Ministry of Education of the Republic of Belarus
Educational institution
Belarusian State University of Informatics and Radioelectronics

UDC 004.934

Do Anh Tu

Software for detecting speech in a signal

Abstract to master's degree thesis

Specialty 7-06-0611-02 Information Security

_____

Supervisor:
Zelmansky O. B., PhD, Associate Professor

_____

Minsk 2026

GENERAL DESCRIPTION OF WORK

The Research Aim and Objectives

The aim of the research is to analyze existing voice activity detection (VAD) algorithms and based on the results of this analysis, to develop a modular software tool capable of detecting speech in audio signals received from audio files or from a microphone in real time.

To achieve the aim, it is necessary to achieve the following objectives:

1) to analyze the existing methods of speech activity detection, including energy-based, zero-crossing rate, statistical, and deep learning approaches, and to identify their advantages and limitations;

2) to analyze the principles of audio signal preprocessing, including resampling, bandpass filtering, and noise reduction using spectral subtraction;

3) to select and justify the choice of two VAD algorithms – the WebRTC VAD (Google) and the Combined Energy-plus-ZCR VAD – for implementation in the developed system;

4) to implement the complete speech processing pipeline in Python, covering signal acquisition, preprocessing, voice activity detection, post-processing, and multi-modal output generation;

5) to implement a multilingual speech recognition module based on the OpenAI Whisper model supporting Russian, English, and Vietnamese;

6) to conduct experimental testing of both VAD algorithms under clean and noisy acoustic conditions using audio recordings in three languages, and to evaluate the results using standard metrics.

Work Connection with Priority Areas of Scientific Research

The topic of the thesis corresponds to paragraph 3 – Development of information and communication technologies, artificial intelligence systems, and software – of the list of priority areas of scientific, scientific-technical and innovative activities for 2021–2025, approved by the Decree of the President of the Republic of Belarus dated 07.05.2020 No. 156.

Thesis Results Approbation

The main results of the thesis were reported and discussed at Conference of Graduate Students, Undergraduates and Students of BSUIR (Minsk, 2025–2026).

INTRODUCTION

In the modern digital era, speech signal processing is gaining increasing importance across various fields of human activity. Voice Activity Detection (VAD) – the task of identifying time intervals containing speech within a continuous audio stream – represents a key preprocessing step for automatic speech recognition, voice-controlled interfaces, telephony noise suppression, speaker diarization, and audio compression. The relevance of this topic is driven by several fundamental factors.

Firstly, there is an exponential growth in the volume of audio data generated in various domains: from telecommunications systems and video conferencing to multimedia archives. Efficient detection and segmentation of speech segments within this data becomes critically important for subsequent processing and analysis.

Secondly, the development of Automatic Speech Recognition (ASR) systems and voice-based interfaces requires accurate extraction of speech fragments from continuous audio streams. The quality of solving these tasks directly depends on the effectiveness of the primary speech detection stage that precedes them.

Despite decades of research, a number of unresolved problems persist, limiting the effectiveness of existing speech detection methods:

1. The problem of processing signals in noisy conditions. Most real-world audio recordings contain various types of background noise – office environments, outdoor settings, electronic interference – which significantly complicates reliable speech detection. Traditional energy-based and zero-crossing-rate methods often demonstrate reduced accuracy as noise levels increase.

2. The trade-off between detection accuracy and computational simplicity. Algorithms based on adaptive statistical models, such as the WebRTC VAD, achieve higher robustness in noisy conditions but are more complex to analyze and tune than simple threshold-based detectors.

3. The lack of practical comparative tools. Although numerous individual VAD algorithms have been described in the literature, few openly available tools allow a direct, side-by-side comparison of a production-grade detector (such as WebRTC VAD) against a classical multi-feature detector under identical preprocessing and test conditions.

4. Real-time implementation constraints. Many high-accuracy methods based on deep neural networks require substantial computational resources, which can hinder their use on standard consumer hardware for real-time microphone-based applications.

5. The problem of defining speech segment boundaries. Accurately determining the start and end of speech segments remains challenging, particularly

during smooth transitions between speech and pauses, and short hesitations within continuous utterances.

The main objective of this Master's thesis is to develop a software tool for speech detection in audio signals that implements and experimentally compares two voice activity detection algorithms of contrasting design – a production-grade adaptive detector and a classical multi-feature detector – within a complete, modular processing pipeline.

To achieve this objective, the following tasks were set:

1. To conduct a comprehensive analysis of existing speech detection methods, including energy-based and zero-crossing-rate techniques, spectral approaches, statistical models (Gaussian Mixture Models, Hidden Markov Models), and modern deep learning architectures, in order to justify the selection of algorithms for implementation.

2. To implement two voice activity detection algorithms: the WebRTC VAD developed by Google, based on an adaptive Gaussian mixture model operating on sub-band energy features, and a Combined VAD that classifies a frame as voiced only when both its short-time energy exceeds an adaptive threshold and its zero-crossing rate falls below a fixed bound.

3. To create software supporting two input modes: processing of pre-recorded audio files in WAV, MP3, FLAC, OGG, and M4A formats, and real-time operation with microphone audio streams with immediate visual feedback.

