2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Noise-Robust Speech-Based Severity Assessment for Emergency Calls

Kanji Okazaki and Keiichi Watanuki

Abstract— **This study aims to automatically classify emergency calls into serious (life-threatening) and minor (non-life-threatening) cases using acoustic features and machine learning models, thereby contributing to automated triage support in emergency response systems. Two enhancement strategies—noise reduction and data augmentation—are investigated to improve robustness in real-world call environments. Accurate triage during emergency calls is critical for optimizing resource allocation and ensuring timely medical response. Building on our previous exploratory analysis of acoustic features, this study advances toward practical deployment by addressing two key challenges: noisy real-world conditions and limited training data. To mitigate background noise and enhance feature stability, Wiener filtering was integrated into the preprocessing pipeline. Data scarcity was addressed through augmentation strategies, including moderate pitch shifting (±2 semitones) as well as comprehensive augmentation with pitch, volume, and noise perturbations. Acoustic features—including fundamental frequency statistics, Mel-frequency cepstral coefficients, and spectral descriptors—were extracted from call recordings provided by the Tokyo Fire Department. Three classifiers (Logistic Regression, Support Vector Machine, and Random Forest) were trained and evaluated using stratified cross-validation. Performance was primarily assessed by area under the ROC curve (AUC) and recall, given the critical importance of minimizing false negatives in emergency triage. Results showed that noise reduction improved robustness, while full augmentation yielded the greatest gains in predictive accuracy, with Random Forest achieving an AUC of 0.93. These findings demonstrate the feasibility of acoustic-based severity classification in emergency calls and highlight the potential of recall-oriented decision-support systems for emergency dispatchers. Future work will focus on real-time implementation and integration into dispatch operations.**

I. Introduction

Previous studies have investigated the automation of emergency call analysis [1, 19], yet most approaches have focused on lexical or linguistic information rather than acoustic cues. Emergency call centers are vital for providing timely aid in life-threatening situations. Accurate severity classification is essential for effective resource allocation; however, operator judgment can be influenced by stress, caller ambiguity, and background noise. Automated speech analysis has therefore gained attention as decision support.

The primary objective of this study is to automatically distinguish serious (life-threatening) from minor (non-life-threatening) emergency calls based on acoustic features extracted from real-world call recordings.

Speech conveys rich acoustic cues beyond lexical content, such as urgency, stress, and emotion. Prior studies have shown the utility of features like fundamental frequency (F0), Mel-frequency cepstral coefficients (MFCCs), and spectral descriptors for classifying emotional and pathological states [2-4, 6, 11]. Furthermore, recent research has also emphasized the interpretability of MFCCs as robust acoustic biomarkers across various speech domains [18]. Our earlier work with Japanese emergency calls revealed significant differences between serious and minor cases and demonstrated the feasibility of severity prediction, though performance was limited by dataset size and variability.

Automated acoustic analysis has also been applied in pathological and urgent speech detection contexts, such as Parkinson's disease [7], COVID-19 detection [9], and evacuation message urgency [6], demonstrating the feasibility of speech-based decision support in real-world environments.

This study advances that work by examining two strategies: (1) noise reduction to address real-world background interference and (2) data augmentation to expand training data and improve robustness. We compare baseline, moderate augmentation (pitch shift), and full augmentation (pitch, volume, noise) across three classifiers—Logistic Regression, Support Vector Machine (SVM), and Random Forest (RF)—evaluated with ROC analysis.

This direction builds on prior work in speech emotion recognition and emergency communication studies [2–4, 17]. Earlier conversational analyses of personal emergency response calls [21] also emphasize the role of emotional and prosodic cues in conveying urgency, further motivating the present focus on acoustic triage modeling. The contributions of this paper are threefold:

1. Developing a robust preprocessing pipeline incorporating noise reduction and systematic augmentation.
2. Evaluating their impact on classification performance using multiple models and ROC-based comparisons.
3. Providing insights into the feasibility of integrating acoustic-based severity detection into real-world emergency triage systems.

