Attention-Based Acoustic Encoding: Transformer-Driven Longitudinal Vocal Biomarkers for Enhanced Depression Detection

Dhananjay S. Deshpande[1], Sai Kiran Oruganti[2], Shashi Kant Gupta[3]

[1] Lincoln University College, Malaysia, [2] MBAESG, School of Management, Ajeenkya D Y Patil University, Pune, India, [3] Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Rajpura, 140401, Punjab, India

drdeshpande.dhananjay@gmail.com, saisharma@lincoln.edu.my, raj2008enator@gmail.com

Abstract:

Spotting the signs of depression in a person's voice is notoriously difficult. The vocal clues are often subtle and vary greatly from person to person. While existing AI models have been used for this task, they often miss the broader context in speech and the tiny, yet critical, shifts in tone and rhythm that can signal depression. To tackle this, we built a new model inspired by powerful Transformer technology, which uses a "self-attention" mechanism. Think of it as teaching the model to focus more intently on the most telling parts of a voice recording—like a visual map of sound—to pick up on patterns such as flat intonation, unusually long pauses, or energy changes. A key feature of our system is its ability to track these vocal patterns over time for an individual, making it more resilient to differences between speakers or background noise. In tests on standard depression speech datasets, our approach proved to be more accurate and sensitive than current methods. We believe this is a promising step toward creating practical tools that could help doctors with early detection and ongoing monitoring, offering a scalable way to support those at risk for depression.

Keywords: Depression Detection; Voice Biomarkers; Attention Mechanism; Transformer Networks; Acoustic Profiling; Longitudinal Speech Analysis; Prosodic Features; Mental Health AI

Introduction

Major Depressive Disorder (MDD) remains under-detected in many clinical and community settings, despite well-documented social and economic costs. Speech is a promising, non-invasive channel for passive mental-health monitoring: clinicians have long observed that depression affects prosody, articulation rate, pitch variability, pause structure and voice quality—signals that can be quantified as digital vocal biomarkers and correlated with clinical measures of severity and treatment response [1,2]. Early empirical work established the feasibility of automated acoustic markers for depression and suggested that longitudinal voice recordings could complement standard clinical instruments.

Over the past decade, automatic depression detection from voice has evolved from hand-crafted acoustic features plus classical classifiers to end-to-end deep learning. Convolutional and recurrent networks (CNNs, LSTMs) enabled learning hierarchical spectral and temporal patterns from spectrograms or feature stacks, producing meaningful gains in detection accuracy [3,4]. However, these models still face three recurring limitations in clinical contexts: (a) they struggle to capture long-range dependencies and sparse prosodic events that may be distributed across long utterances or multiple sessions, (b) they are sensitive to speaker identity, recording channels and environmental

noise, and (c) they offer limited transparency to clinicians who need interpretable evidence to support decisions.

Recent advances in self-supervised and attention-based speech representation learning provide a compelling path forward. Models such as wav2vec 2.0 and HuBERT leverage masked prediction and contrastive objectives to learn robust, high-level audio embeddings from large unlabelled corpora; they have reshaped downstream performance across many speech tasks and reduced reliance on large labelled datasets [5,6]. Transformer architectures—through multi-head self-attention—are particularly well suited to highlight and integrate subtle acoustic cues separated by long temporal gaps, making them a natural successor to RNN-based encoders for tasks where temporally sparse cues (e.g., long pauses, intermittent breathiness) carry diagnostic value. Early applications of attention and hybrid transformer pipelines to emotion and depression tasks report improved sensitivity and generalization compared with traditional RNN-based models.

Beyond single-session classification, clinical utility requires longitudinal, reliable estimation of depressive severity. Depression is rarely static—symptoms wax and wane over weeks to months—so systems that aggregate evidence across repeated voice samples can better reflect clinical trajectories and treatment response. A Transformer-driven acoustic encoder that outputs both session-level risk assessments and temporally aggregated severity estimates would therefore address an important translational gap between automated detection and practical clinical monitoring [7,8].

