clinical diagnosis, and to evaluate only perceptual consistency of the generated caption and constituent paralinguistic states. The holistic depression-related paralinguistic evidence score and extraction confidence were not evaluated by annotators.

I SC-Para Schema and Prompts

I.1 SC-Para Dimensions

SC-Para includes seven utterance-level dimensions, comprising five paralinguistic properties of speech—pitch variability, speaking rate, rhythm regularity, pause pattern, and vocal energy—and

SC-Para dimension	Fully consistent samples	Exact agreement
Pitch variability	99 / 100	99%
Speaking rate	99 / 100	99%
Rhythm regularity	99 / 100	99%
Pause pattern	98 / 100	98%
Vocal energy	98 / 100	98%
Depression-related paralinguistic evidence score	94 / 100	94%
Extraction confidence	95 / 100	95%

Table 6: Stability of PSE extraction under repeated decoding.

Validation target	Positive judgments	Strict pass	Rater agreement
Caption consistency	38/40 (95.0%)	18/20 (90.0%)	18/20 (90.0%)
PSE constituent-state consistency	36/40 (90.0%)	17/20 (85.0%)	18/20 (90.0%)

Table 7: Human validation of the generated captions and the five PSE-derived constituent paralinguistic states on 20 sampled utterances. Positive judgments aggregate two annotators. Strict pass means both annotators marked the item as consistent.

two meta-paralinguistic variables, a depression-related paralinguistic evidence score and an extraction confidence score. The evidence score summarizes the extent to which the observed vocal pattern in the caption is locally consistent with commonly reported depression-related vocal phenomena, while the confidence score reflects the reliability of the extraction. The evidence score is therefore intended as an utterance-level summary of commonly reported depression-related paralinguistic phenomena, not as a subject-level diagnostic prediction.

Utterance-level filtering is applied after PSE extraction. If any categorical state is assigned `unknown`, or missing, the corresponding utterance is excluded from downstream subject-level pooling. This rule prevents incomplete or low-evidence paralinguistic descriptions from being converted into arbitrary numerical values. The subject-level label is not used in this filtering step.

I.2 PSE Prompt Template

Table 9 shows the PSE prompt used to map contextualized vocal descriptions into SC-Para states.

I.3 Prompt for Transcript Masking in Audio Captions

I.4 Output schema and feature definitions

I.5 Example of ALM-generated Audio Caption

I.6 Example of Transcript-masked Audio Caption

J Baseline Representation Details

This section details the extraction configurations and dimension specifications for the baseline representations evaluated in Section 5.2. All representations are ultimately fed into the same attention-based MIL classifier described in Appendix B to ensure a fair comparison.

LLD-based Representations The raw continuous LLDs consist of 88-dimensional eGeMAPS (Yang et al., 2022) features extracted at the utterance level. For *Discretized LLDs*, we generate categorical states via training-set quantile thresholds. Specifically, we apply tertile-based splits for most acoustic dimensions and a median split for rhythm regularity.

SSL-based Representations We evaluate foundational self-supervised speech models by extracting high-dimensional acoustic embeddings. For each utterance, we use the hidden states from the last Transformer layer of each pretrained model and apply temporal pooling to obtain an utterance-level representation. Specifically, all three large models, Wav2Vec2-large-960h [7], HuBERT-large-ls960-ft [8], and WavLM-large [9], provide 1024-dimensional hidden representations.

Transcript-based For Transcript-based representation, each utterance transcript is encoded with a frozen SentenceTransformers all-MiniLM-L6-v2 [10] to obtain an utterance-level text embedding,

[7] https://huggingface.co/facebook/wav2vec2-large-960h
[8] https://huggingface.co/facebook/hubert-large-ls960-ft
[9] https://huggingface.co/microsoft/wavlm-large
[10] all-MiniLM-L6-v2

Criterion	Inconsistent by at least one annotator	Inconsistent by both annotators
Caption consistency	2/20	0/20
PSE constituent-state consistency	3/20	1/20

Table 8: Summary of potential inconsistency cases identified by human annotators during validation.

which is then used as the instance representation in the same attention-based MIL classifier.

For *LLD-PSE w/ transcript*, we provide the PSE with raw numerical LLD values together with the corresponding utterance transcript. The prompt input format is modified to extract the same 7-dimensional state space as SC-Para from acoustic numbers and text.