4. To implement a signal preprocessing stage including bandpass filtering of the speech frequency band (300–3,400 Hz) and noise reduction via spectral subtraction or adaptive spectral gating, in order to reduce false detections caused by background noise.

5. To implement a mechanism for merging and smoothing raw frame-level VAD decisions, reducing both false positives from short noise bursts and false negatives from brief pauses within continuous speech.

6. To integrate a multilingual automatic speech recognition module, based on the OpenAI Whisper model, capable of transcribing the detected speech segments in Russian, English, and Vietnamese.

7. To conduct an experimental evaluation of both VAD algorithms on a set of audio recordings spanning multiple languages and acoustic conditions, using standard precision, recall, and F1-score metrics, and to compare their performance directly against each other.

The scientific novelty of the research lies in the systematic, controlled comparison of a production-grade adaptive VAD algorithm (WebRTC) against a classical multi-feature VAD algorithm (Combined Energy-plus-ZCR) under identical preprocessing conditions, frame parameters, and test data – including a

quantitative agreement analysis expressing the fraction of frames on which the two algorithms produce the same decision. This comparative methodology, combined with multilingual transcription support, distinguishes the developed tool from single-algorithm VAD implementations commonly found in the literature.

The practical significance of the work is determined by the broad application area of the developed software tool:

– automatic speech recognition pipelines, where VAD reduces the audio sent to the recognizer to only its speech-containing portions;

– voice-controlled interfaces and telephony systems requiring real-time speech detection from microphone input;

– multimedia archive processing, where long recordings must be automatically segmented into speech and non-speech regions;

– research and educational contexts requiring a transparent, modular reference implementation for comparing VAD algorithms.

The research employed a comprehensive approach combining theoretical analysis with practical implementation. Methods of digital signal processing – bandpass filtering, the Short-Time Fourier Transform, spectral subtraction – were applied together with the WebRTC adaptive statistical VAD algorithm and a custom-implemented threshold-based Combined VAD. Standard evaluation metrics (precision, recall, F1-score, word error rate) were used together with a set of audio recordings spanning three languages and three acoustic conditions for effectiveness evaluation.

The software tool was implemented in Python 3.10+ using NumPy, SciPy, librosa, soundfile, webrtcvad-wheels, PyAudio, openai-whisper, noisereduce, and matplotlib, ensuring cross-platform compatibility across Windows and macOS and straightforward extension with additional VAD algorithms or output formats in the future.

As a result of the work, a modular software tool for voice activity detection was developed and experimentally evaluated, achieving an average F1-score of 0.931 for the WebRTC VAD and 0.889 for the Combined VAD across 18 test recordings in three languages and three acoustic conditions. The developed system, its comparative evaluation methodology, and its multilingual transcription capability can form the basis for further research in speech signal processing and for the development of more complex speech processing systems.

MAIN PART

The first chapter provides a review of VAD methods and audio signal processing fundamentals – sampling, quantization, the Nyquist-Shannon theorem, bandpass filtering, STFT, and MFCC feature extraction. Threshold-based, statistical (GMM, HMM), and deep learning (CNN, LSTM, Transformer) VAD approaches are analyzed. The librosa library used for audio analysis in the system is also described [1-A].

The second chapter analyzes the two algorithms selected for implementation. The WebRTC VAD (Google) is examined: its adaptive Gaussian mixture model, aggressiveness parameter, and PCM input requirements. The Combined VAD is formulated mathematically: a frame is voiced when short-time energy exceeds an adaptive threshold AND zero-crossing rate falls below a fixed bound. The spectral subtraction noise reduction formula and the Whisper Transformer encoder-decoder architecture are also described [2-A].

The third chapter describes software development. The system is organized into seven Python modules covering the complete pipeline: audio acquisition, preprocessing, VAD, post-processing, and output generation. Structural schema, algorithm flowcharts, a UML class diagram, and sequence diagrams for the three main use cases are presented. Detailed specifications are given for all key functions [2-A].

The fourth chapter presents experimental results. All 12 functional test cases passed. Testing on 18 audio files in Russian, English, and Vietnamese under three acoustic conditions showed WebRTC VAD achieving an average F1-score of 0.931 and Combined VAD achieving 0.889. Noise reduction improved accuracy by 0.036 F1 points on noisy recordings. Whisper achieved average WER of 8.8% on clean speech with 91.7% correct language auto-detection.

CONCLUSION

This work set out to develop a functional software tool for Voice Activity Detection (VAD) – a system capable of detecting speech segments in audio signals loaded from files or captured in real time from a microphone. All primary objectives defined at the outset of the project have been successfully accomplished.

A comprehensive literature review was carried out in Chapter 1, covering the full evolution of VAD methods: from early threshold-based and spectral approaches, through statistical models such as Gaussian Mixture Models and Hidden Markov Models, to modern deep learning architectures including CNNs, LSTMs, and Transformer-based systems. This review established the theoretical foundation necessary for selecting and implementing the algorithms used in the developed system. Chapter 2 provided a detailed analysis of the specific algorithms chosen for implementation – the WebRTC VAD engine and the Combined Energy-plus-ZCR algorithm – examining their mathematical formulations, parameter sensitivity, and practical trade-offs.

