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

UDC 004.934.8

Chen Zhengyu

SOFTWARE FOR RECOGNIZING SPEAKER BY VOICE

Abstract to master degree thesis

in the Specialty 1-98 80 01 Information Security

_____

Supervisor
Associate Professor of the
Department of Information Security,
PhD,
Associate Professor
Zelmansky O.B.

_____

Minsk 2026

GENERAL DESCRIPTION OF WORK

The Research Aim and Objectives

The aim of the research is to design and implement a comprehensive software system for recognizing speakers by voice based on deep learning technology, with the speaker recongition model serving as the core architecture for speaker embedding extraction.

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

1) to analyze the existing speaker recognition methods and technologies, including speech signal processing, audio feature engineering, and deep neural network modeling;

2) to study the CAM++ (Context-Aware Masking plus plus) model architecture, including its front-end convolutional module, densely connected time-delay neural network backbone, and context-aware attention mechanism;

3) to investigate the VoxCeleb1 and VoxCeleb2 datasets, and implement data augmentation techniques including background noise overlay from the MUSAN corpus and reverberation enhancement using room impulse response data;

4) to implement the model training process using FBank feature extraction and AAM-Softmax loss function with Adam optimizer;

5) to implement a complete voiceprint recognition system on the Python platform using PyTorch framework, including functions for voiceprint registration, recognition, user deletion, and real-time recognition;

Work Connection with Priority Areas of Scientific Research

The topic of the thesis corresponds to paragraph 6 Ensuring the security of humans, society and the state (means of technical and cryptographic information protection, cryptology and cybersecurity) 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 dissertation were reported and discussed at the 61st Scientific Conference of Graduate Students, Master Students and Students of the Educational Institution "Belarusian State University of Informatics and Radioelectronics" (Minsk, April 21–25, 2025), and the 62nd Scientific Conference of Graduate Students, Master Students and Students of the Educational Institution "Belarusian State University of Informatics and Radioelectronics" (Minsk, April 13–17, 2026).

INTRODUCTION

Voiceprints, like fingerprints and faces, are unique biological characteristics of each human body. In recent years, with the rapid development of deep learning technology, speaker recognition technology has become more and more widely used, for example in call center operation, criminal investigation, smart speakers and robotics, phone security, and other fields.

This thesis focuses on the design and implementation of a comprehensive software system for recognizing speakers by voice, which encompasses two primary aspects: the selection and training of an appropriate deep learning model, and the design of an intuitive front-end interface to facilitate practical usage. The proposed system is built upon the CAM++ (Context-Aware Masking plus plus) model, which adopts a convolutional module as its front end to extract low-level acoustic features from raw speech signals, and employs densely connected time-delay neural network (TDNN) blocks as its backbone architecture. Within each CAM++ block, an improved context-aware attention mechanism is incorporated, enabling the model to dynamically weight the importance of different temporal frames and feature channels, thereby capturing distinctive speaker characteristics effectively while maintaining high computational efficiency.

Compared to earlier models such as ECAPA-TDNN or traditional x-vector systems, CAM++ significantly reduces the amount of computation required and increases inference speed, making it particularly suitable for real-time or resource-constrained deployment scenarios. Prior to model training, FBank (Filter Bank) features are extracted from the training data to simulate the frequency resolution characteristics of the human auditory system, thereby enhancing the model's perceptual accuracy and overall recognition performance. Extensive experiments were conducted on the combined VoxCeleb1 and VoxCeleb2 databases, which together contain approximately 7,205 distinct speakers, using FBank as the preprocessing method and the Additive Angular Margin Softmax (AAM-Softmax) loss function to enforce discriminative speaker embeddings with clear class separation. Evaluation on the unified VoxCeleb1 and VoxCeleb2 test set yielded an Equal Error Rate (EER) of 2.193% and a minimum Detection Cost Function (MinDCF) of 0.1381, quantitatively demonstrating the effectiveness and robustness of the implemented system.

For the practical deployment of the proposed model in real-world applications, a user-friendly graphical interface has been developed using PyQt5, a robust Python framework for desktop application development, enabling seamless interaction between end users and the underlying recognition engine. Upon launching the program, users are presented with an intuitive window that provides comprehensive functionality for voiceprint management and speaker recognition. Specifically, the interface supports speaker enrollment, allowing users to conveniently register their voice samples along with corresponding identity labels for future identification. It also facilitates identity verification by capturing short-duration speech input and automatically returning the recognition result, eliminating the need for manual

command-line operations. To enable efficient long-term maintenance of enrolled speaker information, a speaker removal function is incorporated, giving users the flexibility to delete obsolete or incorrect entries from the voiceprint library. In addition, the system offers a dedicated real-time recognition mode, in which the microphone input is continuously monitored and processed on the fly, enabling dynamic speaker verification without any manual intervention. This combination of a high-performance CAM++ backend and a thoughtfully designed PyQt5-based graphical front end makes the proposed system both powerful and accessible, rendering it suitable for a wide range of practical voiceprint-based authentication scenarios.