For *Transcript + WavLM*, we concatenate each utterance-level transcript embedding with its corresponding WavLM SSL embedding and feed the fused vector into the same attention-based MIL classifier.

Caption-based Representations For *Caption Embedding*, we directly encode the full ALM-generated natural language captions using the frozen Sentence-transformers [11] (Reimers and Gurevych, 2019), resulting in a 384-dimensional dense vector per utterance.

For *Transcript-masked Caption-PSE*, the captions are first processed by an LLM to suppress transcript-derived lexical and content-level cues, including direct quotes, explicit descriptions of what the speaker says, and topic-level semantic interpretations. Importantly, this masking step is not intended to remove semantically framed descriptions of vocal behavior. It retains descriptions of how the speaker speaks, such as prosody, pausing, rhythm, fluency, energy, perceived affective tone, voice quality, accent, and recording conditions. The prompt used for transcript masking is provided in Table 10, and an example of a transcript-masked caption is shown in Table 14. The resulting transcript-masked descriptions are then processed by the PSE to extract the same 7-dimensional SC-Para schema used in SC-Para.

K Attention-weighted SC-Para consistency

To examine whether the model attends to more internally consistent paralinguistic evidence for depressed subjects, we compute an attention-weighted SC-Para dispersion score for each subject:

$$D_b = \sum_i \sum_j \alpha_{b,i}\alpha_{b,j} d(x_{b,i}, x_{b,j}), \qquad (6)$$

where $d(\cdot,\cdot)$ is the normalized Hamming distance over the five categorical SC-Para states: pitch variability, speaking rate, rhythm regularity, pause pattern, and vocal energy. Lower D_b indicates that high-attention utterances share more similar SC-Para patterns. Since D_b is computed once per subject and attention weights are normalized within each bag, the analysis is robust to variable numbers of utterances across subjects. We compare D_b between depressed and non-depression subjects using a subject-level permutation test and report bootstrap confidence intervals.

As shown in Figure 6, depressed subjects showed lower attention-weighted SC-Para dispersion than non-depression subjects, with mean scores of 0.3617 and 0.3945, respectively. The mean difference was -0.0327 and was significant under the subject-level permutation test, $p = 0.0189$. This result was consistent with the Mann–Whitney U test, $p = 0.0467$, and the bootstrap 95% confidence interval for the mean difference,

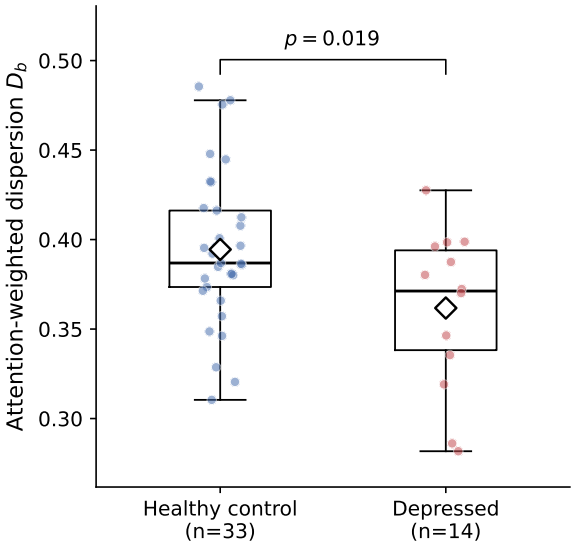

Figure 6: Attention-weighted SC-Para dispersion across non-depression and depressed subjects. Each point denotes one subject, boxes show the interquartile range, and diamonds indicate group means.

[11] all-MiniLM-L6-v2

$[-0.0593, -0.0081]$. Cliff's $\delta = -0.3723$ indicates a moderate effect. Overall, high-attention utterances in depressed sessions showed more internally consistent SC-Para patterns.

