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

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

SMG Labs Research Group

December 2, 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 transcription confidence, semantic content severity, 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.** A complementary insight drives the content severity layer: trained responders and composed bystanders may report life-threatening emergencies without elevated vocal stress, requiring semantic analysis to capture urgency that paralinguistic features miss. 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 [14]. 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 contribution is a three-dimensional triage framework that integrates ASR confidence, bio-acoustic distress, and semantic content severity. Two complementary insights motivate this design:

1. **Uncertainty as signal:** Low ASR confidence, rather than representing system failure, serves as a valuable triage indicator—particularly when combined with elevated vocal distress markers indicating a caller in crisis whose speech may have shifted toward basilectal registers. This reframes accent-induced transcription errors from bugs into features that correlate with genuine caller distress.

2. **Content beyond voice:** Trained first responders, medical professionals, and composed bystanders may report life-threatening emergencies without elevated vocal stress. Semantic analysis of transcript content captures urgency that paralinguistic features alone would miss—ensuring that "children trapped in burning building," spoken calmly, receives appropriate priority.

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 [12] 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. [14] 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 [21]. 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 [18]. 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. [17] 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 outperformed non-fine-tuned Whisper Large, demonstrating that domain-specific data matters more than model size for underrepresented varieties. Their scaling law ($\text{WER} = 158.06 \times \text{M}^{-0.255} \times \text{D}^{-0.269}$) 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 [6]. 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 [23]. 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 [13] 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. [26] 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. [16] 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 [7] 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. [8], 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 [9], 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 [12]. 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 [10] 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. [15] 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 [19] 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 [22]. 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 [11] 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.

