MICROPHONE SPEAKER ANALYSIS: AUDIO SEGMENTATION AND FREQUENCY INSIGHTS

Taisia-Maria COCONU[1], Costin-Alexandru DEONISE[1],
Constantin ANGHEL[2], Cătălin NEGRU[1], Florin POP[1,3,4]

Abstract. *Audio segmentation represents a technical process used for separating a stream of audio recordings, which frequently contain multiple speakers, into uniform sections. This paper explores the implementation of voice-dialing and recognition algorithms to examine and analyze the technology's capability to accurately identify and differentiate speakers in intricate environments. It aims to enhance our understanding of the technology's functionality, including its ability to discern speakers' emotions and gender. Additionally, a hardware simulation is conducted using a two-way microphone and an Arduino board. It seeks to emphasize precision in speaker recognition and diarization, along with the accurate transcription of speeches, by achieving optimal parameters and enhancing existing market models. It also explores the applicability of this technology in various fields by creating applications that mainly use Speech Diarization and Speech Recognition.*

Keywords: Emotion Detection, Gender Detection, Voice Recognition Hardware System.

DOI [10.56082/annalsarsciinfo.2024.1.5](https://doi.org/10.56082/annalsarsciinfo.2024.1.5)

1. Introduction

Speaker diarization is a highly relevant paradigm in the current technological era, focused on identifying, segmenting, and assigning speakers within a continuous speech stream during conversations or speech events. Through speaker diarization, speakers can be detected and identified during conversations, enabling voice frequency analysis to determine the **speaker's gender and emotions**, even in complex scenarios with overlapping speech. From **security** and monitoring to **education**, **health**, **voice assistance** and even **information management**, this concept proves to be essential and overwhelmingly useful [2].

[1] National University of Science and Technology Politehnica Bucharest, Romania
[2] National Institute of Research and Development in Mechatronics and Measurement Technique, Bucharest, Romania
[3] National Institute for Research and Development in Informatics (ICI), Bucharest, Romania
[4] Academy of Romanian Scientists, Bucharest, Romania

Modern speaker diarization approaches incorporate advanced machine learning techniques, including deep neural networks and clustering algorithms, to achieve precise and robust speaker identification and speech-to-text conversion. These methods address the challenges of identifying multiple speakers in an audio recording, accounting for emotional variations that can affect speech patterns. Applied to specific datasets, diarization models can achieve accuracy rates exceeding 95.4%, indicating the proposed method's potential for high accuracy [3]. The most common voice identification and segmentation models are **SpeechBrain**, **InaSpeechSegmenter**, **Picovoice**, **WebRTC** and our **improved model** which can achieve performance of up to 98%. The performances of these models are compared in the following sections.

The paper is structured as follows. Section 2 presents the existing models of speaker diarization with critical remarks. Section 3 highlights the current solutions for emotion and gender detection models. The hardware aspects are presented in section 4, while Section 5 presents the applications for speech recognition and speech diarization. Section 6 concludes our work.

2. Existing models of Speaker Diarization

This section reviews the current approaches addressing the issues targeted by this thesis. We present various existing models for Speech Detection, examining their performance metrics, advantages, and disadvantages. The analysis includes a table summarizing the key characteristics of these models, along with graphs illustrating their performance.

SpeechBrain is a comprehensive conversational AI toolkit that supports speaker recognition, voice-to-speech translation, sound separation, speech recognition, and spoken language understanding, among other functionalities. It encompasses a wide array of audio technologies, including sound event recognition, audio augmentation, and multi-microphone signal processing, leveraging advanced deep learning techniques like self-supervised learning, continuous learning, diffusion models, Bayesian deep learning, and neural networks. The SpeechBrain toolkit offers a user-friendly experience with easy customization and flexibility, integrating numerous conversational AI technologies. As well as performance, this model shows the highest precision, but it can be improved by modifying default threshold for recognizing speech segments to detect them correctly [7].

Picovoice is a platform for creating custom voice solutions that can recognize specified keywords and then interpret the intention behind the subsequent spoken command. It employs the Porcupine engine to detect keyword phrases, providing

offline speech recognition tailored to unique phrases and scenarios. After initialization and processing the audio file, each audio frame is analyzed by Picovoice for real-time interpretation [8].

