TRIDENT: A Redundant Architecture for Caribbean-Accented Emergency Speech Triage

Galbraith, E., Sutherland, C., and Morgan, D.

SMG Labs Research Group

December 1, 2025

Abstract

Emergency speech recognition systems exhibit systematic performance degradation on non-standard English varieties, creating a critical gap in services for Caribbean populations. We present TRIDENT (Triage via Dual-stream Emergency Natural language and Tone), a three-layer triage architecture designed for resilience when automatic speech recognition fails. The system combines Caribbean-accent-tuned ASR, local entity extraction via large language models, and bio-acoustic distress detection to route emergency calls based on both transcription confidence and vocal stress indicators. **Our key insight is that low ASR confidence, rather than representing system failure, serves as a valuable triage signal—particularly when combined with elevated vocal distress markers indicating a caller in crisis whose speech may have shifted toward basilectal registers.** We demonstrate that treating uncertainty as signal enables more equitable triage performance for underrepresented speech varieties. We describe the architectural design, theoretical grounding in psycholinguistic research on stress-induced code-switching, and deployment considerations for offline operation during disaster scenarios. **This paper presents an architectural framework and position paper; empirical validation on Caribbean emergency calls remains future work.** This work establishes a framework for accent-resilient emergency AI that treats dialect variation as a routing feature rather than a transcription bug.

Keywords: automatic speech recognition, Caribbean English, emergency dispatch, vocal stress detection, creole continuum, edge computing, position paper

1 Introduction

When a caller dials emergency services during a crisis, the interaction between human distress and automated systems creates a critical dependency on speech recognition accuracy. Modern ASR systems, however, exhibit well-documented performance disparities across demographic groups [13]. For Caribbean English speakers—a population of over 40 million across the Anglophone Caribbean and diaspora—these disparities compound with a linguistic phenomenon: under acute stress, speakers tend to shift toward basilectal (more creole-heavy) speech registers, precisely the varieties on which ASR systems perform worst.

This paper presents TRIDENT (Triage via Dual-stream Emergency Natural language and Tone), an architectural framework designed not to eliminate ASR errors on Caribbean speech—an unrealistic goal given current technology—but to build a triage system that remains functional when such errors occur. **We frame this work as a position paper and system proposal**, establishing theoretical foundations and design rationale while acknowledging that end-to-end empirical validation on Caribbean emergency calls remains future work.

Our central and novel contribution is reframing low ASR confidence from a system failure into a triage signal that, combined with bio-acoustic distress indicators, routes calls appropriately to human dispatchers. This insight—that ASR uncertainty × acoustic stress fusion provides a robust triage mechanism—represents a conceptual shift from treating accent-induced errors as bugs to treating them as features that correlate with genuine caller distress.

The architecture addresses four gaps in existing emergency AI systems: (1) reliance on cloud-dependent, accent-agnostic ASR; (2) exclusive focus on textual features, ignoring paralinguistic stress signals; (3) no consideration of dialect continua or stress-induced register shifting; and (4) inability to function during infrastructure failures that commonly accompany disasters.

Note on stress-induced register shift: While we focus on Caribbean creole continua, the phenomenon of dialect reversion under cognitive load is not unique to this population. The inhibitory control model of bilingual processing [11] and research on the Lombard effect (speech modifications in noisy environments) suggest our framework may generalize to other bidialectal populations worldwide. Caribbean emergency services serve as our motivating case study, but the architectural principles apply broadly.

2 Related Work

The proposed crisis triage system draws on and extends research across four domains: automatic speech recognition for accented and low-resource speech varieties, artificial intelligence in emergency dispatch, vocal stress detection, and edge computing for disaster resilience. We review each in turn, identifying the gaps that motivate our three-layer architecture.

2.1 The Accent Gap in Automatic Speech Recognition

Modern ASR systems exhibit systematic performance degradation on non-standard English varieties—a disparity with serious implications for equitable access to voice-enabled services. Koenecke et al. [13] conducted the seminal quantitative study, evaluating five major commercial ASR systems across racial demographics. Their findings were stark: word error rates averaged 0.35 for Black speakers compared to 0.19 for White speakers, with 23% of Black speaker audio producing WER exceeding 0.50—functionally unusable transcription—compared to just 1.6% for White speakers. Critically, the researchers traced these disparities to acoustic models rather than language models, as the performance gap persisted even on identical phrases.

The Edinburgh International Accents of English Corpus (EdAcc) benchmark extends this analysis to global accent variation [19]. Testing revealed that OpenAI's Whisper-large model achieved 19.7% WER on EdAcc compared to just 2.7% on LibriSpeech test-clean—a seven-fold performance degradation on accented speech. The study specifically identified Jamaican English among the accents with highest error rates, directly validating concerns about Caribbean speech recognition.