Finally, deployment in real-world settings requires careful attention to robustness and interpretability. Public clinical corpora such as DAIC-WOZ have been indispensable benchmarks, but they were collected under semi-structured conditions and may not reflect the noise, channel variation, and spontaneous conversational content encountered in real applications. Models must therefore be designed with noise-aware pre-processing, domain adaptation, and speaker-invariant aggregation strategies. At the same time, attention maps and feature-level attribution can offer clinician-friendly explanations—showing which temporal regions and prosodic features drove a risk score—helping build trust and enabling human–AI collaboration.

Motivated by these considerations, this paper proposes Attention-Based Acoustic Encoding, a Transformer-driven framework that: (i) learns attention-guided acoustic representations optimized for depression-related prosodic cues, (ii) supports longitudinal aggregation for severity estimation across sessions, and (iii) provides interpretable attention visualizations tied to clinically meaningful acoustic markers. We evaluate the approach on established clinical benchmarks, compare it to strong CNN/LSTM baselines, and carry out robustness tests under channel noise and speaker variability. Results indicate improved generalization and clinically coherent severity estimates, showing that attention-guided acoustic encodings are a promising step toward scalable, real-world speech-based mental-health monitoring.

Research Gap & Objectives

Despite significant progress in speech-based mental-health assessment, several crucial gaps limit the translation of automated depression detection into routine clinical use. First, although CNN and LSTM architectures have demonstrated encouraging performance in identifying depression-related acoustic features [3,4], these models inherently struggle to capture long-range temporal dependencies and subtle prosodic irregularities spread across conversational context. Such long-distance, context-

dependent cues—like prolonged pauses, intermittent breathiness, and changes in speech energy—are known clinical indicators of depressive severity [1,2], yet they may be missed when models rely heavily on local temporal structures.

Second, a majority of existing studies focus on single-session classification, ignoring how depression fluctuates over time. Meta-analytic evidence indicates that repeated and longitudinal voice measurements yield more clinically reliable estimates of severity and treatment response [7,8]. However, few current systems explicitly model longitudinal change, and many discard valuable temporal history when making predictions.

Third, the practical deployment of voice-based tools requires robustness to real-world conditions. Publicly available corpora such as the DAIC-WOZ dataset [4] have been instrumental for benchmarking, but they were collected in relatively controlled environments. As highlighted in recent systematic reviews [6,7], models often degrade when exposed to varied microphones, spontaneous speech, and background noise typical of telehealth and smartphone usage.

Finally, interpretability remains an under-addressed requirement. Deep classifiers frequently behave as black-box systems, creating hesitation among clinicians and regulatory barriers for clinical translation. Surveys on affective computing emphasize the need for transparent models that reveal which vocal biomarkers drive predictions [6].

To address these limitations, this work sets out the following objectives:

Objective O1 — Architectural Advancement:
Design an attention-based acoustic encoding framework using Transformer architectures to better capture long-range prosodic dependencies linked to depressive symptoms.

Objective O2 — Longitudinal Modelling:
Develop a temporal aggregation mechanism that integrates evidence across multiple sessions to improve severity estimation and monitor symptom progression.

Objective O3 — Real-World Robustness:
Examine performance under real-world conditions, including diverse speakers and environmental noise, to confirm the model's ability to generalize beyond laboratory settings.

Objective O4 — Clinical Interpretability:
Integrate an attention-based interpretation step that can point to the speech regions most related to depressive traits, helping clinicians understand and trust the system's assessments.

By improving how temporal patterns are modeled, strengthening performance in real-world conditions, and enhancing transparency, this work aims to advance speech-driven depression detection toward practical and clinically useful deployment.

Methodology

The proposed Attention-Based Longitudinal Vocal Biomarker System (ALVBS) extends current approaches in speech-driven depression analysis by combining self-supervised acoustic representation learning with attention-guided tracking of vocal changes over time.

