Prioritization of 9-1-1 Emergency Calls Using Fusion of Audio Signal Analysis and Speech-to-Text Transcription for Accurate Urgency Classification

Simin Mirzaei, Hamid Reza Tohidypour, Panos Nasiopoulos, Deepan Chakravarthy, Fei Kuan, Leo Kamino, Mabel Wang, and Tayyib Chohan
Electrical & Computer Engineering, University of British Columbia
Vancouver, BC, Canada
Email: {siminmirzaei, htohidyp, panos}@ece.ubc.ca, {dsuresh, fei0316, lkamino, mabelw, tayyibc}@student.ubc.ca

***Abstract*—Overwhelmed 9-1-1 systems during large-scale emergencies can leave calls unanswered, delaying life-saving responses. This paper introduces a machine learning–based framework designed to prioritize calls by urgency. The proposed method integrates audio signal analysis and text transcription pipelines, fuses predictions using logistic regression, and applies a penalty to minimize false negatives. We constructed a dataset of 1351 labeled calls using a mix of public datasets, simulated calls, and data augmentation. Evaluations have shown that our proposed approach achieves 94% accuracy with a 1.5% false-negative rate, surpassing baseline models, and operates in real time. These results highlight the system's potential to enhance the reliability of emergency response by ensuring that the most urgent calls are identified and addressed promptly, thereby reducing delays and improving outcomes during large-scale crises.**

Keywords*—911 systems; speech signal processing; machine learning; speech emotion recognition.

I. INTRODUCTION

Emergency response systems form the backbone of public safety, ensuring rapid and reliable communication between citizens in crisis and trained dispatchers. Traditionally, 9-1-1 calls are screened by human operators, who assess their urgency and dispatch the necessary resources. While this model is effective during normal conditions, it faces severe limitations during large-scale emergencies such as natural disasters, major accidents, widespread power outages, or city-wide infrastructure failures [1]. In these situations, call volumes can surge well beyond the handling capacity of available operators. As a result, many calls may experience long delays in the queue—or worse, may never be answered at all. This overload undermines public trust in the emergency system and, more critically, endangers lives by delaying immediate attention to high-priority incidents.

Currently, call prioritization is performed manually by human operators on a first-come, first-served basis. While operators are highly trained, they are constrained by the sequential nature of human call handling, meaning that a low-priority call answered first may consume valuable time while a high-priority call waits unattended. The absence of an automated front-end screening mechanism leaves emergency communication systems vulnerable to failure precisely when they are needed most.

Although no direct solution currently exists for automated front-end call screening, related research has explored approaches aimed at improving emergency response efficiency. Artificial Intelligence (AI)-assisted dispatch systems utilize machine learning algorithms to analyze incoming emergency calls and provide dispatchers with real-time recommendations, improving resource allocation and response times [2]. Additionally, the integration of Internet of Things (IoT) technologies in smart cities has enabled the development of intelligent infrastructure that can detect and respond to emergencies more efficiently, further enhancing the effectiveness of emergency response systems [3]. Furthermore, Natural Language Processing (NLP) techniques have been employed to effectively bridge language barriers in emergency communication, thereby ensuring that critical information reaches all community members promptly and accurately [4].

To address the lack of an automated front-end screening mechanism, we propose a real-time, AI-driven framework that analyzes every incoming call within the first 10 seconds of audio. Upon answering, the system immediately processes the caller's voice to evaluate both acoustic features (e.g., pitch, intensity, speech rate) and linguistic cues from transcription (e.g., keywords like "fire," "unconscious," "bleeding"). Based on this analysis, the system classifies the urgency of the call and directs it accordingly: high-priority emergencies are immediately routed to a live operator, while low-priority or ambiguous calls are queued for recording and later review. In this way, the proposed solution acts as an intelligent filter at the network edge, ensuring that high-priority calls reach human operators without delay, even under heavy load conditions.

Building on recent advancements in speech emotion recognition [5], natural language understanding [6], and multimodal signal fusion [7], our framework leverages these technologies as the foundation for enabling real-time classification of emergency calls. These developments provide the technical basis that makes our proposed solution both feasible and effective under operational conditions. Our proposed framework aligns with the vision of Next Generation 9-1-1 (NG-911) systems, which transition from analog to Internet Protocol (IP)-based architectures capable of handling not only voice but also video, text, and location data [8]. Our proposed system enhances resilience, reduces operator overload, and improves overall responsiveness of critical public safety infrastructures. In terms of performance, our proposed framework demonstrates strong effectiveness and reliability under operational conditions. Evaluated against industry-standard requirements, it achieved an overall classification accuracy of 94% while maintaining a false-negative rate of just 1.5%. Minimizing false negatives is especially critical in emergency response scenarios, where

misclassifying a high-priority call as low priority could endanger public safety.