Research on African-accented English provides methodologically rigorous comparators. The AfriSpeech-200 corpus encompasses 200 hours of Pan-African English speech across 120 accents from 13 Anglophone countries, with evaluations demonstrating that models achieving 1-3% WER on standard corpora produce 10-90% WER on African-accented subsets [17]. Named entities and domain-specific terminology proved particularly challenging—a finding directly relevant to emergency contexts where accurate location and hazard extraction is critical.

Caribbean English remains especially underserved despite representing millions of speakers. Madden et al. [16] developed the first substantial Jamaican Patois speech corpus (42.58 hours) and derived scaling laws for Whisper performance on this variety. Their results are instructive: pre-trained Whisper Large achieved 89% WER on Patois—functionally useless—while fine-tuned Whisper Medium reduced this to 30% WER. Notably, fine-tuned Whisper Tiny out-

performed non-fine-tuned Whisper Large, demonstrating that domain-specific data matters more than model size for underrepresented varieties. Their scaling law (WER = $158.06 \times M^{-0.255} \times D^{-0.269}$) reveals that dataset increases yield greater gains than model scaling for this population, informing our choice of Whisper Medium with Caribbean-specific fine-tuning.

2.2 AI-Assisted Emergency Dispatch

Emergency services worldwide are exploring AI-powered speech recognition and natural language processing to improve call handling efficiency and triage accuracy. The Emergency Calls Assistant (ECA) framework represents current state-of-the-art, achieving 92.7% accuracy in emergency classification using SVM with linear kernel on textual features [2]. The system operates in two phases—speech-to-text conversion followed by NLP classification—and compares favorably against commercial platforms including RapidSOS, Corti, and AlertGO.

However, critical examination reveals systematic gaps in existing approaches. ECA relies on Google Cloud Speech-to-Text API with no offline capability or accent adaptation. The system processes only transcribed text, ignoring paralinguistic stress markers that may indicate caller distress even when words are unclear. Furthermore, due to privacy restrictions on real emergency recordings, ECA was trained on synthetic datasets—raising questions about generalization to actual crisis communications.

Clinical validation studies demonstrate AI's potential while highlighting implementation challenges. Blomberg et al. [3, 4] evaluated the Corti AI system for cardiac arrest detection, finding that the ML system achieved 84.1% sensitivity compared to dispatchers' 72.5%, with faster time-to-recognition (44 seconds versus 54 seconds median). However, a subsequent randomized clinical trial found no significant improvement in dispatcher recognition when supported by ML alerts—suggesting that human-AI teaming requires careful interface design beyond raw model performance.

A scoping review of 106 AI studies in prehospital emergency care identified underutilization of multimodal inputs as a key gap [5]. No reviewed system integrated audio-based stress detection with text classification—precisely the multi-layer approach we propose. The review also noted the absence of systems designed for infrastructure-independent operation, a critical limitation for disaster response scenarios.

2.3 Vocal Stress Detection

The bio-acoustic analysis layer of our system builds on extensive research establishing acoustic correlates of psychological stress. A systematic review analyzing 38 peer-reviewed studies found that fundamental frequency (F0) is the most consistent stress marker, with 15 of 19 studies reporting significant mean F0 increases under stress conditions [21]. Intensity and amplitude increases showed similarly consistent patterns, while speech rate, jitter, and shimmer produced heterogeneous results across studies.

It is important to note methodological heterogeneity in this literature. While F0 elevation is the most replicated finding, some studies report null or contradictory results depending on stress type (acute vs. chronic), measurement methodology, and population characteristics. Hansen and Patil [12] found that certain stress conditions produce F0 *decreases* in some speakers, particularly under conditions of extreme fatigue or hopelessness. Our system design accounts for this by using F0 as one component of a multi-feature distress score rather than a sole indicator.

Research specifically examining emergency communications provides direct validation for our approach. Van Puyvelde et al. [22] analyzed real-life emergency recordings including cockpit voice recorders and 911 calls, documenting F0 increases from 123.9 Hz to 200.1 Hz during life-threatening emergencies—a 62% increase. F0 range expanded dramatically from 124.2 Hz to

297.3 Hz. Interestingly, jitter *decreased* during emergency stress, contrary to intuition, providing an additional discriminative feature. These findings directly inform our distress detection thresholds.

Studies of actual emergency call centers demonstrate both the promise and limitations of acoustic stress detection. Lefter et al. [15] achieved 4.2% Equal Error Rate for automatic stress detection in emergency telephone calls by fusing prosodic and spectral detectors—compared to 19% EER for individual detectors, highlighting the importance of multi-feature approaches. Demenko and Jastrzębska [6] found over-one-octave pitch shifts in highly stressful Polish police emergency calls, achieving 80-84% classification accuracy.