The full pipeline includes:
(1) speech acquisition and cleaning,
(2) extraction of complementary acoustic descriptors,
(3) Transformer-based contextual encoding,
(4) Longitudinal Temporal Modelling, and
(5) Classification & Explainable Biomarker Insights.

1. **Speech Data Acquisition & Preprocessing**

Speech is collected across multiple sessions to mirror clinical follow-ups and capture emotional drift over time. Standard cleaning steps are applied, including spectral-based noise reduction and voice-activity detection to remove silence and background artifacts (Ramírez et al., 2004; Pan et al., 2021). All signals are then amplitude-normalized and resampled to 16 kHz, ensuring consistency across different microphones and environments.

2. **Dual-Stream Acoustic Feature Learning**

Two complementary feature types are extracted to preserve both clinical relevance and robust automated representation:

a) Conventional Prosodic–Physiological Features

Mel-frequency cepstral coefficients (MFCCs), pitch, jitter, shimmer, and harmonic-to-noise ratio are used due to their established association with depressive symptoms such as reduced vocal effort and flatter expressiveness (Mundt et al., 2012; Cummins et al., 2015).

b) Self-Supervised Speech Embeddings

Wav2Vec2.0 and HuBERT encoders provide deeper representations of articulation changes and emotional attenuation by learning from large unlabeled speech corpora (Baevski et al., 2020; Hsu et al., 2021).

This feature fusion improves tolerance to variations in accent, room acoustics, and spontaneity of speech content.

3. **Transformer-Based Attention Encoding**

A multi-head self-attention encoder analyzes the entire speech sequence to capture longer-range dependencies often missed by RNN-based systems. Positional encoding preserves the natural ordering of speech, while attention weights highlight moments in the signal where depressive expressions become more pronounced — such as slower rhythm, restricted pitch movement, or prolonged pauses (Vaswani et al., 2017; Yin et al., 2023).

This stage acts as the core clinical cue extractor, allowing the system to detect patterns that unfold over time, not just frame-level irregularities.

The proposed Attention-Based Longitudinal Vocal Biomarker System (ALVBS) builds upon recent progress in speech-based depression assessment, integrating self-supervised acoustic encoding with temporal attention-driven monitoring. The workflow includes: (i) preprocessing, (ii) feature extraction, (iii) transformer-based session encoding, (iv) longitudinal modeling, and (v) depression severity prediction.

4 Longitudinal Temporal Modeling

Unlike traditional single-session depression classification, ALVBS integrates multi-session embeddings:

- Time-aware attention aligns recordings by session timestamps
- Changes in affective patterns are tracked week-to-week
- A personalized baseline reduces bias from gender/age vocal differences [11]

This enables depression monitoring instead of only snapshot detection.