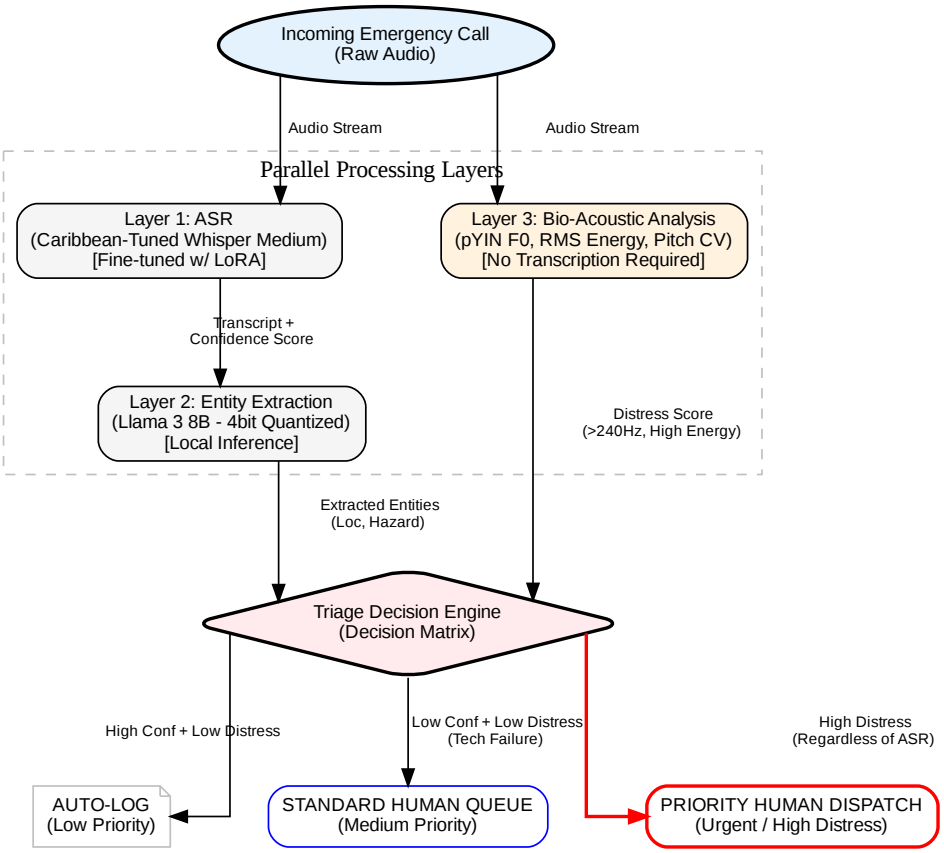

Figure 1: The TRIDENT architecture. The system processes raw audio through two parallel streams: (Left) A Caribbean-adapted ASR and NLP pipeline for semantic extraction and content severity scoring, and (Right) a bio-acoustic analysis layer for determining physiological distress. The Triage Decision Engine integrates three independent signals—ASR confidence, content severity, and vocal distress—ensuring that (1) calls with low transcription confidence but high vocal distress are routed to priority dispatch, and (2) semantically urgent calls from calm reporters are not under-prioritized due to absent vocal stress markers.

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 [17] 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)
- 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:

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

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.

3.2.1 Content Severity Scoring

Beyond entity extraction, the NLP layer computes a **Content Severity Score** ($S_c \in [0, 100]$) that quantifies the urgency implied by the *semantic content* of the transcript, independent of how the caller sounds. This addresses a critical gap: a trained first responder or composed bystander may report a mass casualty event in a calm voice, producing low bio-acoustic distress despite extremely urgent content.

Classification-Based Approach. Rather than brittle keyword matching, we leverage the LLM's semantic understanding to classify transcript content along four dimensions. This approach offers critical advantages for Caribbean speech: the model can recognize that "mi granmodda drop dung an she nah move" conveys the same urgency as "my grandmother collapsed and she's not moving" without requiring an exhaustive enumeration of creole variants. The LLM also handles negation ("no one is trapped"), indirect references ("she nine months pregnant" → vulnerable), and context-dependent interpretation that keyword matching cannot capture.

The LLM outputs structured classifications according to the following schema:

```
{
  "hazard_category": "violent_crime" | "medical" | "fire" |
                     "flood" | "traffic" | "infrastructure" | "other",
  "life_threat_level": "imminent" | "potential" | "none",
  "vulnerable_population": true | false,
  "situation_status": "escalating" | "stable" | "resolved",
  "persons_affected": <integer>
}
```

A deterministic scoring function then maps these classifications to the Content Severity Score, separating the flexibility of neural language understanding from the interpretability of rule-based scoring:

$$S_c = \min\left(100,\ S_{\text{hazard}} + S_{\text{threat}} + S_{\text{vuln}} + S_{\text{scale}}\right) \tag{2}$$

Component 1: Hazard Category (S_{hazard}). Different emergency types carry inherent urgency levels:

Hazard Category	Score
`violent_crime`	30
`medical`	25
`fire`	25
`flood`	20
`traffic`	15
`infrastructure`	10
`other`	5

Table 2: Hazard category severity scores

Component 2: Life-Threat Level (S_{threat}). The LLM classifies the immediacy of danger to life based on semantic understanding of the full transcript context:

- `imminent`: Active, immediate threat to life (trapped, not breathing, active violence, drowning) → +30
- `potential`: Situation could become life-threatening (injuries, spreading fire, chest pain) → +15
- `none`: No apparent threat to life → +0

Component 3: Vulnerable Population (S_{vuln}). Boolean classification indicating presence of children, elderly, pregnant individuals, or persons with disabilities. If `true` $\rightarrow +15$. This reflects both ethical prioritization and reduced self-rescue capacity.

Component 4: Scale and Escalation (S_{scale}). Combines two factors:

- `persons_affected`: +5 per person, capped at +20
- `situation_status = "escalating"`: +10 (fire spreading, water rising, more vehicles involved)

Example severity calculations:

Transcript	LLM Classification	S_c
"Pothole on Nelson Street"	`infrastructure, none, false, stable, 0`	10
"Car accident, one person injured"	`traffic, potential, false, stable, 1`	35
"House fire, spreading to neighbor's yard"	`fire, potential, false, escalating, 0`	50
"Mi granmodda drop dung, she nah breathe"	`medical, imminent, true, stable, 1`	75
"Pickney dem trap inna di fire"	`fire, imminent, true, stable, 2+`	80

Table 3: Content severity scoring via LLM classification. The model's semantic understanding captures urgency from both standard English and Caribbean creole variants without explicit keyword enumeration.

