*Human Systems Engineering and Design (IHSED2025),* Vol. 198, 2025, 512–519
https://doi.org/10.54941/ahfe1006805

# Daily Stress Detection Using Artificial Neural Network Based on Acoustic and Semantic Information From Speech

**Peixian Lu, Xingwei Jiang, and Shiguang Deng**

China Nuclear Power Engineering Co., Ltd., Beijing 100840, China

## ABSTRACT

Cumulative daily stress is harmful to the health of people and leads to productivity loss. Hence, timely detection of daily stress is of vital importance. Natural speech from real life is the recommended information to detect stress as a non-invasive way. This study aims to improve stress detection accuracy by combing the acoustic and semantic information from speech. Based on the speech database with real daily stress, we fused the acoustic and semantic features and developed a daily stress detection model using artificial neural network. The results showed that the model accuracy using acoustic information is 65.50% with a F1-score of 60.21%. The model accuracy using semantic information is 80.00% with a F1-score of 76.65%. By combining the acoustic information and semantic information, the model accuracy was improved to 90.75% with a F1-score of 89.25%. These results indicated the complementary effect of acoustic and semantic information on the daily stress detection. This study validated the effectiveness of detecting daily stress based on the combination of acoustic and semantic information from real speech. The model developed in this study can be applied to daily stress monitoring in daily life, offering valuable insights for stress management intervention to mitigate adverse health impacts.

**Keywords:** Daily stress detection, Artificial neural network, Acoustic, Semantic, Speech

## INTRODUCTION

Daily stress refers to the strain resulting from the challenges of everyday life, encompassing both predictable demands and unforeseen events (Piazza et al., 2013). Prolonged exposure to cumulative daily stress has been associated with adverse health outcomes and reduced productivity (Piazza et al., 2013). Consequently, the timely identification of daily stress is critical for mitigating its negative effects (Cohen et al., 1997).

Natural speech, as a non-invasive data source, is widely regarded as an effective medium for stress detection. From a speech production perspective, stress influences vocal characteristics by increasing muscle tension and altering respiratory patterns (Slavich, Taylor, and Picard, 2019). Additionally, stress can modify lexical and syntactic structures in speech, providing semantic indicators of stress (Scherer and Moors, 2019). Therefore, stress detection can be enhanced by leveraging both acoustic (prosodic and spectral features) and semantic (linguistic content) information (Slavich, Taylor, and Picard, 2019; Akçay and Oğuz, 2020).

 

Artificial Neural Network (ANN) is the predominant approach for analyzing speech signals in stress detection. However, the efficacy of ANN models is highly dependent on data quality, which determines their applicability in real-world settings (Reddy and Kuchibhotla, 2019). Labeled speech databases for stress recognition are typically derived from three sources: acted (simulated), elicited (induced in controlled settings), and natural (spontaneous real-life speech) (Akçay and Oğuz, 2020). While natural speech databases are less common in research, they are considered the most ecologically valid, as they reflect genuine stress expressions in authentic contexts (Sailunaz et al., 2018). Accordingly, this study employs a natural speech database to enhance the reliability of daily stress detection models.

Previous research has predominantly focused on unimodal approaches, developing stress detection models using either acoustic or semantic features in isolation (Akçay and Oğuz, 2020). In contrast, this study seeks to improve detection accuracy by integrating both modalities. The proposed multimodal framework holds practical significance for real-time stress monitoring in daily life, offering actionable insights for stress management interventions to reduce health risks associated with chronic stress.

## METHODS

In this study, based on the natural speech database with daily stress, we developed a daily stress detection model using artificial neural network by combining the acoustic and semantic information.

### Speech Database

The natural speeches were collected from the life-stress-catharsis chat rooms on the 'SOUL' online chat platform. This database contains 400 recordings. The stressor distribution in this database is shown in Table 1 using daily stressor categories in the study (Mauriello et al., 2021). Detailed information of this database is introduced in literature (Lu et al., 2024).