However, a critical reality check comes from Deschamps-Berger et al. [7], who found that while benchmark IEMOCAP data yielded 63% Unweighted Accuracy for emotion recognition, real emergency calls achieved only 45.6%—a substantial domain shift that deployment systems must account for. This finding reinforces our design decision to use bio-acoustic analysis as a triage signal rather than a sole decision-maker, routing high-distress calls to human dispatchers rather than attempting fully automated classification.

Recent work on multimodal fusion in emergency contexts supports our architecture. Feng and Devillers [8], analyzing the French CEMO emergency call center corpus, found that audio components often encode more emotive information than text in crisis contexts, with multimodal fusion yielding 4-9% absolute accuracy gains over unimodal models. This validates our approach of maintaining parallel ASR and bio-acoustic pathways that can compensate for each other's failures.

2.4 Dialect Reversion Under Cognitive Load

A theoretical foundation for Caribbean-specific ASR in emergency contexts comes from psycholinguistic research on bilingual processing under stress. The inhibitory control model establishes that non-target languages remain continuously active and must be suppressed through cognitive effort [11]. For Caribbean speakers navigating the creole continuum—from basilect (most creole features) through mesolect to acrolect (Standard English)—maintaining acrolectal speech requires sustained executive function.

The creole continuum is not simply a stylistic choice but a dynamic system of linguistic control, modulated by cognitive load. Research on cognitive load effects demonstrates that this inhibition fails under stress. Gollan and Ferreira [9] found that under high cognitive load, bilingual speakers use significantly less intraclausal code-switching, instead reverting to monolingual chunks of their dominant language. Importantly, cognitive load also affects lexical access timing—Kroll et al. [14] demonstrated that retrieval of L2 (non-dominant language) vocabulary slows significantly under dual-task conditions, providing a mechanism for stress-induced register shift.

Patrick's [18] foundational sociolinguistic analysis of the Jamaican Creole continuum establishes that stress levels influence speakers' positioning on this spectrum, with most speakers being mesolectal in normal conditions but capable of shifting toward either pole.

The implications for emergency services are significant: a professional who speaks Standard English at work may revert toward basilectal Patois when their house is flooding. Standard ASR systems, trained predominantly on acrolectal varieties, will exhibit precisely the performance degradation documented in the accent gap literature at the moment when accurate recognition is most critical. Our system addresses this by fine-tuning on Caribbean broadcast data that includes mesolectal speech, and by providing bio-acoustic fallback when ASR confidence drops—which may itself serve as a proxy indicator for basilectal reversion.

2.5 Edge Computing for Disaster Resilience

The case for offline-capable emergency AI is made starkly by infrastructure failure during recent disasters. Hurricane Maria's impact on Puerto Rico saw 95% of cell towers fail, with the entire island losing power and over 66% of the population lacking potable water [20]. Communication infrastructure failure caused delays in mortality reporting and created substantial information vacuums, contributing to a disputed death toll ultimately estimated at approximately 3,000. Recovery required over 200 days for full power restoration.

Recent advances in model compression make edge deployment increasingly feasible. Quantization studies demonstrate that 4-bit (INT4) quantization reduces Whisper model size by 45-87% with minimal WER degradation, and may actually reduce hallucinations by acting as a regularizer. Gondi and Pratap [10] demonstrated that transformer-based ASR achieves real-time inference on Raspberry Pi hardware with PyTorch mobile optimization. For the NLP component, 4-bit quantized Llama 3 8B runs at 2-5 tokens per second on Raspberry Pi 5—too slow for real-time conversation but adequate for background entity extraction tasks.

A survey of edge technologies for disaster management identifies prediction, detection, response, and recovery phases where edge computing enables real-time processing without cloud dependency [1]. The survey specifically identifies a gap in offline-capable speech and language processing at the edge—precisely the capability our system provides. Pre-positioned edge computing resources at hospitals, shelters, and emergency coordination centers, loaded with Caribbean-tuned models, could maintain triage capability even during complete grid and network failure.

2.6 Summary: Positioning Our Contribution

The literature reveals a clear opportunity space. Existing emergency dispatch AI systems uniformly exhibit: (1) reliance on cloud-dependent, accent-agnostic ASR; (2) exclusive focus on textual features, ignoring paralinguistic stress signals; (3) no consideration of dialect continua or stress-induced code-switching; and (4) inability to function during infrastructure failures.