The remainder of the paper is organized as follows: Section II outlines our proposed approach and describes its key modules, Section III presents the evaluation results and corresponding discussion, and Section IV concludes the paper.

II. Our Approach

Our proposed approach leverages machine learning, signal processing and audio transcription to classify 9-1-1 calls by priority. Figure 1 shows the block diagram of our proposed end-to-end 911 framework.

As shown, our solution integrates two complementary processing pipelines: 1) an Audio Signal Analysis that extracts features such as pitch, intensity, and speech rate to assess caller stress and urgency, and 2) a Text Transcription Analysis where calls are transcribed into text, and evaluated based on semantic urgency cues (e.g., "unconscious," "bleeding," "fire"). The outputs from the two pipelines are fused producing a final importance score. The rest of the subsections describe in detail our dataset, preprocessing of the data and the key modules of our proposed method that include Audio Analysis, Text Analysis and Fusion.

A. Dataset

Our dataset consists of approximately 1351 audio samples, encompassing three distinct categories of calls. Real calls (n = 485) were collected from publicly available sources, including the Kaggle 911 dataset [9] as well as curated samples from the Austin Police [10] and True911 YouTube channels [11], providing authentic emergency communication data. Simulated calls (n = 416) were self-recorded to supplement the dataset, allowing controlled variation in acoustic conditions, speaker characteristics, and call scenarios not sufficiently represented in real data. Finally, non-phone calls (n = 451) were drawn from the VoxConverse dataset [12], comprising conversational speech samples used to model background, non-emergency, or low-priority audio content, thereby ensuring broader coverage of realistic input conditions for classification tasks.

All 911 calls in the dataset were manually annotated and categorized into two distinct classes according to their urgency and the nature of the incident [13]. High-priority calls were defined as those requiring immediate or rapid response due to their life-threatening or time-sensitive nature, including but not limited to homicides, active fires, medical crises and other critical emergencies posing imminent risk to individuals or property. Low-priority calls, by contrast, comprised incidents that did not necessitate immediate intervention, such as noise complaints, misdials, and prank calls, and were therefore classified as non-emergent in terms of operational response.

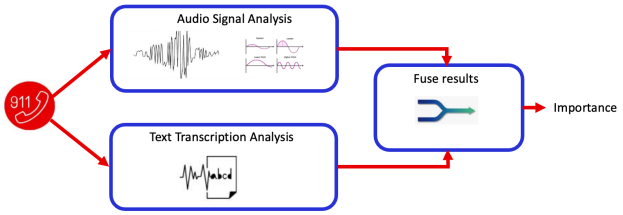

Figure 1. Block diagram of our proposed end-to-end 911 framework.

B. Preprocessing

All data were processed to normalize heterogeneous audio sources and ensure consistency for subsequent analysis. Audio conversion was first applied to normalize input formats: each sample was resampled to 8 kHz and encoded using the Adaptive Multi-Rate (AMR) codec, which reflects telephony-grade quality and closely approximates the spectral and temporal characteristics of real 911 call audio [14]. This step ensures uniformity across all audio samples, thus reducing domain mismatch during model training.

We used diarization to isolate the caller's voice while removing operator speech, hence ensuring that the classifier focuses exclusively on caller-specific acoustic and linguistic cues. To further strengthen model robustness, data augmentation techniques were applied to increase dataset variability and mitigate overfitting. Additive Gaussian noise was introduced at varying Signal-to-Noise Ratios (SNRs) to simulate realistic background interference. Tempo modifications (±5–10%) and temporal shifts were employed to mimic variability in speaking rate and transmission jitter [16][17]. These augmentations enhanced the generalizability of the models to diverse real-world acoustic environments commonly encountered in emergency call centers.

C. Our Methodology

Our solution integrates both acoustic and linguistic information to classify 911 calls into high- and low-priority categories. The overall framework is structured into two parallel pipelines—an audio pipeline and a text pipeline—followed by a fusion module that consolidates predictions to maximize robustness and accuracy. These are detailed in the following subsections.