**Table 1:** Stressor distribution of the speech database.

| Category | Number of Recordings |
|---|---|
| Work | 33 |
| School | 27 |
| Financial Problem | 12 |
| Emotional Turmoil | 31 |
| Social Relationships | 123 |
| Family Issues | 81 |
| Health, Fatigue, or Physical Pain | 22 |
| Everyday Decision Making | 3 |
| Other | 24 |

### Feature Extraction

Acoustic features were extracted using the Munich open-Source Media Interpretation by Large feature-space Extraction (openSMILE) toolkit. To

test the applicability of different feature type sizes, we extracted the most classical baseline feature set B (Li, Dimitriadis, and Stolcke 2019), an expanded version of feature set $B^{+}$ and a reduced version of feature set $B^{-}$. A total of three feature sets were selected (see Table 2).

**Table 2:** Feature type and feature dimension of three acoustic feature sets.

| Feature Set | Feature Type | Feature Dimension |
|---|---|---|
| B | Root-Mean-Square Signal Frame Energy, Mel-Frequency Cepstral Coefficients MFCC 1-12, Fundamental Frequency (F0), Voicing Probability, Zero-Crossing Rate, Loudness, F0 Envelope, Line Spectral Frequency | 988 |
| $B^{+}$ | Expanded based on feature set B by including: Spectral Parameters, Voice Quality | 6373 |
| $B^{-}$ | Reduced based on feature set B by excluding: Loudness, F0 Envelope, Line Spectral Frequency | 384 |

Semantic features were extracted using the BERT-Base-Chinese model of the pretrained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019). The semantic features in each speech text were extracted by concatenating the output features of the last 4 layers of the BERT pre-training model, as recommended in previous study (Devlin et al., 2019). In total, 3072-dimensional feature vector containing semantic information were extracted from the text of each recording.

The feature fusion of acoustic and semantic features was implemented using a standard feature-level concatenation approach, as $\mathbf{x}_{\mathbf{fused}} = [\mathbf{x}_{\mathbf{acoustic}}, \mathbf{x}_{\mathbf{semantic}}]$. In this paper, three distinct acoustic feature sets were individually fused with semantic features, yielding three fused feature sets: fused feature set B, fused feature set $B^{+}$, and fused feature set $B^{-}$.

## Algorithm and Validation

In this study, we employed ANN algorithm for stress detection. The underlying theoretical framework of the ANN architecture is illustrated in Figure 1. The ANN comprised two hidden layers (32 neurons each, ReLU activation) and was trained via Adam optimizer (MSE loss). Hyperparameters were tuned via grid search (batch size = 10, epochs = 50), with final evaluation using 10-fold stratified cross-validation.

To assess classifier performance, we employed a 10-fold cross-validation approach. Considering the possible impact of data imbalance, we implemented stratified random sampling to partition the data into 10 subsets while preserving the original label distribution. In each iteration, nine subsets were used for model training, with the remaining subset reserved for testing (see Figure 2). The training set and testing set were pre-processed separately, including feature normalisation and feature dimensionality reduction. To assess model robustness, we performed ten repetitions of the 10-fold cross-validation procedure for each configuration. Performance evaluation was

conducted using four standard classification metrics: accuracy, precision, recall, and F1-score, offering a comprehensive assessment of model effectiveness across different aspects of classification performance.