The Content Severity Score provides the third dimension of the triage decision matrix (Section 3.5), ensuring that semantically urgent calls receive priority routing even when vocal distress markers are low.

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 [23, 26, 27], we focus on features that capture physiological arousal through vocal production changes.

3.3.1 Feature Extraction

Using librosa, we extract the following acoustic features:

1. **Fundamental Frequency (F0):** Mean pitch extracted via autocorrelation method
   - Typical baseline: 85–180 Hz (male), 165–255 Hz (female) [24]
   - Stress indicator: Elevation above speaker baseline
2. **F0 Coefficient of Variation (CV):** Pitch instability measure
   - Computed as $CV = \sigma_{F0}/\mu_{F0}$
   - Normalizes for baseline differences across speakers
   - Stress indicator: $CV > 0.3$ suggests vocal instability
3. **Energy (RMS amplitude):** Mean intensity across utterance
   - Normalized to 0–1 scale relative to recording gain

- Stress indicator: Elevated intensity during distress vocalizations

4. **Jitter:** Cycle-to-cycle variation in F0 period
   - Relatively independent of prosodic patterns [26]
   - Pathology threshold: >1.04% [5]

3.3.2 Distress Score Calculation

The distress score combines multiple acoustic indicators into a composite metric. We weight features according to their documented reliability and sex-independence:

$$D = w_{\text{pitch}} \cdot P + w_{\text{var}} \cdot V + w_{\text{energy}} \cdot E + w_{\text{jitter}} \cdot J \tag{3}$$

where:

$$P = \min\left(1.0, \max\left(0, \frac{\bar{F}_0 - 180}{120}\right)\right) \quad \text{(pitch elevation)} \tag{4}$$

$$V = \min\left(1.0, \frac{CV_{F0}}{0.5}\right) \quad \text{(pitch instability)} \tag{5}$$

$$E = \min\left(1.0, \frac{\bar{E}}{0.1}\right) \quad \text{(energy)} \tag{6}$$

$$J = \min\left(1.0, \frac{\text{jitter}}{0.02}\right) \quad \text{(perturbation)} \tag{7}$$

The weights reflect relative reliability from the literature:

- $w_{\text{pitch}} = 0.30$ — F0 elevation is the most consistent stress marker but is sex-dependent
- $w_{\text{var}} = 0.35$ — F0 coefficient of variation is sex-normalized and robust
- $w_{\text{energy}} = 0.20$ — intensity elevation accompanies distress
- $w_{\text{jitter}} = 0.15$ — perturbation measures are prosody-independent

3.3.3 Threshold Classification

- **High Distress:** $D > 0.5$
- **Low Distress:** $D \leq 0.5$

These thresholds are calibrated against Van Puyvelde et al.'s [26] findings on vocal markers in emergency versus baseline speech.

3.3.4 Sex Differences in Baseline F0

A methodological consideration is the substantial difference in baseline fundamental frequency between male speakers (typically 85–175 Hz) and female speakers (165–270 Hz) [24, 25]. Setting a single absolute F0 threshold for distress detection risks differential sensitivity: a female speaker might cross a high absolute threshold with moderate stress, while a male speaker might require extreme distress to reach the same value.

We address this limitation through two strategies. First, we prioritize **normalized and relative measures**: the coefficient of variation ($CV = \sigma_{F0}/\mu_{F0}$) captures pitch instability

independent of baseline, and jitter measures cycle-to-cycle perturbations that research shows are "relatively independent from prosodic patterns" [26]. Second, we reduce the weight assigned to absolute F0 elevation relative to normalized measures.

