A Model for Keyword Spotting in a Voice Signal to Counter Social Engineering and Disinformation

Andrii Didus[1] and Ihor Tereikovskyi[1]

[1]*National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 03056, Kyiv*

Abstract

The voice communication channel is a significant vector for social engineering attacks and the spread of disinformation. Existing countermeasures that rely on cloud services have substantial drawbacks, including high latency, dependence on network connectivity, and privacy risks, making them unsuitable for real-time applications. This paper proposes a resource-efficient modular keyword spotting model designed for autonomous operation on resource-constrained edge devices. The model's architecture is based on the transformation of sequences of Mel-frequency cepstral coefficients into compact string "fingerprints" using differentiated weighting of informative features, followed by classification using the Levenshtein distance. Experimental validation on a Ukrainian-language command corpus demonstrated high performance: the F1-score reached 0.92 in ideal conditions and 0.78 at a signal-to-noise ratio of 5 dB. The proposed model significantly surpasses baseline and classical counterparts in the balance of accuracy, speed, and resource efficiency, which confirms its suitability for creating autonomous systems for proactive detection of auditory threats.

Keywords

keyword spotting, voice processing, social engineering, disinformation, resource-constrained systems

1. Introduction

The modern information security landscape is characterized by the expansion of attack vectors beyond traditional digital channels. The voice communication channel has evolved into a full-fledged space for attacks, actively used for social engineering (particularly voice phishing), spreading disinformation, and unauthorized system access. The dominant paradigm for countering such threats is based on the centralized analysis of audio data in cloud environments. This approach has several fundamental limitations. Firstly, the high latency associated with data transmission makes immediate real-time threat response impossible. Secondly, the reliance on a stable network connection makes such systems vulnerable in tactical situations or in the absence of reliable infrastructure. Thirdly, transmitting confidential audio data to third-party servers creates significant privacy and security risks. [1]

This work proposes a paradigm shift towards autonomous on-device analysis. A resource-efficient keyword spotting (KWS) model has been developed to eliminate the shortcomings of cloud-based solutions. The model provides a private, low-latency, and connection-resilient framework for threat detection, turning technical characteristics into strategic advantages in the field of cybersecurity. [2-4]

[1] Corresponding authors

didusavd@gmail.com (A. Didus); terejkowski@ukr.net (I. Tereikovskyi)

0009-0004-2235-6742 (A. Didus); 0000-0003-4621-9668 (I. Tereikovskyi)

2. Model Architecture and Methods

The proposed model is built on a modular principle, ensuring flexibility and adaptability. The architecture includes sequential stages of signal processing, feature extraction, "acoustic fingerprint" creation, and classification.

2.1. Signal Processing and Feature Extraction

The initial processing converts the raw audio signal into a set of numerical features. This stage includes:

Signal preparation: Voice Activity Detection (VAD) to isolate speech from background noise and amplitude normalization to compensate for volume variations.

Feature extraction: Calculation of Mel-frequency cepstral coefficients (MFCCs). To enhance reliability, static MFCCs are supplemented with dynamic features—the first (delta) and second (delta-delta) derivatives—which encode the rate and acceleration of spectral changes. This improves robustness to variations in speech tempo.

2.2. "Acoustic Fingerprint" Formation

The key element of the model is the transformation of a sequence of feature vectors into a compact string representation (an "acoustic fingerprint"). The process is based on the hypothesis of the unequal informativeness of different MFCCs. A weight vector W is applied to amplify the coefficients most significant for phonetic content. The weighted feature vectors are averaged, quantized, and serialized into a final string F according to Formula 1 below:

$$F = \text{Serialize}\left(Q\left(\frac{1}{T}\sum_{t=1}^{T}(M_t \odot W)\right)\right), \quad (1)$$

where M is the feature vector at time t, T is the number of frames, Q is the quantization function, and $\odot$ – is element-wise multiplication.

2.3. Classification by Levenshtein Distance

The final classification is performed by comparing the input "fingerprint" with reference templates from a pre-compiled dictionary. The Levenshtein distance is used as a similarity metric, which is a deterministic and computationally efficient method. The choice of this method ensures the transparency and reproducibility of results, which is critical for security systems. [5]

3. Experimental Study

To validate the model, a series of experiments was conducted using a software prototype implemented in Python.