Kanji Okazaki is with the Graduate School of Science and Engineering, Saitama University, Saitama 338-8570, Japan (corresponding author: phone: +81-90-7225-4245; e-mail: okazaki.k.293@ms.saitama-u.ac.jp).

Keiichi Watanuki is with the Graduate School of Science and Engineering and the Advanced Institute of Innovative Technology, Saitama University, Saitama 338-8570, Japan.

978-1-6654-5784-2/26/$31.00 ©2026 IEEE

II. Material and Methods

A. Dataset

This study utilized anonymized emergency call recordings provided by the Tokyo Fire Department. Two classes of cases were considered:

- Serious: life-threatening conditions such as cardiac arrest and severe trauma.
- Minor: non-life-threatening conditions judged at dispatch.

Approximately 200 call segments (balanced between the two classes) were included. All recordings were originally sampled at 48 kHz with 16-bit PCM resolution. Ethical approval was obtained in accordance with institutional guidelines.

B. Preprocessing

All audio was downsampled to 16 kHz using high-quality Kaiser-windowed sinc interpolation. Stereo files were averaged to mono, and signals were amplitude-normalized to the range [–1, 1] to prevent clipping.

Framing parameters were fixed as follows:

- Window length: 25 ms (400 samples)
- Hop length: 10 ms (160 samples)
- FFT size: 512
- Window function: Hamming or Hann depending on the feature type
- Mel filterbank: 40 bands (20–8000 Hz)

These parameters ensured consistency across MFCC, Mel-spectrogram, and other spectral analyses. For noise reduction experiments, we applied a Wiener filter (scipy.signal.wiener) as a front-end processing step.

C. Data Augmentation

To mitigate limited data size and improve robustness, two augmentation strategies were implemented:

1. Moderate augmentation:

- Pitch shift: ±2 semitones

2. Full augmentation:

- Pitch shift: ±2 semitones
- Volume perturbation: scaling by 0.8× and 1.2×
- Noise injection: Gaussian noise ($\sigma = 0.005$)

Each augmented signal was processed identically to the raw signal, and features were re-extracted. Output datasets were stored separately (Raw / Moderate / Full).

D. Feature Extraction

Acoustic features were extracted using Librosa and Parselmouth (Praat API). Extracted features included:

- Fundamental frequency (F0): Estimated using probabilistic YIN (librosa.pyin) with fmin=60 Hz, fmax=400 Hz. Derived statistics: mean, standard deviation, interquartile range, min, max, range, valid ratio, voiced frame count, delta standard deviation.
- Cepstral features (MFCCs): 13 MFCCs computed with 40 Mel filters, summarized by mean, SD, IQR, min, max, and range.
- Mel-spectrogram features: 40 Mel-band energies averaged across frames.
- Spectral descriptors: Centroid (mean, SD), bandwidth (mean, SD), roll-off (95%), zero-crossing rate (mean), RMS energy (mean, 90th percentile).

Perturbation measures such as jitter and shimmer [10] were attempted but excluded due to unstable estimation in noisy emergency call environments (frequent NaN values).

E. Statistical Analysis

Univariate feature comparisons between serious and minor groups were performed using:

- Mann–Whitney U test (non-parametric, two-tailed)
- Effect sizes: rank-biserial correlation (r), Cliff's δ, Cohen's d
- Multiple testing correction: Benjamini–Hochberg FDR

Promising features were identified using dual thresholds ($p < 0.05$, $|r| \geq 0.2$).

F. Machine Learning Modelling

Three classifiers were evaluated:

- Logistic Regression (L2 penalty, liblinear solver)
- Support Vector Machine (linear kernel)
- Random Forest (200 estimators)

All models were trained with Stratified 5-fold Cross-Validation. Missing values were imputed by mean, and features were standardized (z-scoring).

Performance metrics: Accuracy, Precision, Recall, F1, and Area Under the ROC Curve (AUC). ROC curves were aggregated across folds for visualization.