Importantly, research on psychosocial stress indicates that the *direction* and *pattern* of vocal changes under stress show "striking parallels in men and women" [20]—both sexes exhibit increased pitch mean, minimum, and variation during acute stress. The challenge is not that stress manifests differently, but that baseline values differ.

Limitation. The current threshold values are derived from literature on non-Caribbean, predominantly Western populations. Automatic sex identification from voice is itself an imperfect classifier, particularly for voices near the overlap region of male and female F0 distributions. Rather than introduce a potentially error-prone sex classification step, we employ sex-normalized features (CV, jitter) weighted more heavily than absolute F0. A validation study with sex-stratified analysis on Caribbean emergency calls is essential to: (1) calibrate population-appropriate thresholds, (2) confirm that normalized measures maintain sensitivity across speaker demographics, and (3) determine whether Caribbean populations exhibit different baseline distributions requiring adjustment.

3.4 The Complementarity Principle

The theoretical foundation for our multi-layer design rests on what we term the **Complementarity Principle**: the three triage dimensions capture distinct failure modes and urgency signals that compensate for each other's blind spots.

Dimension 1: ASR Confidence. The conditions that degrade ASR performance (high stress, code-switching to basilect, environmental noise) are precisely the conditions that often accompany genuine emergencies. Low confidence is not merely a technical limitation—it correlates with caller distress.

Dimension 2: Bio-Acoustic Distress. Vocal stress markers (elevated pitch, intensity, instability) provide a parallel assessment channel that operates on raw audio, independent of transcription success. A caller whose speech is entirely unintelligible to ASR will still produce detectable distress signals.

Dimension 3: Content Severity. Semantic analysis of transcript content captures urgency that vocal characteristics may miss. Trained professionals, repeat callers, and composed bystanders often report critical emergencies without elevated vocal stress—their calm delivery masks the urgency that only content analysis reveals.

This creates a robust triage space with complementary coverage:

- **High Confidence + Low Distress + Low Severity:** Routine call, automated processing appropriate
- **High Confidence + Low Distress + High Severity:** The composed reporter—urgent content from a calm caller requires priority handling despite absent vocal stress markers
- **High Confidence + High Distress + Low Severity:** Anxious caller, minor issue—human review to de-escalate
- **High Confidence + High Distress + High Severity:** Confirmed emergency, all signals aligned
- **Low Confidence + Low Distress + Low Severity:** Likely technical issue, re-prompt or review
- **Low Confidence + Low Distress + High Severity:** Garbled but fragments suggest urgency—human ears needed

- **Low Confidence + High Distress + Low Severity:** Distressed caller, unintelligible speech—priority human review
- **Low Confidence + High Distress + High Severity:** Maximum urgency—all indicators elevated, immediate dispatch

Two cells represent our key insights. The **Low Confidence + High Distress** cases (regardless of content) capture callers in crisis whose speech has shifted toward basilectal registers—we interpret ASR failure combined with vocal stress as valuable triage information rather than system failure. The **High Confidence + Low Distress + High Severity** cell captures the complementary blind spot: the trained first responder or composed bystander whose calm voice belies the urgency of their report. Together, these insights ensure that neither paralinguistic nor semantic signals alone determine routing.

3.5 Triage Decision Matrix

The final routing decision integrates three independent signals: ASR confidence (transcription reliability), bio-acoustic distress (caller physiological state), and content severity (semantic urgency of the message). This three-dimensional approach addresses a critical limitation of two-dimensional triage: a calm, composed caller reporting a mass casualty event would be under-prioritized if routing relied solely on vocal distress markers.

3.5.1 Three-Dimensional Triage Space

Each call is mapped to a point in triage space defined by:

- **ASR Confidence** (C): High (≥ 0.7) or Low (< 0.7)
- **Bio-Acoustic Distress** (D): High (> 0.5) or Low (≤ 0.5)
- **Content Severity** (S_c): High (≥ 50) or Low (< 50)

The $2 \times 2 \times 2$ combination yields eight triage cells, shown in Table 4.