The developed software implements a complete five-stage processing pipeline: audio acquisition, signal preprocessing, voice activity detection, post-processing, and multi-modal output generation. The system accepts input from audio files in any format supported by the librosa library – including WAV, MP3, FLAC, OGG, and M4A – as well as from microphone devices captured in real time via PyAudio. The preprocessing stage applies a fifth-order Butterworth bandpass filter covering the 300–3,400 Hz speech band and optionally applies noise reduction using spectral subtraction or the noisereduce library.

Two VAD algorithms were implemented and made available to the user. The WebRTC VAD, developed by Google and deployed in production in Chrome and Google Meet, uses an adaptive Gaussian mixture model operating at the frame level with a configurable aggressiveness parameter. The Combined VAD applies short-time energy and zero-crossing rate criteria simultaneously, labeling a frame as voiced only when both thresholds are satisfied. A post-processing stage common to both algorithms smooths the raw frame-level decisions by filling short silence gaps and removing spurious short speech bursts, and groups the resulting voiced frames into speech segment lists.

The system provides a comparison mode in which both algorithms are run on the same signal and their results are evaluated side by side, computing per-algorithm statistics – speech duration, number of segments, average segment duration, and speech percentage – as well as a pairwise agreement matrix expressing the fraction of frames on which both algorithms agree. This comparison functionality directly supports the research objective of evaluating and contrasting VAD approaches under identical conditions.

The output layer of the system produces up to five simultaneous artifacts: a waveform plot with highlighted speech regions saved as PNG, a Mel spectrogram and MFCC visualization saved as PNG, a WAV file containing only the detected speech segments, a plain-text transcript produced by the OpenAI Whisper speech

recognition model, and a comparison dashboard plot when both algorithms are evaluated together. Whisper supports over 90 languages including Russian, English, and Vietnamese, and can automatically detect the language of the recording without requiring the user to specify it.

Experimental testing was conducted on 18 audio files spanning three languages (Russian, English, and Vietnamese) and three acoustic conditions (clean studio recordings, office noise, and outdoor noise), with file durations between 30 and 58 seconds. The WebRTC VAD achieved an average F1-score of 0.931 across all test files, outperforming the Combined VAD (average F1 = 0.889) in all conditions. The performance gap between the two algorithms was smallest in clean conditions (0.022 F1 points) and widest in outdoor noisy conditions (0.055 F1 points), confirming that the WebRTC algorithm's adaptive noise model handles non-stationary noise significantly better than the fixed-threshold energy and ZCR approach. Noise reduction preprocessing was shown to improve WebRTC VAD accuracy by an average of 0.036 F1 points on noisy recordings. The Whisper base model achieved word error rates of 8.8% on clean recordings and 18.3% on noisy recordings, with correct language auto-detection in 91.7% of test cases.

The software is organized into seven independent Python modules – main.py, vad_detector.py, noise_reduction.py, mic_recorder.py, vad_comparison.py, stt_module.py, and cli_ui.py – following a functional, stateless design that allows any module to be tested or replaced independently. The command-line interface supports both an interactive menu mode, where all settings are collected step by step, and a direct argument mode for scripted or automated use. A real-time ASCII waveform display provides immediate visual feedback during microphone recording, showing signal amplitude and VAD decisions frame by frame in the terminal.

Several directions exist for extending the current work. The most impactful improvement would be the integration of a deep learning VAD model such as Silero VAD, which would substantially increase detection accuracy in noisy and outdoor conditions without requiring manual parameter tuning. A graphical user interface built with Tkinter or PyQt would make the system more accessible to users without command-line experience. The Whisper model size could be made adaptive – automatically selecting a larger model when audio quality is sufficient and a smaller model when speed is prioritized. Finally, a speaker diarization module could be added to segment multi-speaker recordings not only by speech activity but also by speaker identity, broadening the system's applicability to meeting transcription and interview analysis.

In summary, the developed VAD system successfully combines two classical speech detection algorithms with modern speech recognition, noise reduction, and visualization capabilities in a modular, cross-platform Python application. The system fulfills all stated objectives and provides a solid foundation for further research and development in the field of speech signal processing.

LIST OF OWN PUBLICATIONS

1-A. Do A. T., Feature Extraction - MFCC for Speech Processing / A. T. Do // Information Security: Collection of Materials of the 61th Scientific Conference of Graduate Students, Undergraduates and Students of BSUIR, Minsk, April 21–25, 2025 / Belarusian State University of Informatics and Radioelectronics. – Minsk, 2025. – P. 38-43.

2-A. Do A. T., Speech Activity Detection in Real Noise / A. T. Do // Information Security: Collection of Materials of the 62th Scientific Conference of Graduate Students, Undergraduates and Students of BSUIR, Minsk, April 13–17, 2026 / Belarusian State University of Informatics and Radioelectronics. – Minsk, 2026. – P. 13-18.