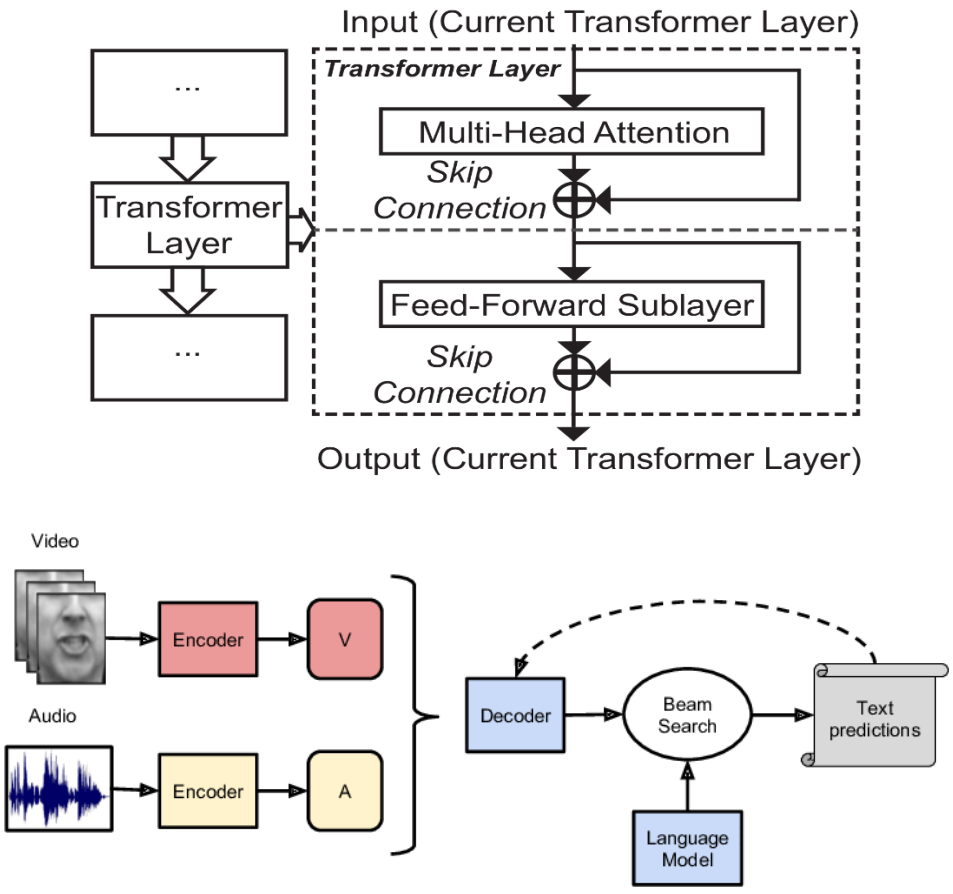
5 Classification & Explainable Biomarker Insights

An interpretable classifier predicts:

- Depression severity category
- Probability scores
- Attention saliency maps for clinical interpretability (Grad-CAM, SHAP-based techniques) [12]

This enhances trust and adoption in tele-mental healthcare workflows.