ASR Conf.	Distress	Content	Triage	Rationale
High	Low	Low	**STANDARD**	Routine call, clear transcription
High	Low	High	**ELEVATED**	Calm reporter, urgent content*
High	High	Low	**ELEVATED**	Distressed caller, minor content
High	High	High	**URGENT**	Clear emergency, confirmed urgent
Low	Low	Low	**UNCLEAR**	Possible technical issue
Low	Low	High	**PRIORITY**	Fragments suggest urgency
Low	High	Low	**PRIORITY**	Distress despite poor transcription†
Low	High	High	**CRITICAL**	Maximum urgency, all indicators elevated

Table 4: Three-dimensional triage decision matrix. *Addresses trained responder/composed bystander scenario. †Preserves original insight regarding stress-induced dialect shift.

3.5.2 Triage Categories

CRITICAL: Immediate human dispatch. Highest priority queue position.

URGENT: Immediate human dispatch. High priority.

PRIORITY: Human dispatcher reviews audio promptly. Flagged for potential dialect reversion or content ambiguity.

ELEVATED: Priority queue placement. Human review recommended.

STANDARD: Normal queue with extracted metadata available to dispatcher.

UNCLEAR: System prompts caller for clarification or flags for audio quality review.

3.5.3 Preserving the Core Insight

The original two-dimensional insight—that low ASR confidence combined with high vocal distress signals a caller in crisis whose speech has shifted toward basilectal registers—remains encoded in the matrix. The Low/High/Low and Low/High/High cells both route to priority or critical handling. The addition of Content Severity provides a parallel pathway for urgent calls that would otherwise be missed: the High/Low/High cell captures the trained professional or composed bystander reporting a genuine emergency without elevated vocal stress markers.

3.5.4 Dispatcher Interface Examples

Figure 2 and Figure 3 illustrate the dispatcher interface for contrasting triage scenarios. The interface presents real-time triage indicators including the three-dimensional signal values, extracted location metadata, and confidence scores to support rapid human decision-making.

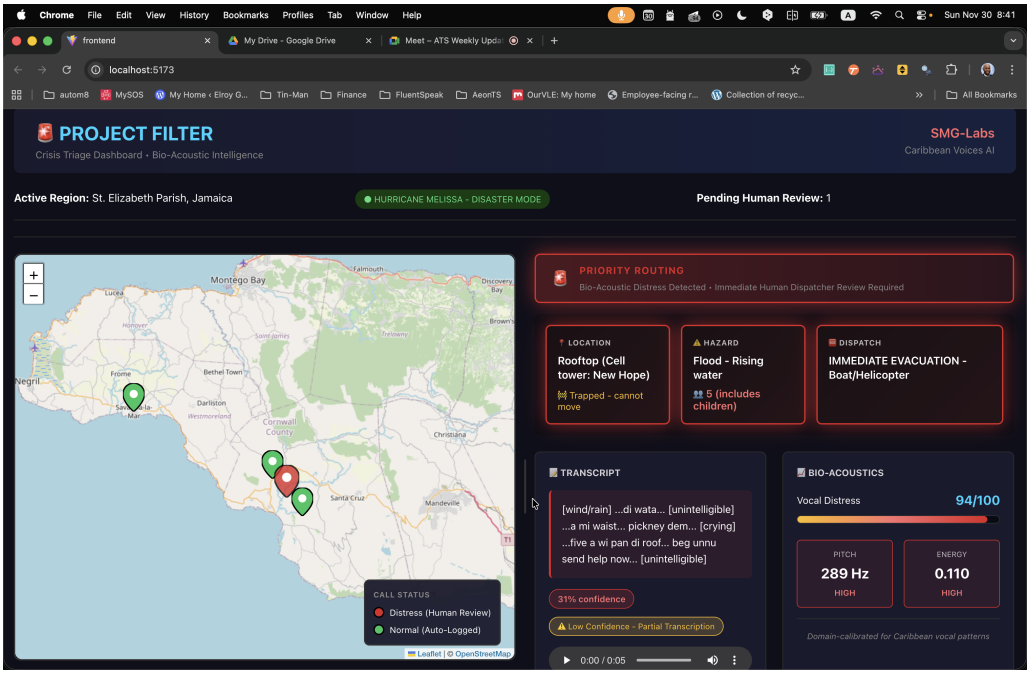

Figure 2: Dispatcher interface for a low-risk scenario (STANDARD triage). The system displays high ASR confidence, low distress, and low content severity, with successfully extracted location metadata. The call queues normally with automated metadata available to the dispatcher.