Traditional machine learning models—Logistic Regression, Support Vector Machine, and Random Forest—were chosen due to the limited dataset size (~200 samples) and the need for interpretability at the feasibility stage. Deep learning approaches generally require substantially larger datasets and extensive parameter optimization, which are planned for future work to further improve model generalization.

G. Example Data Visualization

To illustrate the dataset characteristics, Figure 1 presents a serious case with waveform, spectrogram, and fundamental frequency (F0) contour. This example demonstrates the

substantial acoustic variability and background noise encountered in real emergency calls, where overlapping speech, emotional fluctuation, and environmental interference often coexist. Such complexity motivates the incorporation of noise reduction and data augmentation strategies to ensure robust and practical model performance in deployment settings.

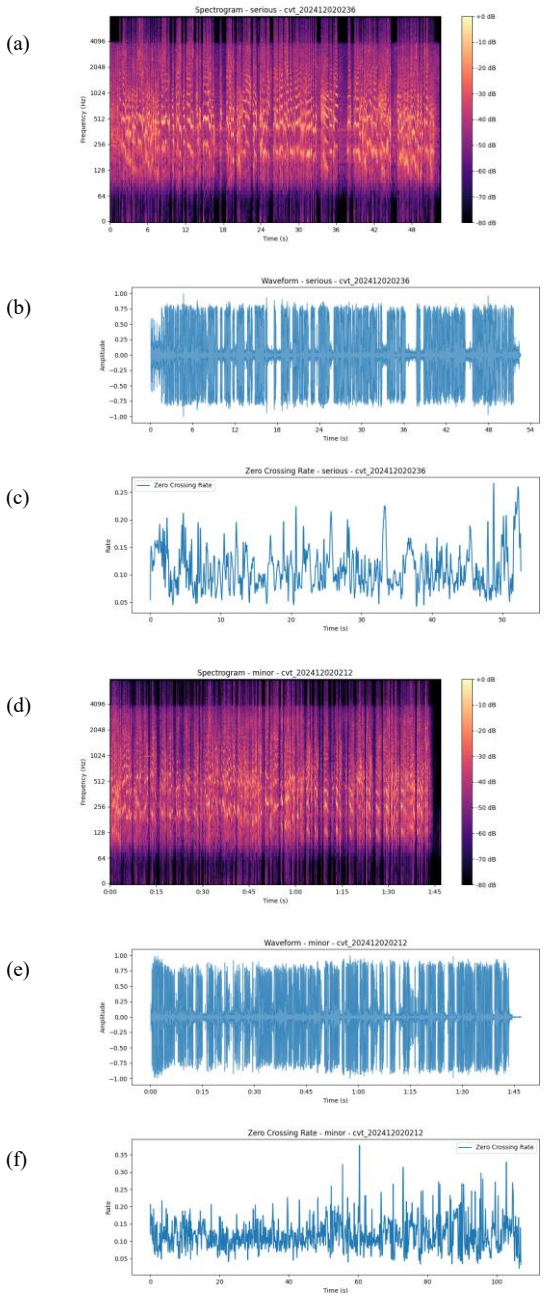

Figure 1. Examples of speech samples from both classes: (a–c) Serious case showing higher energy and F0 variability; (d–f) Minor case showing lower intensity and flatter F0 contour. These differences visually demonstrate the acoustic contrast between severity levels.

III. Results

A. Univariate Statistical Analysis

Univariate comparisons between the serious and minor classes revealed statistically significant differences across several acoustic features. In particular, Mel-frequency cepstral coefficients (MFCCs) and fundamental frequency (F0) descriptors exhibited strong discriminative power after false discovery rate (FDR) correction ($p < 0.05$). Figure 2 illustrates boxplots of the most salient features, whereas Figure 3 depicts a ranking of features based on –log10(FDR) values combined with rank-biserial correlation. Together, these results indicate that cepstral and spectral descriptors serve as critical acoustic markers for distinguishing between severity levels in emergency calls.