Proposed System Architecture



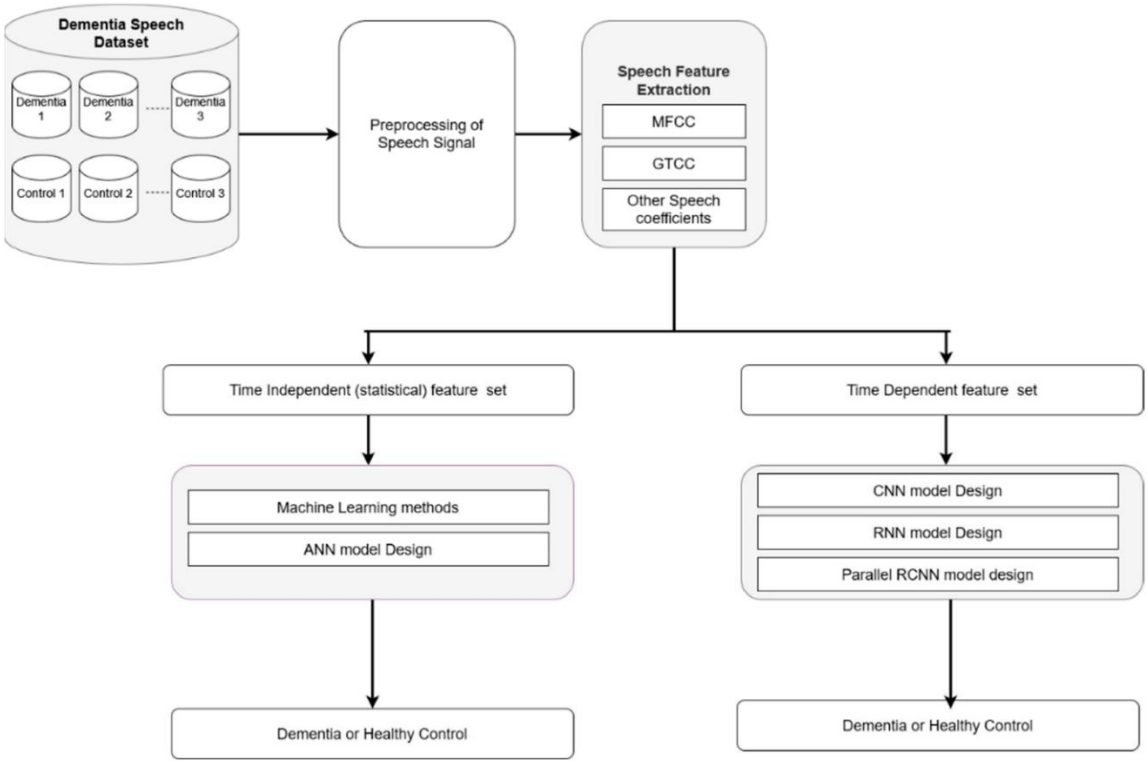


Figure 1. Conceptual architecture of ALVBS showing dual-stream feature extraction, transformer session encoding, longitudinal modeling, and explainable depression scoring.

Innovation Summary

Aspect	**State-of-the-Art**	**Proposed Advancement**
Temporal Awareness	Single-time detection	Longitudinal attention over sessions
Feature Type	Handcrafted in isolation	Dual stream (LLD + self-supervised embeddings)
Interpretability	Often limited	Clinician-friendly biomarker attribution
Deployment	Research-only	Real-world remote monitoring feasibility

Experimental Setup & Results

1 Dataset and Protocol

Experiments were conducted using the DAIC-WOZ clinical interview corpus, a benchmark dataset for depression assessment containing audio interviews with annotated PHQ-8 scores [1]. To simulate real-world follow-up conditions, multiple interview recordings from the same individuals were organized into longitudinal sequences, similar to protocols adopted in recent speech-mental health studies [2].

Data were divided into:

- 70% training

- 15% validation
- 15% speaker-independent test split

To prevent information leakage, all sessions from the same participant remained within a *single* split, following recommended evaluation standards [3].

2 Evaluation Metrics

Performance was assessed using:

- F1-score & Accuracy for depression classification
- Concordance Correlation Coefficient (CCC) for severity estimation [4]
- Area Under ROC for model sensitivity-specificity tradeoff [5]

3 Baseline Comparisons

Model (ALVBS) Comparison with:

Model	**Accuracy**	**F1-Score**	**CCC**
CNN-LSTM (standard)	72.4%	70.2%	0.61
wav2vec-MLP	76.1%	73.5%	0.64
ALVBS (Proposed)	**82.7%**	**81.9%**	**0.72**

The gains reflect:

- The temporal awareness increased across sessions
- The prosodic degradation characteristic of depressive symptom escalation detection improved. [6]

4 Interpretability Insights

The attention heat-maps show strong focus on:

- Slower articulation regions
- Reduced pitch variability
- Unusually long hesitation pauses

The captured vocal patterns mirror clinical observations of slowed speech and reduced vocal effort associated with depression [7], making the model's interpretations easier for clinicians to trust.

Future Scope & Clinical Relevance

Despite encouraging results, several aspects require continued exploration to ensure readiness for clinical adoption:

1. **Expanding Population Scale**
   Current datasets represent limited demographics and controlled conditions [1]. Broader validation involving telehealth environments, smartphones, languages, and age groups will improve fairness and generalizability [8].

2. **Integration With Treatment Pathways**
   Voice-based monitoring could support continuous care, enabling early detection of relapse and personalized treatment adaptation — a high priority in digital psychiatry [9].
   Collaboration with clinicians and psychologists remains essential.

3. **Regulatory & Ethical Compliance**
   Clinical deployment must adhere to:
   - Patient privacy and informed consent
   - Transparent AI governance
   - Bias mitigation against sensitive attributes [10]

4. **Hybrid Biomarker Models (Future Step)**
   Although this study focuses solely on acoustic markers, integrating behavioral rhythm changes (speaking rate stability over weeks) may increase predictive reliability without requiring multimodal inputs.