Our three-layer architecture directly addresses each gap. Caribbean-accent-tuned Whisper provides the foundation that makes downstream NLP viable. Local Llama 3-based entity extraction operates without internet connectivity. Bio-acoustic distress detection provides a parallel triage signal that functions even when ASR fails—transforming low transcription confidence from a system failure into a routing feature. The result is the first crisis triage system designed specifically for Caribbean emergency services, capable of operating when communication infrastructure is most degraded.

3 System Architecture

TRIDENT implements a three-layer architecture where each component provides independent value while contributing to a unified triage decision. Figure 1 illustrates the system flow.

3.1 Layer 1: Caribbean-Tuned ASR

The ASR layer employs OpenAI's Whisper Medium model (769M parameters) fine-tuned with Low-Rank Adaptation (LoRA) on Caribbean broadcast speech. We selected Whisper Medium over Large based on Madden et al.'s [16] scaling law, which demonstrates diminishing returns from model size compared to domain-specific data for Caribbean varieties.

Fine-tuning Configuration:

- Base model: openai/whisper-medium
- Adaptation: LoRA (rank=16, alpha=32)

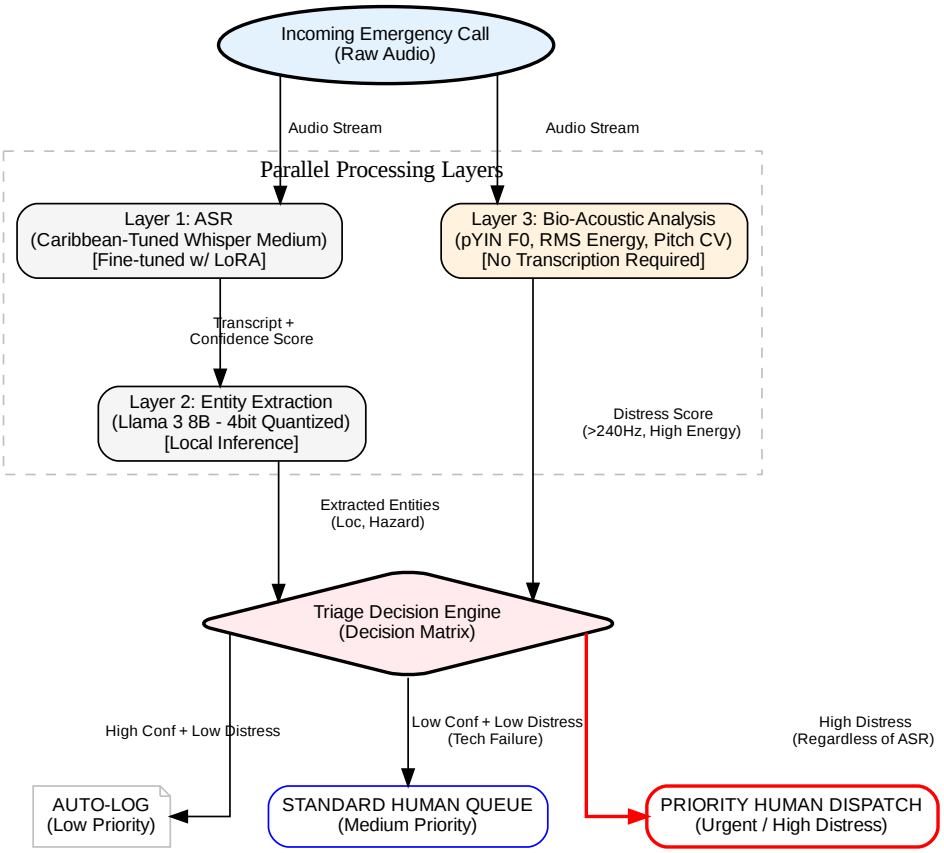

Figure 1: The TRIDENT architecture. The system processes raw audio through two parallel streams: (Left) A Caribbean-adapted ASR and NLP pipeline for semantic extraction, and (Right) a bio-acoustic analysis layer for determining physiological distress. The Triage Decision Engine integrates ASR confidence scores with vocal distress markers, ensuring that calls with low transcription confidence but high vocal distress (indicative of basilectal shift or acute crisis) are routed to priority dispatch rather than discarded.

- Training data: BBC Caribbean broadcast corpus (∼28,000 clips)
- Trainable parameters: ∼0.5% of total model

Confidence Scoring: The system computes **utterance-level** confidence as the mean log-probability across all decoded tokens, normalized to a 0-1 scale. Specifically:

$$\text{confidence} = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log P(t_i|t_1 \ldots t_{i-1}, \text{audio})\right) \tag{1}$$

We use utterance-level rather than token-level confidence because emergency triage requires a holistic assessment of transcription reliability. Token-level confidence would require additional aggregation logic and may miss systematic degradation patterns (e.g., consistently low confidence across an entire basilectal utterance).