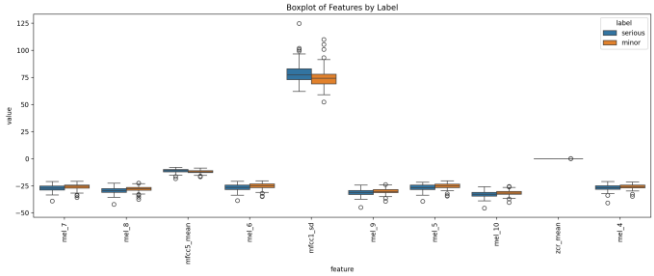

Figure 2. Boxplots of the top-ranked acoustic features (after FDR correction, $p < 0.05$) comparing serious and minor cases.

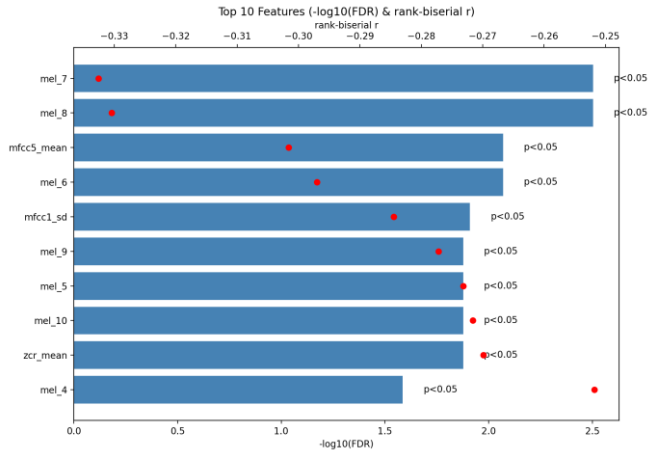

Figure 3. Ranking plot of acoustic features based on –log10(FDR) and rank-biserial correlation.

B. *Effect of Noise Reduction*

Baseline (raw audio) features were compared against Wiener-filtered features. Across all models (Logistic Regression, Support Vector Machine, Random Forest), Wiener filtering consistently improved prediction AUC. For example, Logistic Regression increased from AUC = 0.632 (raw) to 0.705 (Wiener). The overall comparison is summarized in Table I, while the corresponding ROC curves are shown in Fig. 4, highlighting the enhanced separability achieved through noise reduction. These results suggest that even basic noise suppression can stabilize feature distributions and enhance discriminative performance.

TABLE I. CLASSIFICATION PERFORMANCE (AUC) WITH AND WITHOUT WIENER NOISE

Model	Dataset	
	Wiener-filtered	***Raw***
LogReg	0.705	0.632
RF	0.653	0.634
SVM	0.693	0.596

a. Values represent mean AUC across 5-fold cross-validation. AUC: Area Under the ROC Curve; LogReg: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine.

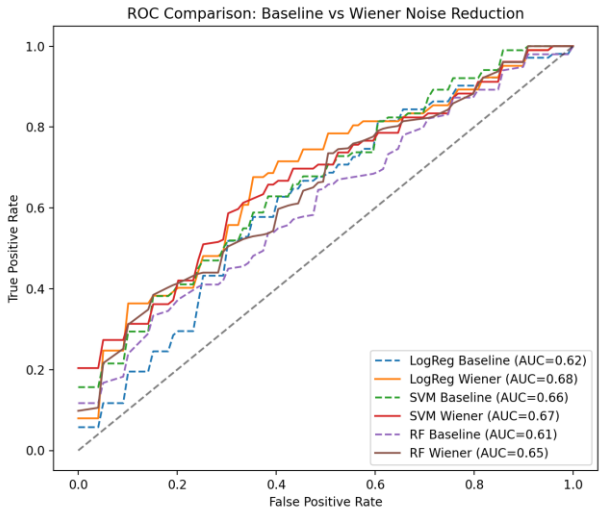

Figure 4. ROC curves demonstrating the effect of wiener noise reduction on classification performance

C. *Impact of Data Augmentation*

To address data scarcity, two augmentation strategies were evaluated:

- Moderate augmentation (pitch shifting ±2 semitones),
- Full augmentation (pitch shift, volume perturbation, noise injection).