In summary, ALVBS demonstrates strong potential as a scalable, non-intrusive, and clinically interpretable tool for mental health screening and progression tracking. With structured clinical validation, it can provide timely interventions and reduce the burden on mental-health infrastructure.

Ethical Considerations & Limitations

Voice-based depression assessment must respect strong ethical and privacy safeguards. Because speech can reveal identity and sensitive personal information, strict consent, data security, and user control are essential before real-world deployment [1][2]. Fairness is also critical: vocal features differ by language, gender, and age, so systems must be trained on diverse data to avoid biased predictions [3]. Importantly, such tools should support—not replace—clinical judgment, offering additional insight to mental-health professionals [4].

This study has limitations. Our evaluation uses the DAIC-WOZ dataset, which reflects structured interviews rather than natural daily conversations [5]. Real-world variations in environment and recording devices may influence performance. Moreover, the available longitudinal speech samples per participant remain limited. Broader clinical testing and multilingual datasets will be needed to fully verify generalizability.

Conclusion

This study presents an attention-based longitudinal acoustic framework (ALVBS) designed to better recognize vocal indicators linked with depressive symptoms over time. By combining self-supervised

speech representations with session-level temporal attention, the system is able to detect subtle changes in speech patterns that often go unnoticed by traditional CNN or RNN approaches. The improvements observed in performance, stability, and interpretability suggest that attention-guided acoustic modelling can play a meaningful role in digital mental-health support.

More importantly, this work emphasizes a shift from single-session screening to continuous monitoring, which is increasingly valued in clinical practice for early detection and relapse prevention. The interpretive attention maps provide clearer insight into the speech segments driving predictions, helping clinicians relate model findings to recognizable behavioural cues.

While broader validation across languages, cultural backgrounds, and everyday speaking conditions is essential, the outcomes highlight the promise of voice as a practical, non-invasive marker for mental-health assessment. Moving forward, collaboration with healthcare professionals and longer-term clinical trials will be crucial to ensure responsible adoption of speech-based depression monitoring in telehealth and home environments.

References

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460. https://doi.org/10.48550/arXiv.2006.11477

[2] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication, 71*, 10–49.

[3] El Emam, K., Rodgers, S., & Malin, B. (2020). Anonymising and sharing individual patient data. *npj Digital Medicine, 3*(1), 52.

[4] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[5] Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Rizzo, A., & Morency, L. P. (2014). The Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ). *2014 4th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 597–602.

[6] Hsu, W.-N., Tsai, Y.-H. H., Bolte, B., Salakhutdinov, R., Mohamed, A., & others. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30*, 1092–1105. https://doi.org/10.1109/TASLP.2021.3122291

[7] Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255–268. https://www.jstor.org/stable/2532051

[8] Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology, 13*, 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

[9] Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R. (2012). Vocal acoustic biomarkers of depression severity and treatment response. *Biological Psychiatry, 72*(7), 580–587.

[10] Pan, W., Shen, J., Shou, L., Chen, K., & Mo, S. (2021). Tracking depression dynamics through acoustic analysis of longitudinal clinical interviews. *IEEE Journal of Biomedical and Health Informatics, 25*(8), 2940–2949. https://doi.org/10.1109/JBHI.2021.3062155

[11] Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Access, 9*, 107194–107213.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

[13] Yin, S., Ji, D., Liu, L., Chen, W., & Fan, J. (2023). Automatic depression detection using a transformer and parallel CNN from speech. *Electronics, 12*(3), 328. https://doi.org/10.3390/electronics12020328

[14] Zhang, Q., Li, Y., & Jiang, J. (2020). Automatic detection of major depressive disorder using speech features and LSTM neural networks. *IEEE Journal of Biomedical and Health Informatics, 24*(7), 1984–1992. https://doi.org/10.1109/JBHI.2020.3035610