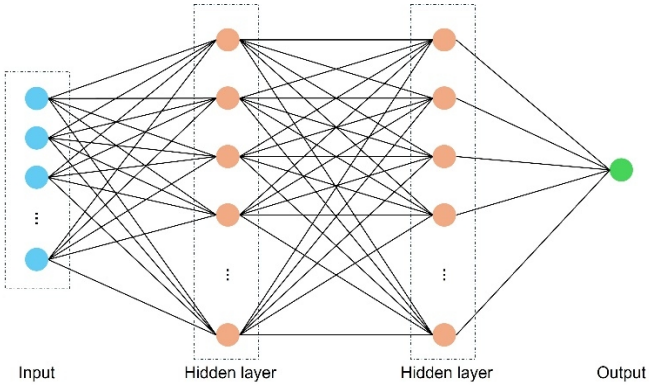


**Figure 1:** The theoretical framework of the ANN architecture.

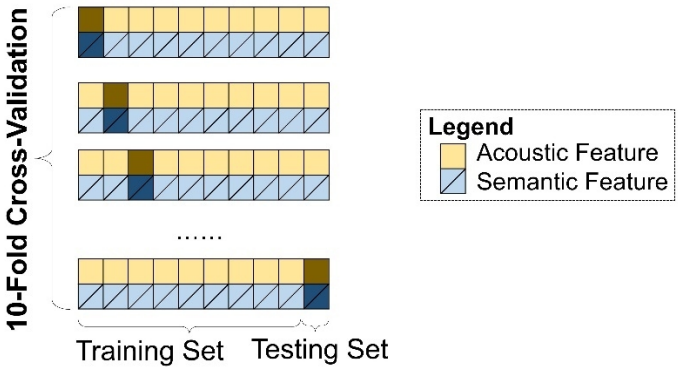


**Figure 2:** Illustration of 10-fold cross validation.

## RESULTS

### Model Performance

The model performances of acoustic, semantic, and fused feature sets are shown in Table 3, with their mean and SD values.

For the accuracy, semantic feature set achieved a value of 80.00%. The accuracy values of different acoustic feature sets varied from 62.50% to 65.50%, and the acoustic feature set $B^{+}$ achieved the highest accuracy. The accuracy values of different fused feature sets varied from 83.25% to 90.75%, and the fused feature set $B^{-}$ achieved the highest accuracy.

For the recall, semantic feature set achieved a value of 72.75%. The recall values of different acoustic feature sets varied from 56.43% to 68.42%, and the acoustic feature set B achieved the highest recall. The recall values of

different fused feature sets varied from 83.10% to 86.52%, and the fused feature set B⁻ achieved the highest recall.

For the precision, semantic feature set achieved a value of 82.57%. The precision values of different acoustic feature sets varied from 61.00% to 64.33%, and the acoustic feature set $B^+$ achieved the highest precision. The precision values of different fused feature sets varied from 84.78% to 92.93%, and the fused feature set B⁻ achieved the highest precision.

For the F1-score, semantic feature set achieved a value of 76.65%. The F1-score values of different acoustic feature sets varied from 57.40% to 63.61%, and the acoustic feature set B achieved the highest F1-score. The F1-score values of different fused feature sets varied from 80.95% to 89.25%, and the fused feature set B⁻ achieved the highest F1-score.

**Table 3:** Model performances of different feature sets.

| Feature Set | Feature Type | Accuracy | Recall | Precision | F1-Score |
|---|---|---|---|---|---|
| Acoustic | B⁻ | 62.50% (5.14%) | 56.43% (14.32%) | 61.00% (9.61%) | 57.40% (8.01%) |
| | B | 64.75% (6.92%) | 68.42% (14.65%) | 61.31% (9.77%) | 63.61% (8.95%) |
| | $B^+$ | 65.50% (2.84%) | 57.69% (10.46%) | 64.33% (4.52%) | 60.21% (5.36%) |
| Semantic | | 80.00% (3.33%) | 72.75% (10.57%) | 82.57% (5.71%) | 76.65% (5.60%) |
| Fused | B⁻ | 90.75% (6.78%) | 86.52% (12.58%) | 92.93% (5.69%) | 89.25% (8.68%) |
| | B | 87.50% (4.41%) | 83.10% (7.42%) | 89.99% (9.19%) | 85.97% (4.78%) |
| | $B^+$ | 83.25% (6.35%) | 77.69% (8.48%) | 84.78% (7.84%) | 80.95% (7.40%) |

## Effectiveness of Feature Fusion

Overall, the model performances were improved significantly by fusing acoustic and semantic features (see Figure 3). The unimodal approaches demonstrated baseline performance, with acoustic features achieving 65.50% accuracy and semantic features reaching 80.00% accuracy. Following feature fusion, the highest accuracy was improved to 90.75% using the fused feature set B⁻. The highest F1-score of acoustic features and semantic features were 63.61% and 76.65%, respectively. After feature fusion, the highest F1-score was improved to 89.25% using the fused feature set B⁻. Notably, feature set B⁻ consistently delivered maximum improvements across all metrics: improving accuracy by 28.25 percentage points, recall by 30.09, precision by 31.93, and F1-score by 31.85. These results demonstrate that the multimodal fusion strategy yields substantially better performance than unimodal approaches, with fused feature set B⁻ emerging as the optimal configuration.