Table II presents the mean AUC values across Raw, Moderate, and Full datasets. Both augmentation strategies improved model performance relative to the Raw baseline. The Full augmentation condition yielded the highest overall AUCs, particularly for the Random Forest model, which improved from AUC = 0.610 (Raw) to AUC = 0.929 (Full).

Table III summarizes recall values across Raw, Moderate, and Full augmentation conditions. The results highlight a clinically and operationally important trend: augmentation consistently reduced false negatives, allowing more severe cases to be correctly identified. Notably, Random Forest under Full augmentation achieved the highest recall (0.815), underscoring the importance of recall-sensitive evaluation for reliable emergency triage.

D. *ROC Analysis of Classification Models*

Figure 5 illustrates the ROC curves for all three datasets—Raw, Moderate, and Full—within this subsection for direct comparison with the numerical results presented in Table II and Table III. While Table I provides numerical comparisons, Figure 5 illustrates ROC curves for all three datasets. The ROC curves confirm that data augmentation not only increased AUC values but also improved recall, especially under the Full augmentation condition. Among the models, Random Forest exhibited the most substantial improvement, achieving the most favorable balance between sensitivity and specificity.

TABLE II. CLASSIFICATION PERFORMANCE (AUC) WITH DIFFERENT AUGMENTATION STRATEGIES

Model	Dataset		
	Full	***Moderate***	***Raw***
LogReg	0.789	0.705	0.632
RF	0.929	0.653	0.634
SVM	0.788	0.693	0.596

a. Values represent mean AUC across 5-fold cross-validation. AUC: Area Under the ROC Curve; LogReg: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine.

TABLE III. CLASSIFICATION PERFORMANCE (RECALL) WITH DIFFERENT AUGMENTATION STRATEGIES

Model	Dataset		
	Full	***Moderate***	***Raw***
LogReg	0.706	0.657	0.618
RF	0.815	0.647	0.529
SVM	0.721	0.578	0.559

a. Values represent mean AUC across 5-fold cross-validation. AUC: Area Under the ROC Curve; LogReg: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine.

IV. Discussion

A. Effect of Data Augmentation

The application of Wiener filtering demonstrated consistent improvements in classification performance across all models. Even a relatively simple denoising method enhanced feature stability, leading to higher AUC values compared to the raw baseline. These results highlight the importance of incorporating noise reduction techniques to handle the unpredictable acoustic environment of emergency calls.

B. Classifier Performance and Recall

Among the classifiers evaluated, Random Forest consistently achieved the best performance under augmented conditions, reaching an AUC of 0.929. Importantly, recall also improved substantially, rising from 0.529 in the Raw condition to 0.578 under Moderate augmentation and 0.815 under Full augmentation. This indicates that augmentation not only enhanced overall discriminative ability but also significantly reduced false negatives, allowing more severe cases to be correctly identified.

Although Random Forest exhibited the largest gains, Logistic Regression and SVM also showed notable recall improvements—for example, Logistic Regression increased from 0.618 (Raw) to 0.706 (Full), and SVM from 0.559 (Raw) to 0.721 (Full). These findings suggest that the benefits of augmentation are not model-specific but broadly applicable across classifiers. From a societal perspective, minimizing false negatives is critical, as overlooking severe cases can delay emergency response and lead to adverse outcomes. Consequently, recall-sensitive models such as Random Forest, when supported by augmentation, represent a particularly promising direction for deployment in emergency communication systems.

C. Robustness to Real-World Conditions

Emergency call environments are inherently noisy and unpredictable, often containing overlapping speech, emotional variation, and environmental interference. The present results indicate that augmentation-driven models generalize better to such challenging conditions. This finding aligns with prior research on speech recognition in noisy domains [7, 8, 15] and on data augmentation for robust speech modeling [13, 14], but extends its implications to the high-stakes context of emergency triage. In practice, integrating augmentation-enhanced models into existing call center infrastructure could provide real-time severity prediction, supporting dispatchers in making faster and more reliable decisions.