Dataset: A specially recorded Ukrainian-language corpus containing commands for controlling a ground drone was used. Scalability was tested on lexicons of 10, 100, and 200 words.

Baseline model for comparison: To evaluate the effectiveness of the proposed improvements (dynamic features, weighting), a baseline model was implemented that used only static, unnormalized MFCCs.

4. Results

The model demonstrated high accuracy and reliability. In ideal conditions (clean signal, known speaker), the F1-score for a 100-word lexicon was 0.92. The system showed high resistance to noise, maintaining an F1-score of 0.78 at a signal-to-noise ratio of 5 dB. [2-4]

For contextualization of the results, a comparative analysis was conducted with the baseline model, the classic Dynamic Time Warping (DTW) algorithm, and a cloud-based ASR (Automatic Speech Recognition) service. The results for the 100-word lexicon are presented in Table 1. [6]

Table 1
Comparative analysis of KWS architecture performance

Model	Memory Footprint	Inference Time	RTF	F1-Score	F1-Score (5 dB noise)
Baseline	~150KB	~3 ms	0.003	0.75	0.45
Proposed	~250KB	~5 ms	0.005	0.92	0.78
Classical DTW	~2MB	~20 ms	0.02	0.88	0.65
Cloud-based	N/A	~450 ms	0.45	0.97	0.91

Conclusions

This paper presents a validated framework for transferring auditory threat detection tasks to peripheral devices. The proposed KWS model is lightweight, fully autonomous, and effective for critical applications where access to cloud infrastructure is limited or unavailable. Key findings of the study demonstrate that the "acoustic fingerprint" approach ensures privacy, ultra-low latency for real-time response, and resilience to network failures. Empirical validation confirmed that the model achieves an optimal balance between accuracy and computational efficiency. This enables a shift from a reactive, centralized security model to a proactive and decentralized one, strengthening the protection of the voice communication channel at the edge device level.

References

[1] D. O'Shaughnessy, "Trends and developments in automatic speech recognition research," Computer Speech & Language, vol. 83, 2024, 101538, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2023.101538.

[2] S. Alharbi et al., "Automatic Speech Recognition: Systematic Literature Review," IEEE Access, vol. 9, pp. 131858-131876, 2021, doi: 10.1109/ACCESS.2021.3112535.

[3] G. Chen, C. Parada and G. Heigold, "Small-footprint keyword spotting using deep neural networks," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 4087-4091, doi: 10.1109/ICASSP.2014.6854370.

[4] A. Mahmood and U. Köse, “Speech recognition based on convolutional neural networks and MFCC algorithm”, Adv. Artif. Intell. Res., vol. 1, no. 1, pp. 6–12, 2021.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989, doi: 10.1109/5.18626.

[6] Дичка, І. А., Терейковський, І. А., Дідус, А. В., Терейковська, Л. О., & Бояринова, Ю. Є. (2023). Оцінка ефективності засобів розпізнавання ключових слів у голосовому сигналі. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія: Технічні науки, 34(3), 1-7. https://doi.org/10.32782/2663-5941/2023.3.1/19.

[7] D. Seo, H. -S. Oh and Y. Jung, "Wav2KWS: Transfer Learning From Speech Representations for Keyword Spotting," IEEE Access, vol. 9, pp. 80682-80691, 2021, doi: 10.1109/ACCESS.2021.3078715.

[8] S. Dua et al., "Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network," Applied Sciences, vol. 12, no. 12, 2022, Art. no. 6223, https://doi.org/10.3390/app12126223.

[9] J. Oruh, S. Viriri and A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," IEEE Access, vol. 10, pp. 30069-30079, 2022, doi: 10.1109/ACCESS.2022.3159339.

[10] C. -H. H. Yang et al., "Decentralizing Feature Extraction with Quantum Convolutional Neural Network for Automatic Speech Recognition," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6523-6527, doi: 10.1109/ICASSP39728.2021.9413453.