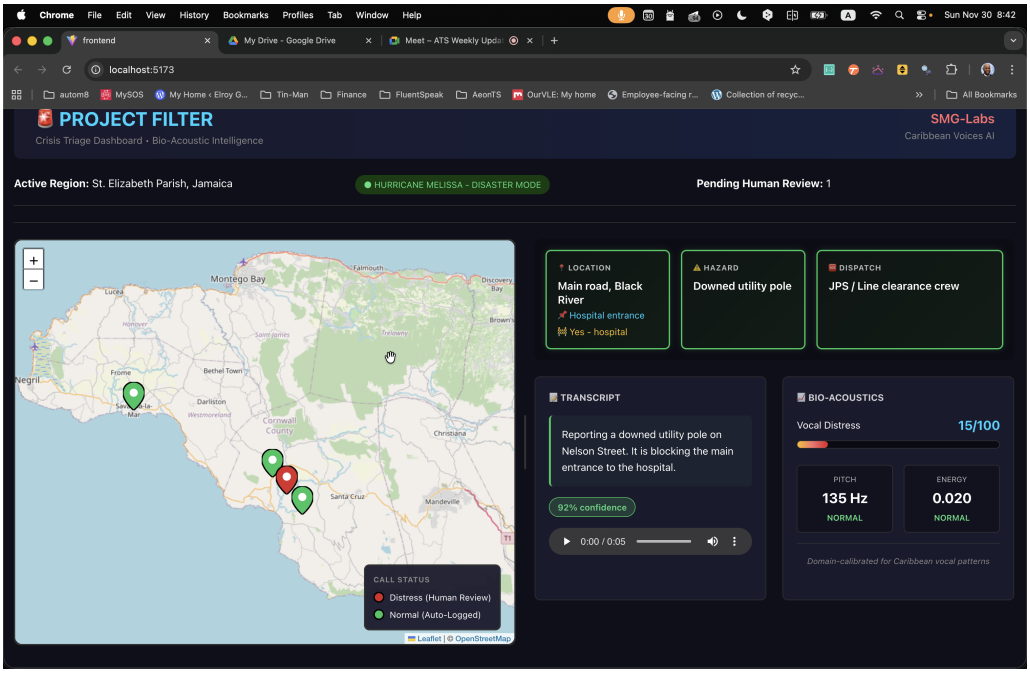

Figure 3: Dispatcher interface for a high-risk scenario (CRITICAL or URGENT triage). Elevated distress markers, reduced ASR confidence, and high content severity trigger immediate priority routing, with visual indicators alerting dispatchers to potential dialect shift or acute crisis requiring immediate human attention.

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. [17] 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. [8] 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 [10]
2. Impaired executive function leads to reduced inhibition of dominant language varieties [12]
3. For Caribbean speakers, dominant varieties include basilectal forms underrepresented in ASR training [19, 17]
4. Stress simultaneously elevates bio-acoustic markers (F0, intensity) that can be detected independently of speech content [26]

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 5: 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. The three-dimensional triage matrix (ASR confidence × distress × content severity) is theoretically motivated but has not been validated against expert dispatcher judgments.

Training data constraints. Caribbean emergency speech corpora do not exist. ASR fine-tuning was performed on broadcast speech, which differs significantly from emergency call acoustics in noise profiles, emotional content, and register distribution. The gap between training domain (broadcast) and deployment domain (emergency calls) may introduce systematic errors not captured in current evaluation.