1) Audio Signal Analysis Pipeline: In our audio signal analysis module, acoustic features were extracted from all emergency call recordings using the open-source Speech and Music Interpretation by Large-space Extraction (openSMILE) toolkit, a widely adopted framework in speech and paralinguistic research [18]. We employed the Computational Paralinguistics Challenge (ComParE) configuration, which generates 6373 Low-Level Descriptors (LLDs) and statistical functionals. These include prosodic features (e.g., pitch, jitter, shimmer) that capture stress-induced variations in voice, spectral features (e.g., Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, spectral flux) that describe the timbre and energy distribution of the signal, and voice quality measures (e.g., Harmonics-to-Noise Ratio (HNR)) that reflect vocal stability. All extracted features were normalized to zero mean and unit variance to ensure comparability across callers and recording conditions.

To reduce the number of extracted audio features, we employed a Random Forest classifier to handle classification

without explicit dimensionality reduction. Random Forests were chosen because they are robust to noisy and redundant features, can capture non-linear relationships, and provide interpretable probability outputs for decision-making. In our pipeline, openSMILE first extracts a comprehensive set of acoustic features, which are then input directly into the Random Forest classifier. The classifier outputs a probability vector that is subsequently used to estimate the urgency of each call, as illustrated in Figure 2.

2) Text Transcription Analysis Pipeline: All audio calls were first transcribed using OpenAI Whisper [20], which provides accurate, timestamped speech-to-text conversion under telephony-grade conditions. Each transcript was then processed with the MiniLM-L6-v2 sentence transformer [21], which maps sentences into a 384-dimensional embedding space, capturing their semantic meaning. These embeddings served as input to a three-layer feedforward neural network with two hidden layers (256 and 128 units, ReLU activations) and a final softmax layer that outputs a probability distribution over urgent versus non-urgent classes. To mitigate class imbalance and reduce the risk of misclassifying urgent calls, we trained the network with a weighted cross-entropy loss, assigning higher penalties to false negatives. Optimization was carried out using the Adam optimizer with a learning rate of 10^{-4}, and early stopping was employed to prevent overfitting. Figure 3 illustrates the proposed transcription analysis module, from speech transcription to semantic embedding and final classification.

3) Fusion Model: To integrate complementary cues from both acoustic and linguistic pipelines, we employed a late-fusion strategy. The audio pipeline produced class probabilities via a Random Forest, while the text pipeline generated predictions through a neural network. These outputs were concatenated and passed to a logistic regression meta-classifier, which learned an optimal weighting between modalities. This design allows the system to remain robust in cases where one pipeline alone may be ambiguous (e.g., calm tone but urgent verbal content).

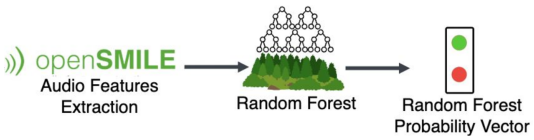

Figure 2. Block diagram of our proposed Audio Signal Analysis module.

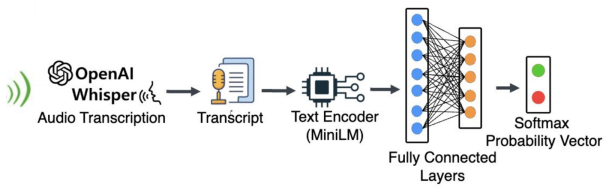

Figure 3. Block diagram of our proposed Text Transcription Analysis pipeline.

To account for the high cost of misclassifying safety-critical events, a cost-sensitive scheme was introduced by applying a threefold penalty when high-priority calls were predicted as low-priority. This adjustment biases the model toward recall of urgent cases, reflecting operational priorities in emergency response.

All pipelines (openSMILE for audio, Whisper for transcription, and sentence-transformers for text embeddings) rely on open-source tools, ensuring reproducibility, cost-efficiency, and local execution without dependence on proprietary cloud services. Figure 4 illustrates the overall architecture, comprising the audio analysis pipeline, text analysis pipeline, and the fusion stage.

III. Evaluation And Discussion

To assess the effectiveness of the proposed system, we conducted a series of experiments designed to reflect realistic emergency call scenarios.