Confidence Threshold: We set the "low confidence" threshold at 0.7 based on initial calibration experiments, though sensitivity analysis is needed to optimize this value (see Limitations).

3.2 Layer 2: Local NLP Entity Extraction

When ASR produces usable transcription (confidence ≥ 0.7), the NLP layer extracts structured emergency information using Llama 3 8B running locally via Ollama. The extraction schema targets four entity types critical for emergency dispatch:

- **LOCATION:** Street addresses, landmarks, geographic references
- **HAZARD:** Emergency type (fire, flood, medical, violence, etc.)
- **PERSONS:** Number of people involved, injuries mentioned
- **URGENCY:** Temporal markers ("right now," "hurry," breathing patterns)

Handling Garbled Input: A critical design question is how the NLP layer behaves when ASR produces low-quality transcriptions. We address this through confidence-aware prompting:

```
SYSTEM: You are extracting emergency information from a
speech transcript. The transcription confidence is
{confidence_score}.

If confidence is below 0.7, the transcript may contain errors.
Extract what you can, but:
1. Mark uncertain extractions with [UNCERTAIN]
2. Do not hallucinate or guess missing information
3. Prioritize extracting any recognizable location names
4. Note phonetically similar alternatives for garbled terms

TRANSCRIPT: {asr_output}
```

Example of garbled transcript handling:

ASR Output (confidence=0.52)	**NLP Extraction**
"mi house a bun down pan [unintelligible] road near di gas station"	LOCATION: "[UNCERTAIN] road, near gas station"; HAZARD: "fire (house burning)"; PERSONS: "unknown"; URGENCY: "high"

Table 1: Example of NLP extraction from low-confidence ASR output

When confidence is very low (<0.4), the NLP layer produces minimal structured output and flags the call for immediate human review, relying on the bio-acoustic layer to provide triage guidance.

3.3 Layer 3: Bio-Acoustic Distress Detection

The bio-acoustic layer operates on raw audio, independent of ASR success, extracting features correlated with psychological distress. Based on the vocal stress literature [21, 22], we focus on two primary features:

Feature Extraction (using librosa):

1. **Fundamental Frequency (F0):** Mean pitch extracted via autocorrelation method
    - Typical baseline: 85-180 Hz (male), 165-255 Hz (female)
    - Stress indicator: Elevation >20% above speaker baseline

2. **Energy (RMS amplitude):** Mean intensity across utterance
    - Normalized to 0-1 scale relative to recording gain
    - Stress indicator: energy_avg > 0.05 (normalized units)

Distress Score Calculation:
The distress score combines pitch and energy deviations into a single metric:

$$\text{distress\_score} = w_{\text{pitch}} \cdot \text{pitch\_component} + w_{\text{energy}} \cdot \text{energy\_component} \tag{2}$$

where:

$$\text{pitch\_component} = \min\left(1.0, \max\left(0, \frac{\bar{F}_0 - 180}{120}\right)\right) \tag{3}$$

$$\text{energy\_component} = \min\left(1.0, \frac{\bar{E}}{0.1}\right) \tag{4}$$

$$w_{\text{pitch}} = 0.6 \quad \text{(based on literature showing F0 as most reliable marker)}$$

$$w_{\text{energy}} = 0.4$$

Where $\bar{F}_0$ is mean fundamental frequency in Hz and $\bar{E}$ is mean RMS energy (normalized).
Threshold Rationale:

- "High Distress" threshold: distress_score > 0.7
- "Moderate Distress" threshold: $0.4 <$ distress_score ≤ 0.7
- "Low Distress" threshold: distress_score ≤ 0.4

These thresholds are calibrated against Van Puyvelde et al.'s [22] findings on F0 ranges in emergency versus baseline speech. The 180 Hz baseline in the pitch component represents an approximate population mean; see Limitations for discussion of gender normalization requirements.

3.4 The Complementarity Principle

The theoretical foundation for our multi-layer design rests on what we term the **Complementarity Principle**: the conditions that degrade ASR performance (high stress, code-switching to basilect, elevated emotion) are precisely the conditions that elevate bio-acoustic distress signals.

This creates a natural redundancy:

- **High ASR confidence + Low distress:** Standard call, NLP extraction reliable
- **High ASR confidence + High distress:** Urgent but comprehensible, prioritize
- **Low ASR confidence + Low distress:** Technical issue (noise, distance), re-prompt
- **Low ASR confidence + High distress:** Critical combination—caller in crisis, speech shifted to basilect, route to human dispatcher immediately

The fourth quadrant—low confidence combined with high distress—represents our key insight. Rather than treating this as a system failure, we interpret it as valuable triage information: this caller needs immediate human attention precisely *because* automated systems cannot process their speech.