Table 9: PSE prompt template for converting ALM-generated contextualized vocal descriptions into structured SC-Para dimensions.

```
You are an expert in paralinguistic analysis of depression-related speech patterns.

IMPORTANT CONSTRAINTS:
- You CANNOT infer identity, mental diagnosis, or life situation.
- You are NOT making a medical diagnosis.
- Your task is to estimate the strength of depression-related paralinguistic evidence at
↪  the utterance level ONLY.

Input:
A caption generated by an audio-language model for a single utterance.
The caption may contain:
- acoustic or vocal descriptions
- semantic or contextual inference

Your task:
1. Extract acoustically observable vocal and paralinguistic information.
2. Based on vocal and paralinguistic patterns, estimate the presence and strength of
↪  depression-related speech characteristics.
3. If vocal information is sparse or ambiguous, use "unknown" and reduce confidence.

Depression-related paralinguistic patterns may include (but are not limited to):
- reduced pitch variability or flattened intonation
- monotonous or restricted prosodic range
- slow or irregular speaking rate
- increased pauses or hesitation
- reduced vocal energy or intensity
- breathy, tense, or hoarse voice quality
- unstable or reduced articulatory clarity

Output STRICTLY in the following JSON format.
Each JSON corresponds to ONE utterance.

{
  "prosody": {
    "pitch_variability": "flat/moderate/high/unknown",
    "speaking_rate": "slow/medium/fast/unknown",
    "rhythm": "steady/irregular/unknown",
    "pauses": "few/moderate/many/unknown",
    "energy": "low/medium/high/unknown"
  },
  "depression_related_paralinguistic_evidence": {
    "value": 0-100,
    "confidence": 0.0-1.0
  }
}

If information is insufficient, use 'unknown' and lower confidence.

IMPORTANT: Output ONLY valid JSON. No extra text outside JSON.
```

Table 10: LLM prompt for masking transcript-derived lexical and content cues in audio captions while preserving vocal-behavioral descriptions.

```
You are an audio description filtering expert. Your task is to filter out speech content
↪  transcription from audio descriptions and keep only the audio feature descriptions.

Content to REMOVE:
1. Specific words and phrases spoken by the speaker (direct quotes)
2. Detailed descriptions of conversation content
3. Explanations of semantic meaning and topics discussed

Content to KEEP:
1. Speaker's voice characteristics (timbre, tone, pitch, voice quality)
2. Speaking patterns (pace, rhythm, pauses, cadence)
3. Emotional features (calm, excited, hesitant, etc.)
4. Environmental sounds (background noise, echo, room acoustics)
5. Accent and language characteristics
6. Non-verbal sounds (breathing, coughing, laughter, clicks)
7. Recording technical quality descriptions

Return ONLY the filtered description without any explanations or additional comments. If
↪  the entire description is transcription content, return an empty string.

USER_PROMPT_TEMPLATE = """Filter the following audio description, removing all
↪  transcription content and keeping only audio feature descriptions:

{caption}

Filtered description:"""
```

Table 11: SC-Para output schema and feature definitions used by PSE.

```
Feature definitions:

prosody.pitch_variability:
   - "flat": reduced pitch variation, monotonous or restricted intonation.
   - "moderate": normal or balanced pitch variation after semantic calibration.
   - "high": large pitch variation that cannot be explained only by transcript context.
   - "unknown": insufficient evidence.

prosody.speaking_rate:
   - "slow": slow speech tempo after semantic calibration.
   - "medium": normal or balanced speaking rate.
   - "fast": fast speech tempo after semantic calibration.
   - "unknown": insufficient evidence.

prosody.rhythm:
   - "steady": regular, consistent speech rhythm with no notable disruptions.
   - "irregular": disrupted, uneven rhythm due to hesitations, repairs, or fragmented
   ↪  phrasing.
   - "unknown": insufficient evidence.

prosody.pauses:
   - "few": little evidence of pausing or hesitation.
   - "moderate": some pauses or mild hesitation.
   - "many": frequent pauses, long pauses, or disrupted fluency.
   - "unknown": insufficient evidence.

prosody.energy:
   - "low": weak vocal intensity, reduced loudness, or low vocal activation.
   - "medium": normal or balanced vocal energy.
   - "high": strong vocal intensity that is not only due to emphasis or emotion.
 - "unknown": insufficient evidence.

depression_related_paralinguistic_evidence.value:
   - A calibrated depression-related paralinguistic evidence score from 0 to 100.
   - Higher values mean stronger depression-related paralinguistic cues.


depression_related_paralinguistic_evidence.confidence:
   - A value from 0.0 to 1.0.
   - Higher confidence means the caption-based paralinguistic evidence is clear and
   ↪  reliable.
   - Lower confidence means paralinguistic evidence is weak, ambiguous, missing, or
   ↪  strongly dependent on transcript interpretation.

Output STRICTLY in the following JSON format.
Each JSON corresponds to ONE utterance.

{
  "prosody": {
    "pitch_variability": "flat/moderate/high/unknown",
    "speaking_rate": "slow/medium/fast/unknown",
    "rhythm": "steady/irregular/unknown",
    "pauses": "few/moderate/many/unknown",
    "energy": "low/medium/high/unknown"
  },
  "depression_related_paralinguistic_evidence": {
    "value": 0-100,
    "confidence": 0.0-1.0
  }
}
```