Furthermore, recent applications of acoustic sensing in emergency management have also demonstrated the feasibility of large-scale deployment in operational settings. For instance, Shams et al. [16] implemented an acoustic data detection system in emergency vehicle dispatch, showing that real-time processing can be integrated into large-scale infrastructure. This supports the practicality of incorporating severity-classification models into actual emergency response workflows.

D. Effect of Noise Reduction

This study was exploratory and constrained by a relatively small dataset, even after augmentation. While synthetic perturbations improved robustness, they may not fully capture the complexity of real-world variability, such as extreme emotional stress or simultaneous multi-speaker scenarios. Future work will therefore focus on expanding datasets through collaborative collection with emergency services, as well as incorporating advanced denoising and speech separation techniques (e.g., deep learning–based enhancement). Furthermore, in line with the recall-sensitive evaluation strategy, future studies will adopt cost-sensitive metrics and pursue field trials to validate real-world applicability. Similar noise suppression frameworks based on deep neural networks have been explored in recent studies [12], suggesting the potential for further improvement through hybrid or learned denoising approaches.

V. Conclusion

This study demonstrated that acoustic features extracted from emergency calls can be effectively leveraged to predict case severity using machine learning models. Baseline models trained on raw audio exhibited limited discriminative power, whereas data augmentation substantially improved performance, with the full augmentation condition yielding the best results. Notably, recall-sensitive evaluation confirmed that augmentation reduced false negatives, thereby lowering the risk of overlooking severe cases.

From a practical standpoint, these findings suggest that augmentation-enhanced models can serve as valuable decision-support tools for emergency dispatchers, enabling faster and more reliable triage in real-world environments. Integrating such models into call center operations could reinforce public safety infrastructure and mitigate adverse outcomes associated with delayed medical response.

Future work will focus on expanding dataset size through collaborations with emergency services, exploring advanced noise reduction and speech separation techniques, and validating model performance in field trials. Such collaborative efforts are expected to yield a more diverse and representative dataset for model generalization. These efforts are essential to transition from proof-of-concept to robust, socially deployable solutions. In future work, we also plan to evaluate deep learning architectures such as convolutional or transformer-based models once a sufficiently large corpus of emergency call data is available, thereby advancing toward a fully data-driven triage support system.

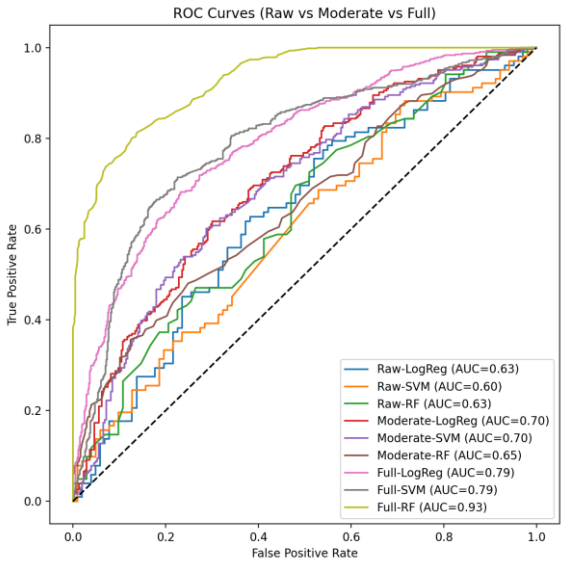

Figure 5. ROC curves showing the effect of data augmentation on classification on performance

ACKNOWLEDGMENT

The authors would like to acknowledge the Tokyo Fire Department for providing access to anonymized emergency call recordings used in this study. We also thank our colleagues and collaborators for their valuable technical advice and discussions.

REFERENCES

[1] Abi Kanaan, M., "A methodology for emergency calls severity prediction", [Online]. Available at: https://publiweb.femto-st.fr/tntnet/entries/19812/documents/author/data (Accessed: 24 August 2025).