MAIN PART

The first chapter provides a detailed introduction to the background and significance of voiceprint recognition technology, highlighting its unique advantages as a biometric modality compared to traditional methods such as fingerprints or facial recognition. The chapter describes the research approach and framework, including platform construction using Python, PyTorch, and CUDA libraries, as well as the overall design of the CAM++ model as the core architecture for speaker embedding extraction. The main components of the model, including the front-end convolutional module and the densely connected time-delay neural network, are briefly introduced.

The second chapter presents the relevant theories and main technologies used in the voiceprint recognition system. It covers speech signal processing methods such as framing, windowing, Fourier transform, and FBank feature extraction. The chapter also explains deep neural network modeling, including convolutional neural networks and deep neural networks, and provides a brief introduction to classic voiceprint recognition models ECAPA-TDNN[1-A], as well as related technology libraries such as PyTorch and CUDA. The theoretical foundation for subsequent model design and implementation is established.

The third chapter describes the dataset acquisition and processing pipeline. The study adopts the combined VoxCeleb1 and VoxCeleb2 datasets as the training set, containing approximately 7,205 distinct speakers. Data augmentation techniques are introduced, including background noise overlay from the MUSAN corpus and reverberation enhancement using room impulse response data[2-A].

The fourth chapter presents the overall structure of the proposed CAM++ model. The feature extraction module, FCM module, ResNet and ResBlock modules, as well as TDNN and Dense-Block components are described. The chapter explains how densely connected time-delay neural networks and context-aware attention mechanisms work together to extract discriminative speaker embeddings from speech signals.

The fifth chapter describes the training process of the model. The loss function adopts AAM-Softmax, which introduces an additive angular margin to enforce separation between different speaker classes. The training data, preprocessing steps, training procedure, and implementation details.

The sixth chapter presents the model evaluation framework. The evaluation adopts the VoxCeleb1 and VoxCeleb2 standard protocols, using Equal Error Rate and minimum Detection Cost Function as core metrics. Experimental results show that the developed system achieves a low Equal Error Rate and a low minimum Detection Cost Function, demonstrating satisfactory accuracy and robustness. The chapter also discusses the convergence characteristics and optimal discrimination threshold of the model during training.

The seventh chapter describes the system design and implementation. The system adopts a three-layer architecture consisting of a user interface layer built with Tkinter and PyQt5, a business logic layer that includes the CAM++ model and

cosine similarity matching, and a data storage layer. The chapter explains the voiceprint comparison, registration, recognition, user deletion, and real-time continuous recognition modules. The real-time recognition mechanism employs a multi-threaded architecture with a sliding window, enabling low-latency speaker identification.

CONCLUSION

Speaker recognition is an important means of biometric identification in daily life and has found wide applications in areas such as security authentication, smart homes, call center verification, and forensic investigations. This thesis presents a comprehensive study on the design, implementation, and evaluation of a deep learning-based voiceprint recognition system. The research covers the entire pipeline from theoretical foundations to practical deployment, adopting the CAM++ model as the core architecture for speaker embedding extraction.

Currently, the most important technical methods in speaker recognition include speech signal processing and feature extraction techniques such as FBank and MFCC, deep neural network modeling approaches including convolutional neural networks and time-delay neural networks, as well as attention-based context-aware masking mechanisms.

The thesis provides a detailed analysis of the CAM++ model architecture, highlighting three key aspects: the densely connected time-delay neural network backbone, the context-aware masking attention mechanism, and the AAM-Softmax loss function, which enforces angular separation between different speaker classes.

Model training and evaluation are conducted on the combined VoxCeleb1 and VoxCeleb2 datasets, which contain approximately 7,205 distinct speakers. A user-friendly graphical interface is developed using Tkinter and ttk, offering functions such as voiceprint registration, recognition, user deletion, and real-time continuous monitoring. The real-time recognition mechanism employs a dual-threaded producer-consumer architecture with a sliding window, enabling low-latency speaker identification.

Nevertheless, the application environment of voiceprint recognition systems remains complex and diverse, and several challenges still need to be addressed. Most current research on speaker recognition focuses on short-duration speech and either text-dependent or text-independent tasks. Compared with short-duration speech, long-duration conversational speech and multi-speaker scenarios exhibit much more complex characteristics.

LIST OF OWN PUBLICATIONS

1-A. Zhengyu, C. Software for recognizing speaker by voice / Chen Zhengyu // 61-я науч. конф. аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» : материалы докл. науч. конф., Минск, 21-25 апреля 2025 г. / БГУИР. – Минск, 2025. – С. 14-17.

2-A. Zhengyu, C. Software for recognizing speaker by voice / Chen Zhengyu // 62-я науч. конф. аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» : материалы докл. науч. конф., Минск, 13-17 апреля 2026 г. / БГУИР. – Минск, 2026