ASR Confidence	Distress Score	Triage Decision
High (≥0.7)	Low (≤0.4)	**STANDARD:** Queue with extracted metadata
High (≥0.7)	Moderate (0.4-0.7)	**ELEVATED:** Priority queue, human review recommended
High (≥0.7)	High (>0.7)	**URGENT:** Immediate human dispatch
Low (<0.7)	Low (≤0.4)	**UNCLEAR:** Re-prompt caller, check audio quality
Low (<0.7)	Moderate (0.4-0.7)	**PRIORITY REVIEW:** Human dispatcher reviews audio
Low (<0.7)	High (>0.7)	**CRITICAL:** Immediate human dispatch, flag as potential dialect

Table 2: Triage decision matrix based on ASR confidence and bio-acoustic distress

3.5 Triage Decision Matrix

The final routing decision combines ASR confidence and distress score according to the following matrix:

4 Theoretical Foundations

4.1 Why Accent-Tuned ASR Is Necessary But Insufficient

Fine-tuning Whisper on Caribbean speech will improve transcription accuracy, but it cannot eliminate the accent gap entirely. Madden et al. [16] achieved 30% WER on Jamaican Patois with fine-tuning—a dramatic improvement from 89% baseline, but still far above the <5% WER typical for standard English. In emergency contexts, even 30% WER means nearly one-third of words may be incorrect, potentially including critical location or hazard information.

Moreover, fine-tuning on broadcast speech cannot fully capture emergency speech characteristics: elevated noise (sirens, screaming, wind), emotional vocal qualities, and the stress-induced basilectal reversion discussed above. A system relying solely on ASR, no matter how well-tuned, will fail precisely when it is needed most.

4.2 Why Bio-Acoustic Analysis Is Necessary But Insufficient

Conversely, bio-acoustic distress detection alone cannot provide the semantic information needed for emergency dispatch. A caller may exhibit extreme vocal stress while saying "my house is on fire" or "I lost my keys"—the distress signal is identical, but the appropriate response differs dramatically.

Furthermore, as Deschamps-Berger et al. [7] demonstrated, laboratory accuracy of emotion recognition systems (63%) drops substantially in real emergency calls (45.6%). Bio-acoustic features provide reliable *gradient* information about caller state but cannot substitute for semantic content.

4.3 The Integration Thesis

Our architecture integrates these complementary information sources based on the following thesis: **In emergency contexts, the correlation between ASR failure and genuine distress creates an opportunity to use recognition uncertainty as a routing signal rather than an error to be minimized.**

This thesis rests on the psycholinguistic literature establishing that:

1. Stress triggers cognitive load effects that impair executive function [9]
2. Impaired executive function leads to reduced inhibition of dominant language varieties [11]

3. For Caribbean speakers, dominant varieties include basilectal forms underrepresented in ASR training [18, 16]

4. Stress simultaneously elevates bio-acoustic markers (F0, intensity) that can be detected independently of speech content [22]

The logical conclusion: when ASR confidence drops and bio-acoustic distress rises, the system has detected a caller in genuine crisis whose speech has shifted beyond standard recognition capabilities. This combination should trigger immediate human review—not because the system has failed, but because it has successfully identified a caller who needs human attention most.

5 Deployment Considerations

5.1 Hardware Requirements

The complete system is designed for deployment on Raspberry Pi 5 (8GB RAM) or equivalent edge hardware:

Component	Model	Size	Inference Speed
ASR	Whisper Medium (INT4)	∼400MB	∼10s per 30s audio
NLP	Llama 3 8B (4-bit)	∼4GB	2-5 tokens/sec
Bio-acoustic	librosa + numpy	<50MB	Real-time

Table 3: Hardware requirements for edge deployment

Total system footprint: ∼4.5GB, well within Raspberry Pi 5 8GB capacity.

5.2 Latency Analysis

Important Clarification: TRIDENT is designed as a **batch triage engine for queue management during disaster surges**, not a real-time conversational assistant. The system processes completed call recordings (or call segments) to assign priority scores for dispatcher review.