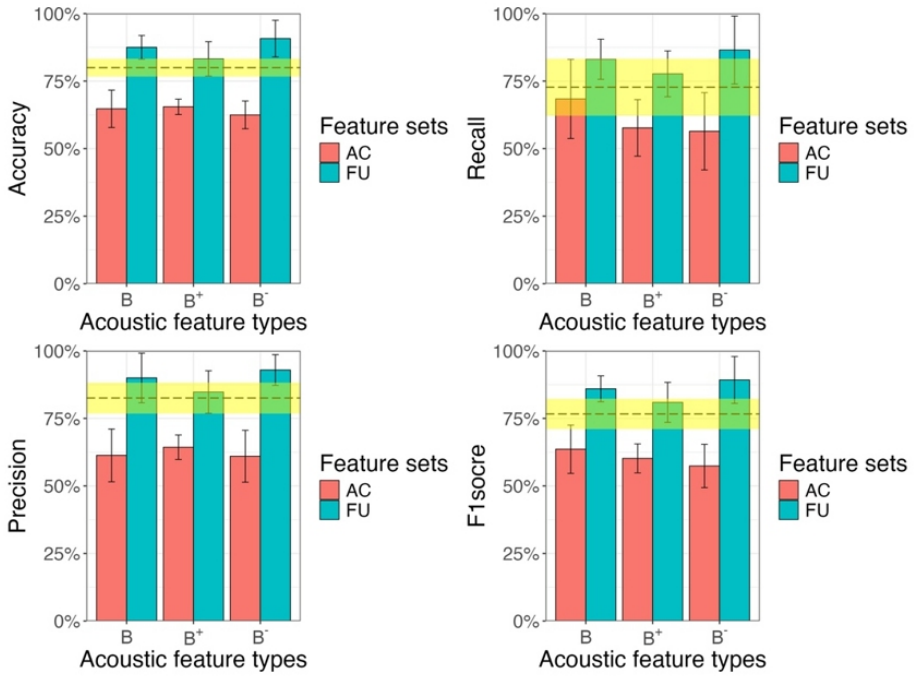


**Figure 3:** Comparison of model performances among different feature sets. The abbreviation in legend of "AC" and "FU" represent "acoustic feature set" and "fused feature set", respectively. The error bar represents standard deviation of model performance. The dash line and yellow shadow area represent the mean and standard deviation of semantic feature set, respectively.

## DISCUSSION

This study presents a novel bimodal approach for daily stress detection using natural speech. Leveraging a real-world daily stress speech database, we extracted complementary feature sets through: (1) acoustic analysis using openSMILE, and (2) semantic representation via the BERT pre-training model. The fusion of these modalities created an enriched bimodal input that significantly enhanced stress detection performance compared to unimodal approaches (acoustic-only or semantic-only). These results indicated the complementary nature of these modalities, where each capture distinct stress-related patterns. The experiment results validated the practical feasibility of detecting daily stress based on bimodal signals of acoustic and semantic features in real-world scenarios.

The experiment results demonstrate that models utilizing fused feature sets consistently outperformed unimodal approaches relying solely on either acoustic or semantic features. This finding not only validates our initial hypothesis (Lu et al., 2024) but also aligns with established research in emotion recognition (Wu & Liang, 2011). The performance improvement suggests that acoustic and semantic features capture complementary aspects of stress expression in speech - while acoustic features reflect physiological changes in vocal production, semantic features encode cognitive and

linguistic markers of stress. This complementary nature enables more comprehensive stress detection when both modalities are combined, supporting the theoretical framework of stress manifestation in speech communication.