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

Sex differences in F0 baseline. Fundamental frequency is sexually dimorphic: male voices typically range 85–175 Hz while female voices range 165–270 Hz [24]. We partially address this through architectural choices:

- Prioritizing sex-normalized features: F0 coefficient of variation ($CV = \sigma_{F0}/\mu_{F0}$) and jitter, which measure relative instability rather than absolute values
- Weighting normalized features (CV: 0.35, jitter: 0.15) more heavily than absolute F0 elevation (0.30)

However, residual bias likely remains. Research confirms that stress manifests with "striking parallels in men and women" [20]—both sexes show increased pitch mean and variation under stress—but our reliance on any absolute F0 component creates risk:

- **False positive risk:** A relaxed female speaker near the upper baseline range may contribute to elevated distress scores
- **False negative risk:** A stressed male speaker with naturally low F0 may not contribute sufficiently to the pitch component

A validation study with sex-stratified analysis is essential to quantify this bias and determine whether further threshold adjustment or feature re-weighting is required.

Content severity classification. The Content Severity Score depends on LLM classification quality. While leveraging Llama 3's semantic understanding avoids brittle keyword matching, it introduces new failure modes:

- Classification errors propagate deterministically to severity scores
- Caribbean creole expressions not well-represented in LLM training data may be misclassified
- The model may fail to recognize culturally-specific threat indicators or landmarks

Empirical evaluation of classification accuracy on Caribbean emergency transcripts is needed, with particular attention to false negatives (urgent content classified as non-urgent).

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. Speaker changes mid-call could confuse bio-acoustic analysis and entity extraction continuity.

Threshold sensitivity. Multiple thresholds govern system behavior: ASR confidence (0.7), distress score (0.5), and content severity (50). These values were selected based on literature and initial calibration but have not been rigorously optimized. Sensitivity analysis examining system performance across threshold combinations is needed to understand precision-recall tradeoffs for each triage category.

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, with appropriate privacy protections and community consent.

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

- Comparison of three-dimensional triage against two-dimensional (confidence $\times$ distress) baseline
- Sex-stratified analysis of bio-acoustic distress detection accuracy
- Assessment of Content Severity classification on Caribbean creole transcripts

Ablation studies. Rigorous testing to quantify the contribution of each architectural component:

- Does bio-acoustic analysis improve triage over ASR-only approaches?
- Does Content Severity catch urgent calls missed by distress detection alone?

- What is the marginal value of Caribbean-tuned ASR versus off-the-shelf Whisper?

Sex-adaptive distress detection. Implementing and validating approaches to further reduce sex bias:

- Within-call F0 *change* detection rather than absolute thresholds
- Automatic speaker characteristic estimation for threshold adaptation
- Ensemble approaches combining multiple normalization strategies

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

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, though each language introduces its own ASR adaptation and content classification challenges.

Edge deployment optimization. While the architecture is designed for offline operation, current latency profiles (45–60 seconds per call) limit real-time applicability. Optimization for edge hardware (Raspberry Pi, embedded GPU) would enable deployment at emergency coordination centers with degraded connectivity.

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 bioacoustic 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] Paul Boersma and David Weenink. *Praat: doing phonetics by computer*, 2013. Version 5.3.51.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

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

[20] Katarzyna Pisanski, Joanna Nowak, and Piotr Sorokowski. Multimodal stress detection: Testing for covariation in vocal, hormonal and physiological responses to Trier Social Stress Test. *Hormones and Behavior*, 106:52–61, 2018.

[21] 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.

[22] 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.

[23] 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.

[24] Ingo R. Titze. Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, 85(4):1699–1707, 1989.

[25] Hartmut Traunmüller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults. *Journal of the Acoustical Society of America*, 97(4):2634–2639, 1995.

[26] 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.

[27] André Veiga et al. The fundamental frequency of voice as a potential stress biomarker: A systematic review and meta-analysis. *Stress and Health*, 2025.

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.