End-to-end processing time for a 30-second call segment:

- Audio preprocessing: ∼2 seconds
- ASR transcription: ∼10 seconds
- Bio-acoustic extraction: ∼1 second (parallel with ASR)
- NLP entity extraction: ∼30-45 seconds
- Triage decision: <1 second
- **Total: ∼45-60 seconds**

This latency is unsuitable for real-time call answering (picking up the phone), but appropriate for:

- **Surge triage:** When call volume exceeds dispatcher capacity, the system prioritizes the queue
- **Post-call analysis:** Reviewing recorded calls for quality assurance or pattern detection
- **Voicemail triage:** Processing voicemail messages left during high-volume periods

For real-time operation, a production deployment would require GPU acceleration (e.g., NVIDIA Jetson) to reduce ASR latency to <3 seconds.

5.3 Offline Operation

All components operate without internet connectivity:

- Whisper model weights stored locally
- Llama 3 served via local Ollama instance
- Bio-acoustic analysis uses standard signal processing libraries
- Triage logic implemented in local Python

This enables deployment at emergency coordination centers that may lose internet connectivity during disasters while maintaining local power (generator/battery backup).

6 Limitations and Future Work

6.1 Current Limitations

Validation gap (most critical). This paper presents an architectural framework with theoretical grounding but limited empirical validation on real emergency calls. Performance claims for each layer are based on component evaluations and related literature rather than end-to-end system testing.

Training data constraints. Caribbean emergency speech corpora do not exist. Fine-tuning was performed on broadcast speech, which differs significantly from emergency call acoustics in noise profiles, emotional content, and register distribution.

Bio-acoustic threshold calibration. Distress detection thresholds are derived from literature on non-Caribbean populations. Baseline vocal characteristics may vary across Caribbean demographics, requiring population-specific calibration.

Gender and F0 baseline variation. Fundamental frequency is sexually dimorphic: relaxed male voices average 85-180 Hz while relaxed female voices average 165-255 Hz. Our current distress detection uses a single baseline (180 Hz), which creates systematic bias:

- **False positive risk:** A relaxed female speaker may naturally produce F0 values that trigger "elevated distress" classification
- **False negative risk:** A highly stressed male speaker may not exceed the 240 Hz "high distress" threshold

Addressing this requires either: (a) speaker gender classification followed by gender-normalized thresholds, (b) "relative to baseline" calculation requiring speaker enrollment, or (c) detection of F0 *change* within a call rather than absolute values. Each approach has tradeoffs that require empirical evaluation. **This represents a significant risk of gender bias in triage logic that must be addressed before deployment.**

Single-speaker assumption. The current architecture assumes single-speaker input. Multi-party calls, common in emergencies ("put your mother on the phone"), are not handled.

Confidence threshold sensitivity. The 0.7 ASR confidence threshold was selected based on initial calibration but has not been rigorously optimized. Sensitivity analysis examining system performance across threshold values (0.5-0.9) is needed to understand the precision-recall tradeoff for triage decisions.

6.2 Future Work

Caribbean Emergency Speech Corpus. The most critical enabler for future progress is a dedicated corpus combining Caribbean-accented speech with emergency domain content and stress annotations. We are exploring partnerships with Caribbean emergency services to develop such a resource.

Empirical validation. End-to-end evaluation with emergency dispatch professionals rating system triage decisions against expert judgment.

Ablation studies. Rigorous testing to demonstrate that the bio-acoustic layer actually improves triage accuracy over ASR-only approaches—proving the value of the "low confidence as signal" insight.

Dialect density estimation. Augmenting the triage logic with automatic estimation of creole feature density, providing dispatchers with guidance on expected communication challenges.

Multilingual extension. Caribbean emergency services handle calls in English, Spanish, French, Dutch, and various creoles. Extending the architecture to multilingual operation would significantly expand impact.

Gender-normalized distress detection. Implementing and validating speaker-dependent F0 baseline estimation to address the gender bias limitation described above.

7 Conclusion

TRIDENT presents a defensive architecture for Caribbean emergency speech processing that treats ASR limitations not as failures to be eliminated but as signals to be incorporated into triage logic. By combining accent-adapted speech recognition, local NLP extraction, and bio-acoustic distress detection, the system maintains functionality across a range of conditions—including the high-stress, dialect-shifted speech most likely to defeat traditional ASR approaches.

The key insight is that low ASR confidence combined with high vocal distress is not a system failure but a system feature: the signature of a caller in genuine crisis whose speech patterns have shifted beyond the reach of standard recognition. Routing these calls to priority human review ensures that the most vulnerable callers receive the most urgent attention.

We hope this architectural framework contributes to more equitable emergency AI systems—not just for Caribbean populations, but for the billions of speakers worldwide whose accents and dialects remain underserved by current speech technology.

References

[1] Mohamed Aboualola, Khalid Abualsaud, Tamer M. S. Khattab, Nizar Zorba, and Hossam S. Hassanein. Edge technologies for disaster management: A survey of social media and artificial intelligence integration. *IEEE Access*, 11:73782–73802, 2023.