A representative dataset of anonymized emergency call recordings was used. Calls were categorized into multiple urgency levels, providing the ground truth labels. For evaluation, audio was segmented into 10-second chunks, with both acoustic features (MFCCs, prosody) and linguistic features (transcribed text) extracted from each segment. System predictions were then compared against ground truth.

Our system achieved an overall classification accuracy of 94% across all urgency categories. More importantly, the false-negative rate—defined as the proportion of urgent (high-priority) calls incorrectly classified as low priority—was limited to 1.5%. This metric is critical in emergency contexts, as overlooking a high-priority incident poses direct risks to public safety. Compared to unimodal baselines (audio-only and text-only models), the proposed multimodal fusion reduced misclassifications, confirming the benefit of integrating acoustic and linguistic cues.

The prototype operated on 10-second increments, with each segment processed and classified within sub-second latency. This design ensures that ongoing calls can be re-assessed dynamically as new information emerges, fulfilling operational requirements for NG-911 infrastructures.

All components were implemented with on-premises feasibility in mind. While Whisper's API was used during prototyping for speech-to-text transcription, the model can be deployed locally in production using open-source alternatives, ensuring compliance with strict privacy and data protection regulations [23].

To contextualize performance, we compared our system against state-of-the-art Large Language Models (LLMs). For instance, ChatGPT-4 achieved 92.5% classification accuracy when operating on transcripts only, demonstrating strong language understanding, but falling below our solutions performance. Transcript-only systems cannot exploit acoustic features (e.g., stress, intonation) and face challenges in meeting real-time latency constraints. In contrast, our lightweight multimodal fusion model provided higher

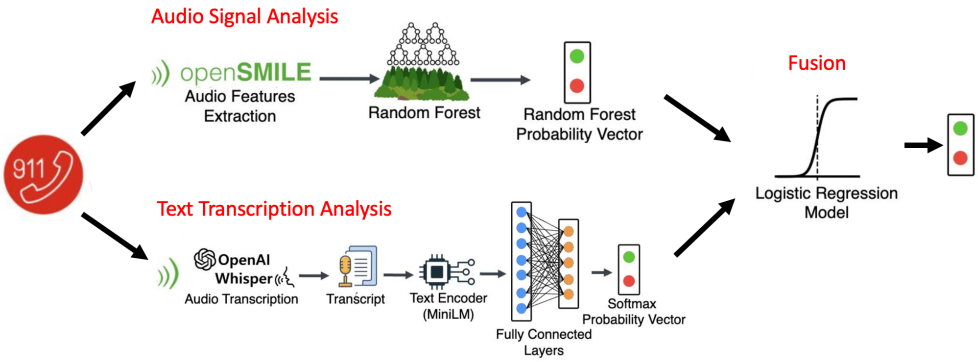

Figure 4. Block diagram of our proposed Text Transcription Analysis module.

reliability, particularly in reducing false negatives, while remaining deployable within real-world operational settings.

These results validate the design choice of employing a multimodal, cost-sensitive fusion model. By incorporating both signal and semantic features, our solution mitigates risks of misleading signals (e.g., calm tone during urgent emergencies) that could confound unimodal classifiers. Furthermore, the system demonstrates that interpretable, resource-efficient models—rather than large opaque architectures—can deliver mission-critical performance in emergency response scenarios.

IV. Conclusion and Future Work

This paper presented a framework for 9-1-1 emergency calls that fuses audio signal analysis with text transcription to assess call urgency. By leveraging both acoustic cues and semantic information, the system achieved 94% overall classification accuracy with a false-negative rate of only 1.5%, satisfying stringent public safety requirements. The framework operates in near real time, ensures privacy through on-premises deployment, and surpasses transcript-only approaches by integrating linguistic and paralinguistic signals. These results demonstrate that lightweight, interpretable multimodal models can provide dependable decision support for next-generation emergency response systems, alleviating operator overload and enhancing system resilience during high-demand scenarios.

Future work will expand the dataset and investigate advanced preprocessing and denoising techniques to further improve classification performance. Additional directions include evaluating model robustness under noisy or incomplete calls, exploring cross-lingual and dialect adaptation for diverse caller populations, and integrating real-time stress or emotion recognition. Another promising avenue is the development of adaptive models that learn continuously from operator feedback while maintaining transparency and accountability in decision-making.