[2] Deschamps-Berger, T., Lamel, L. and Devillers, L., "End-to-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings", arXiv preprint, arXiv:2110.14957, 2021.

[3] Fagherazzi, G., et al., "The use of vocal biomarkers from research to clinical practice", NPJ Digital Medicine, vol. 4, 12, 2021.

[4] Figueroa, C., et al., "Comparative study of acoustic parameters of voice and speech in adolescents with and without suicidal risk", Journal of Affective Disorders, vol. 354, pp. 154–163, 2024.

[5] Harada, S., Saito, Y. and Saruwatari, H., "Analysis of urgency in evacuation voice messages and its application to speech synthesis", Proceedings of the Acoustical Society of Japan, 2022.

[6] Ishihara, I., "Detection of pathological voices using the cepstrum method", Journal of the Acoustical Society of Japan, vol. 78, no. 9, pp. 496–499, 2022.

[7] Kadiri, S.R., Kodali, M. and Alku, P., "Severity classification of Parkinson's disease from speech using single frequency filtering-based features", arXiv preprint, arXiv:2308.09042, 2023.

[8] Kamble, M.R., Patiño, J., et al., "Exploring auditory acoustic features for the diagnosis of COVID-19", arXiv preprint, arXiv:2201.09110, 2022.

[9] Kim, S., et al., "COVID-19 detection model with acoustic features from cough sounds", Applied Sciences, vol. 13, no. 4, 2378, 2023.

[10] Li, X., "Stress and emotion classification using jitter and shimmer", Doctoral dissertation, Marquette University, 2007. Available at: https://epublications.marquette.edu/data_drdolittle/9/

[11] Li, G., Hou, Q., Zhang, C., et al., "Acoustic parameters for the evaluation of voice quality in patients with voice disorders", Annals of Palliative Medicine, vol. 10, no. 1, pp. 104–112, 2021.

[12] Nogales, A., Caracuel-Cayuela, J., and García-Tejedor, Á. J., "Analyzing the Influence of Diverse Background Noises on Voice Transmission: A Deep Learning Approach to Noise Suppression", Applied Sciences, vol. 14, 740, 2024.

[13] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition", Proc. Interspeech 2019, arXiv:1904.08779.

[14] Park, D. S., Zhang, Y., Chiu, C.-C., Chen, Y., Li, B., Chan, W., Le, Q. V., and Wu, Y., "SpecAugment on Large Scale Datasets", arXiv preprint, arXiv:1912.05533, 2019.

[15] Rashid, M., Alman, K.A., Hasan, K., Hansen, J.H.L. and Hasan, T., "Respiratory distress detection from telephone speech using acoustic and prosodic features", arXiv preprint, arXiv:2011.09270, 2020.

[16] Shams, M.Y., et al., "Acoustic data detection in large-scale emergency vehicle dispatch", Expert Systems with Applications, 2024.

[17] Shiotani, K., Kimura, K., Kitakoya, K. and Komatani, T., "Analysis of speech behavior in 119 emergency calls: Characterizing information transmission structures", Bulletin of Kansai University Society of Human Health, vol. 14, pp. 45–56, 2023.

[18] Tracey, B., Volfson, D., Glass, J., et al., "Towards interpretable speech biomarkers: exploring MFCCs", Scientific Reports, vol. 13, 22787, 2023.

[19] Valizada, A., "Development of speech recognition systems in emergency call centers", Symmetry, vol. 13, no. 4, 634, 2021.

[20] Yanuma, M., "Communication characteristics in reports of cardiac arrest by elderly callers: Examination of delay factors in telephone-assisted CPR", Bulletin of Kokushikan University Faculty of Letters, vol. 52, no. 2, pp. 33–42, 2021.

[21] Young, V., "Exploratory analysis of real personal emergency response call conversations", PMC, 2016. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5109662 (Accessed: 24 August 2025)..