[2] Afraa Attiah and Manal Kalkatawi. AI-powered smart emergency services support for 9-1-1 call handlers using textual features and svm model for digital health optimization. *Frontiers in Big Data*, 8:1594062, 2025.

[3] Stig Nikolaj Blomberg et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation*, 138:322–329, 2019.

[4] Stig Nikolaj Blomberg et al. Effect of machine learning on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: A randomized clinical trial. *JAMA Network Open*, 4(1):e2032320, 2021.

[5] Marcel Lucas Chee, Mark Leonard Chee, Haotian Huang, Katelyn Mazzochi, Kieran Taylor, Han Wang, Mengling Feng, Andrew Fu Wah Ho, Fahad Javaid Siddiqui, Marcus Eng Hock Ong, Nan Liu, et al. Artificial intelligence and machine learning in prehospital emergency care: A scoping review. *iScience*, 26(8):107407, 2023.

[6] Grażyna Demenko and Magdalena Jastrzębska. Analysis of voice stress in call center conversations. In *Proceedings of Speech Prosody 2012*, pages 183–186, 2012.

[7] Théo Deschamps-Berger, Lori Lamel, and Laurence Devillers. End-to-end speech emotion recognition: Challenges of real-life emergency call centers data recordings. In *Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2021.

[8] Théo Deschamps-Berger, Lori Lamel, and Laurence Devillers. Exploring attention mechanisms for multimodal emotion recognition in an emergency call center corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*, pages 1–5, 2023.

[9] Tamar H. Gollan and Victor S. Ferreira. Should I stay or should I switch? A cost-benefit analysis of voluntary language switching in young and aging bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):640–665, 2009.

[10] Santosh Gondi and Vineel Pratap. Performance evaluation of offline speech recognition on edge devices. *Electronics*, 10(21):2697, 2021.

[11] David W. Green. Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1(2):67–81, 1998.

[12] John H. L. Hansen and Sanjay Patil. Speech under stress: Analysis, modeling and recognition. *Speaker Classification I*, pages 108–137, 2007. Chapter in Springer Lecture Notes in Computer Science.

[13] Allison Koenecke et al. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[14] Judith F. Kroll, Susan C. Bobb, and Zofia Wodniecka. Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition*, 9(2):119–135, 2006.

[15] Iulia Lefter, Leon J. M. Rothkrantz, David A. van Leeuwen, and Pascal Wiggers. Automatic stress detection in emergency (telephone) calls. *International Journal of Intelligent Defence Support Systems*, 4(2):148–168, 2011.

[16] Jordan Madden, Matthew Stone, Dimitri Johnson, and Daniel Geddez. Towards robust speech recognition for jamaican patois music transcription. *arXiv preprint arXiv:2507.16834*, 2025.

[17] Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. AfriSpeech-200: Pan-african accented english speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685, 2023.

[18] Peter L. Patrick. *Urban Jamaican Creole: Variation in the Mesolect*. John Benjamins Publishing, Amsterdam, 1999.

[19] Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. The edinburgh international accents of english corpus: Towards the democratization of english asr. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*, pages 1–5, 2023.

[20] Carlos Santos-Burgoa, John Sandberg, Erick Suárez, Ann Goldman-Hawes, Scott Zeger, Alejandra Garcia-Meza, Cynthia M. Pérez, Kenneth Rivera, Adriana Colón Ramos, Jose Figueroa, et al. Differential and persistent risk of excess mortality from hurricane maria in puerto rico: A time-series analysis. *The Lancet Planetary Health*, 2(11):e478–e488, 2018.

[21] Lilien Schewski, Mathew Magimai Doss, Guido Beldi, and Sandra Keller. Measuring negative emotions and stress through acoustic correlates in speech: A systematic review. *PLOS ONE*, 20(7):e0328833, 2025.

[22] Martine Van Puyvelde, Xavier Neyt, Francis McGlone, and Nathalie Pattyn. Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, 9:1994, 2018.

A Implementation Details

Repository: https://github.com/smg-labs/project-filter *(to be made public upon acceptance)*

Dependencies:

- Python 3.11+
- openai-whisper
- transformers, peft (LoRA fine-tuning)
- ollama (Llama 3 serving)
- librosa (audio feature extraction)
- jiwer (WER evaluation)

Hardware requirements:

- Training: NVIDIA GPU with 16GB+ VRAM recommended
- Inference: CPU-only operation supported; 8GB RAM minimum

B Acknowledgments

This work was developed during the Caribbean Voices AI Hackathon organized by the UWI AI Innovation Centre. We thank the organizers for creating the competition and providing the BBC Caribbean speech corpus that motivated this research.