Study on speech detection method under heavy noise conditions using spectro-temporal modulation analysis

2310439 TRAN, Quynh Nhu

Detecting and locating survivors in disaster zones remains a critical and challenging task, particularly in Japan, where earthquakes, landslides, and floods occur with notable frequency. Deploying drones equipped with microphone arrays offers a promising means of rapidly scanning hazardous areas. By capturing human voices within disaster sites, such systems can significantly shorten search times and improve survival chances. However, disaster environments present extremely adverse acoustic conditions—often dominated by intense noise from wind, rain, machinery, or collapsing structures—that severely challenge conventional speech detection methods. In such scenarios, the signal-to-noise ratio (SNR) can fall well below -10 dB, sometimes reaching -20 dB, drastically limiting the effectiveness of traditional speech processing techniques.

Existing methods, including direction of arrival (DOA) estimation and voice activity detection (VAD), perform reasonably well under moderate noise but struggle in extremely low SNR conditions. Their detection accuracy drops sharply once SNR falls below -10 dB, rendering them ineffective in the most challenging disaster scenarios. This gap highlights the urgent need for a speech detection approach capable of sustaining robust performance under severe noise interference.

This study aims to develop a speech detection method tailored specifically to the heavy noise environments typical of disaster response operations. The proposed framework integrates spectro-temporal modulation (STM) analysis with deep neural networks to detect speech across SNRs from 20 dB down to -20 dB, with particular emphasis on the difficult range below -10 dB. STM features characterize both spectral and temporal modulation patterns, which are distinct for speech and noise. These patterns tend to remain relatively stable even under intense noise, making STM a promising choice for robust feature representation. Leveraging these properties, the proposed method seeks to achieve reliable speech detection even in the harshest conditions.

The novelty of this work lies in its dedicated focus on extremely low SNR environments and the systematic combination of STM analysis with attention-augmented deep learning architectures. Unlike conventional spectrogram-based approaches, STM represents audio in a joint modulation frequency domain, offering a unified multi-resolution perspective of the signal's temporal and spectral dynamics. Speech generally exhibits coherent spectro-temporal patterns, whereas noise—especially stationary noise—produces more random

or diffuse modulation structures. Exploiting these differences enables the system to distinguish between speech and non-speech even when noise energy dominates. Furthermore, incorporating the Convolutional Block Attention Module (CBAM) into the ResNet18 architecture enhances the model's ability to concentrate on modulation regions most indicative of speech presence. This attention mechanism operates across both spatial and channel dimensions, directing computational focus toward salient spectro-temporal areas while suppressing noise-dominated regions.

The methodology consists of two principal stages: feature extraction and classification. In the feature extraction stage, clean speech and noise are mixed at set SNR levels to create noisy speech samples. Spectrograms are generated using various filterbank configurations—Short-Time Fourier Transform (STFT), Constant-band, Mel, and Gammatone filterbanks—to assess their effect on STM quality and detection accuracy. These spectrograms are then converted into STM representations via a two-dimensional Fast Fourier Transform, capturing modulation information across both time and frequency. Multiple normalization and compression techniques are applied to improve robustness.

In the classification stage, several deep learning models are employed. A baseline convolutional neural network (CNN) serves as an initial benchmark, followed by ResNet18, which offers greater depth and residual connections for enhanced feature learning. To further improve low-SNR detection, the CBAM attention module is integrated into ResNet18. This augmented model learns to prioritize speech-relevant modulation patterns while suppressing noise-heavy features. The system is trained and evaluated on the JVS corpus for clean speech, with white noise used as a controlled stationary noise source to simulate disaster-site interference. Performance is evaluated over a broad SNR range using accuracy, equal error rate (EER), and other detection metrics.

Experimental findings confirm the effectiveness of the proposed approach. The ResNet18-CBAM model using STM features from the Gammatone filterbank delivered the highest overall performance. It achieved an accuracy of 92.79% across all SNR conditions, with near-perfect detection at higher SNRs. A comparative analysis against rVAD-fast and WebRTC VAD demonstrated the proposed method's superior performance, which achieved an EER below 10% at -10 dB SNR—a condition where conventional methods typically fail. Even at -15 dB, an extremely challenging condition, the EER remained around 20%, significantly outperforming the approximately 50% EER of the other methods. Although performance declined at -20 dB due to the overwhelming noise, the system maintained measurable detection capability, underscoring its robustness relative to existing techniques.

A comparative study of filterbanks revealed that the Gammatone filterbank consistently outperformed STFT, Mel, and Constant-band alternatives. This advantage stems from its auditory-inspired frequency resolution, which emphasizes low-frequency components critical to speech perception. Such emphasis is especially valuable in noisy conditions, where low-frequency speech cues are more resistant to masking. The results demonstrate that the choice of filterbank plays a substantial role in STM quality and, in turn, detection accuracy.

This research contributes to the field of robust speech detection in several ways. First, it offers a comprehensive evaluation of STM features across multiple filterbank designs, providing practical insight into optimal feature extraction under noise. Second, it validates the integration of attention mechanisms into deep residual networks as a means of enhancing detection in extreme noise conditions. Third, it proposes a scalable framework that operates effectively across a wide SNR range, making it adaptable to diverse disaster-site environments.

In conclusion, this work addresses the urgent demand for reliable speech detection in disaster scenarios characterized by exceptionally high noise levels. By merging spectro-temporal modulation analysis with attention-enhanced deep learning, the proposed system achieves high detection accuracy even at SNRs far below the threshold where conventional methods fail. While it performs particularly well with Gammatone-based STM features, challenges persist at extremely low levels such as -20 dB. Future research may incorporate non-stationary noise types, realistic disaster soundscapes, and multimodal fusion with visual or thermal inputs to further improve reliability. The outcomes of this study hold direct relevance for the development of autonomous drone-based survivor detection systems, potentially accelerating rescue operations and saving lives in disaster-stricken regions.