Dataset	Split	Overall	Utterance discard rate		Per-subject discard rate	
			Non-depression	Depression	Non-depression	Depression
DAIC-WOZ	Train	5.7%	6.0%	4.9%	6.1 (3.8)%	4.4 (3.5)%
	Dev	4.7%	4.6%	4.8%	5.0 (3.0)%	4.5 (2.2)%
	Test	6.2%	6.2%	6.2%	6.0 (3.7)%	7.0 (3.0)%
MODMA	Train	1.9%	1.4%	2.6%	1.4 (3.9)%	2.6 (4.7)%
	Dev	2.5%	2.8%	2.2%	2.8 (2.8)%	2.3 (2.8)%
	Test	0.6%	0.0%	1.1%	0.0 (0.0)%	1.1 (2.2)%

Table 12: Discard rates after PSE filtering. Overall reports the utterance-level discard rate over all utterances in each split. Utterance discard rate reports the discard rate separately for non-depression and depression groups. Subject discard rate reports the mean and standard deviation of per-subject discard rates within each PHQ label group. For MODMA, discard rates are reported for one representative cross-validation fold whose train/dev/test partitions together cover the full dataset; the main classification results use subject-level 5-fold cross-validation.

Table 13: Example caption generated by the frozen Qwen3-Omni-30B-A3B Captioner for an utterance-level audio clip.

```
The audio clip is a 7-second, high-quality recording made in a quiet, indoor environment,
↪ likely a home or small office. A single adult female, with a calm and measured
↪ demeanor, speaks in a General American English accent. The voice is close to the
↪ microphone, producing clear, detailed speech with a slight natural room reverb, but no
↪ significant background noise or distortion.

She lists five self-care activities in a slow, deliberate manner: ``read, take a long
↪ walk, hot bath, meditate, just close my eyes sometimes.'' Each item is separated by a
↪ pause, and the final item is delivered with a slightly softer, more introspective
↪ tone, as if reflecting on a personal habit. The last phrase trails off with a gentle,
↪ downward inflection, suggesting a sense of closure or personal reflection.

The choice of activities, reading, walking, taking a hot bath, meditating, and simply
↪ closing her eyes, points to a focus on mental and emotional well-being, relaxation,
↪ and stress reduction. The delivery is neutral and unhurried, lacking any strong
↪ emotion but implying a thoughtful, methodical approach to self-care. The overall
↪ context suggests this is an excerpt from a broader conversation, interview, or
↪ personal reflection, possibly relating to mental health, wellness, or coping
↪ strategies. There are no cultural or regional markers beyond General American English,
↪ and the clip is presented in a clean, unembellished style typical of contemporary,
↪ informal recording practices.
```

Table 14: Example transcript-masked audio caption that preserves semanticized vocal-behavioral and acoustic-context information while removing transcript-derived lexical and content-level cues.

```
The audio clip is a 7-second, high-quality recording made in a quiet, indoor environment,
↪ likely a home or small office. A single adult female, with a calm and measured
↪ demeanor, speaks in a General American English accent. The voice is close to the
↪ microphone, producing clear, detailed speech with a slight natural room reverb, but no
↪ significant background noise or distortion.

She lists several items in a slow, deliberate manner. Each item is separated by a pause,
↪ and the final item is delivered with a slightly softer, more introspective tone, as if
↪ reflecting on a personal habit. The last phrase trails off with a gentle, downward
↪ inflection, suggesting a sense of closure or personal reflection.

The delivery is neutral and unhurried, lacking any strong emotion but implying a
↪ thoughtful, methodical approach. There are no cultural or regional markers beyond
↪ General American English, and the clip is presented in a clean, unembellished style
↪ typical of contemporary, informal recording practices.
```