References

[1] M. Chien, S. Chen, and L. Chen, “Emergency call systems under disaster scenarios: Challenges and solutions,” *IEEE Communications Magazine*, vol. 56, no. 12, pp. 79–85, Dec. 2018.

[2] C. Talukdar, “Real-Time Routing in IoT Networks for Emergency Response in Smart Cities,” *Risk Assessment and Management Decisions*, vol. 1, no. 2, pp. 322–328, Dec. 2024. [Online]. Available: doi: 10.48314/ramd.v1i2.55.

[3] Z. Li, “Leveraging AI automated emergency response with natural language processing: Enhancing real-time decision making and communication,” *Applied and Computational Engineering*, vol. 71, pp. 1–6, Aug. 2024.

[4] A. S. Santos et al., “Smart resilience through IoT-enabled natural disaster management: A COVID-19 response in São Paulo state,” *IET Smart Cities*, vol. 6, no. 3, pp. 211-224, Sep. 2024. [Online]. Available: doi:10.1049/smc2.12082.

[5] S. Sondhi, M. Khan, R. Vijay, and A. Salhan, “Vocal Indicators of Emotional Stress,” *Int. J. Comput. Appl.*, vol. 122, no. 15, pp. 38-43, Jul. 2015.

[6] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv:1907.11692, 2019.

[7] Z. Tu, B. Liu, W. Zhao, R. Yan, and Y. Zou, “A feature fusion model with data augmentation for speech emotion recognition,” *Appl. Sci.*, vol. 13, no. 7, p. 4124, Mar. 2023. [Online]. Available: https://doi.org/10.3390/app13074124.

[8] CRTC, “Telecom regulatory policy CRTC 2017-182,” 2017. [Online]. Available: https://crtc.gc.ca/eng/archive/2017/2017-182.htm. [Retrieved: September, 2025].

[9] L. Teitelbaum, “911 recordings,” Kaggle dataset, 2023. [Online]. Available: https://www.kaggle.com/datasets/louisteitelbaum/911-recordings. [Retrieved: September, 2025].

[10] Austin Police Department, “Austin Police YouTube Channel,” 2023. [Online]. Available: https://www.youtube.com/austinpolice. [Retrieved: September, 2025].

[11] intheblackmedia, “True 911 calls,” 2023. [Online]. Available: https://www.youtube.com/@True911calls. [Retrieved: September, 2025].

[12] A. Ye, “VoxConverse dataset,” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/washingtongold/voxconverse-dataset. [Retrieved: September, 2025].

[13] J. Arnett, “911 Dispatch Call Processing Protocols,” CSG Justice Center, 2023. [Online]. Available: https://csgjusticecenter.org/wp-content/uploads/2023/11/911-Dispatch-Call-Processing-Protocols-Key-Tools-for-Coordinating-Effective-Call-Triage.pdf. [Retrieved: September, 2025].

[14] VoiceAge, “AMR (Adaptive Multi-Rate) standard,” [Online]. Available: https://voiceage.com/AMR-NB.AMR.html. [Retrieved: September, 2025].

[15] OpenAI, “Whisper,” GitHub repository, 2025. [Online]. Available: https://github.com/openai/whisper. [Retrieved: September, 2025].

[16] R. M. Ben-Sauod, R. S. Alshwehdi, and W. I. Eltarhouni, “The Impact of data augmentation techniques on speech emotion recognition,” in International Conference on Information and Communication Technology for Intelligent Systems, Singapore, pp. 225-241, 2024.

[17] J. A. Nicolás, J. d. Lope, and M. Graña, “Data augmentation for deep learning in speech emotion recognition,” *Springer*, 2022.

[18] F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE – the Munich versatile and fast open-source audio feature extractor,” in Proc. ACM Multimedia (MM), Florence, Italy, pp. 1459-1462, 2010.

[19] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://link.springer.com/content/pdf/10.1023/a:1010933404324.pdf

[20] J. W. K. (OpenAI), “Whisper turbo release,” GitHub discussion, 2024.

[21] HuggingFace, “MiniLM-L6-v2: Sentence transformer,” 2025. [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. [Retrieved: September, 2025].

[22] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, Feb. 2013. [Online]. Available: https://www.wiley.com/en-us/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473. [Retrieved: September, 2025].

[23] OpenAI Whisper team, “Whisper: Automatic speech recognition,” 2025. [Online]. Available: https://github.com/openai/whisper/blob/main/README. [Retrieved: September, 2025].