The results further reveal that semantic-only models outperformed acoustic-only approaches, suggesting that advanced semantic feature extraction (e.g., through BERT's deep transformer architecture) may potentially achieve comparable detection performance without acoustic analysis — particularly advantageous for resource-constrained or real-time applications. However, acoustic features retain unique value in scenarios where semantic analysis proves unreliable, such as when speakers intentionally mask stress through language or when utterances contain ambiguous content. This underscores the importance of developing enhanced acoustic feature extractors to complement semantic analysis, as their synergistic combination could yield further improvements in stress detection robustness, particularly for challenging cases where either modality alone might fail.

The comparative analysis revealed distinct performance patterns among the fused feature sets. The ANN model utilizing fused feature set $B^{—}$ achieved optimal performance, contrasting with acoustic-only experiments where feature set $B^{+}$ demonstrated superiority. This divergence can be attributed to several factors: while feature set $B^{+}$ contains more comprehensive acoustic features, it may also introduce additional noise. In unimodal acoustic analysis, ANN's inherent noise robustness effectively leveraged feature set $B^{+}$'s informational advantage. However, in the bimodal context, semantic features likely superseded the additional information from feature set $B^{+}$'s extended features while its noise component persisted, explaining feature set $B^{—}$'s superior performance. These findings yield that more features do not invariably improve performance.

## CONCLUSION

This study validates the efficacy of combining acoustic and semantic features for daily stress detection using natural speech. The findings demonstrate that the multimodal integration of vocal characteristics and linguistic content significantly enhances detection accuracy compared to unimodal approaches. The developed model presents a viable solution for real-world stress monitoring applications, enabling continuous, non-invasive assessment of stress levels during daily activities. The practical implementation of this model offers substantial potential for early stress identification and timely intervention, which may help prevent stress-related health complications and support mental wellbeing management.

## ACKNOWLEDGMENT

The authors would like to acknowledge Naiyu Gao for his guidance on neural network algorithms.

## REFERENCES

Akçay, M. B., and K. Oğuz. 2020. "Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers." *Speech Communication* 116: 56–76. doi: 10.1016/ j.specom.2019.12.001.

Cohen, S., Kessler, R. C., & Gordon, L. U. (Eds.). (1997). Measuring stress: A guide for health and social scientists. Oxford University Press on Demand.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Minneapolis, Minnesota: Association for Computational Linguistics. 1: 4171–4186.

Li, B., Dimitriadis D., and Stolcke. A. 2019. "Acoustic and Lexical Sentiment Analysis for Customer Service Calls." In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Brighton, United Kingdom, 5876–5880. doi: 10.1109/ICASSP.2019.8683679.

Lu, P., Tsao, L., & Ma, L. (25 Nov 2024). Daily stress detection from real-life speeches using acoustic and semantic information, Ergonomics, doi: 10.1080/ 00140139.2024.2430370.

Mauriello, M. L., T. Lincoln, G. Hon, D. Simon, D. Jurafsky, and P. Paredes. 2021. "SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems." In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, 1–7. doi: 10.1145/3411763.3451799.

Piazza, J. R., Charles, S. T., Sliwinski, M. J., Mogle, J., & Almeida, D. M. (2013). Affective reactivity to daily stressors and long-term risk of reporting a chronic physical health condition. Annals of Behavioral Medicine, 45(1), 110–120.

Reddy, L. L., and S. Kuchibhotla. 2019. "Survey on Stress Emotion Recognition in Speech." 4.

Sailunaz, K., M. Dhaliwal, J. Rokne, and R. Alhajj. 2018. "Emotion Detection from Text and Speech: A Survey." *Social Network Analysis and Mining* 8 (1): 28. doi: 10.1007/ s13278-018-0505-2.

Scherer, K. R., and A. Moors. 2019. "The Emotion Process: Event Appraisal and Component Differentiation." *Annual Review of Psychology* 70 (1): 719–745. doi: 10.1146/ annurev-psych-122216-011854.

Slavich, G. M., S. Taylor, and R. W. Picard. 2019. "Stress Measurement Using Speech: Recent Advancements, Validation Issues, and Ethical and Privacy Considerations." *Stress (Amsterdam, Netherlands)* 22 (4): 408–413. doi: 10.108 0/10253890.2019.1584180.

Wu, C. H., and Liang, W. B. 2011. "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels." *IEEE Transactions on Affective Computing* 2 (1): 10–21.