InaSpeechSegmenter, an audio segmentation toolbox based on CNN, divides audio signals into segments such as noise, music, and speech-like sections. It identifies speech segments based on the speaker's gender (male or female). The tool is designed to facilitate detailed research on speaker gender detection by calculating the proportion of speaking time occupied by men and women. It displays multiple intervals indicating the presence of male or female voices, periods of no noise, and intervals where background music is detected [10].

WebRTC utilizes the Gaussian Mixture Model (GMM) as its foundation, known for its speed and accuracy in distinguishing between noise and silence. However, its performance may decline when differentiating speech from background noise. The VAD operates by analyzing short audio frames and providing results for each frame. Enhancements in VAD effectiveness can be achieved by configuring parameters like silenceDurationMs and speechDurationMs, allowing for the detection of longer utterances and minimizing false positives during pauses between sentences [9][13][14].

SpeechBrain achieved the highest scores, boasting an average recall of 0.97 and an average precision of 0.96. **Picovoice** also demonstrated strong performance, while **InaSpeechSegmenter** delivered acceptable results. In contrast, **WebRTC** performed less effectively. These results underscore the importance of selecting the right VAD model based on specific requirements and input data. Below, we synthesized the most important performance metrics of these models to decide which has better performance and deserves to be improved.

Model/ Tool	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Loss	MER (%)
SpeechBrain	96	97	91	92	0.23	2.29
Picovoice	94	96	91	91	0.3	6
WebRTC	90	92	87	87	0.5	5
InaSegmenter	93	94	88	89	0.2	5.5

Table 1: Comparison of Performance Metrics for Different Speech Processing Models.

In previous research, we enhanced **our speech recognition model** using the SpeechBrain framework with a vast multilingual dataset exceeding 2 terabytes. Real-time noise augmentation during training bolstered robustness. Training over 100 epochs yielded significant improvement in performance. Advanced preprocessing and hyperfeature optimization were pivotal in achieving paradigm-shifting voice activity detection (VAD). Our model excelled with **0.98 recall** and **0.97 precision**, outperforming others. Architectural upgrades included **RNN layers and expanded CNN channels,** capturing intricate audio details. Testing on the KAIST dataset [4] affirmed SpeechBrain's superiority in detecting speech amidst noise [1].

Below we can see a table which contains the most important characteristics of discussed models.

Characterisitic	SpeechBrain	Picovoice	InaSegmenter	WebRTC
Platform	Open source, based on PyTorch	Complete, fully runs on-device	Audio segmentation toolkit based on CNN	Based on a Gaussian Mixture Model
Key Features	Speech recognition, enhancement, sound separation, text-to-speech	User recognition from naturally spoken phrases, offline speech recognition functionalities	Audio segmentation into speech, music, and noise, speaker gender classification	Speech and silence detection in short audio frames
Performance	Supports modern deep learning technologies, language model training	Outperforms cloud-based alternatives by significant margins	Provides precise segmentation and speaker gender classification	Efficient in distinguishing between noise and silence
Accesibility	Open source, with extensive documentation and tutorials	Offers free start without limited trial	Simple-to-use API	Offers parameters to enhance speech detection capability
Advantages	Wide range of functionalities, extensive documentation and tutorials	Fully runs on-device, offering data control, efficient	Accurate and classification in audio segmentation	Efficient in real-time speech and silence detection

Table 2: Comparison of Speech Processing Tools

3. Emotion and Gender Detection Models

Emotion recognition in speech finds application in various fields, including voice assistance, assessing users' emotional states in mental health applications, and monitoring emotions during interactions with automated systems [6].

The core task of a voice emotion recognition system involves transforming speech patterns into parametric representations at lower data rates by using SVC to construct and train a model for emotion classification. For training and testing this model, we used the **RADVESS dataset**. It's critical to balance the data set and assess how well the model performs on the test and training sets. The speech emotion recognition system allows users to experiment with different emotions available, such as *calmness, happiness, neutrality, boredom, pleasure, anger, sadness, disgust, fear* despite other existing models being able to detect four or eight emotions.

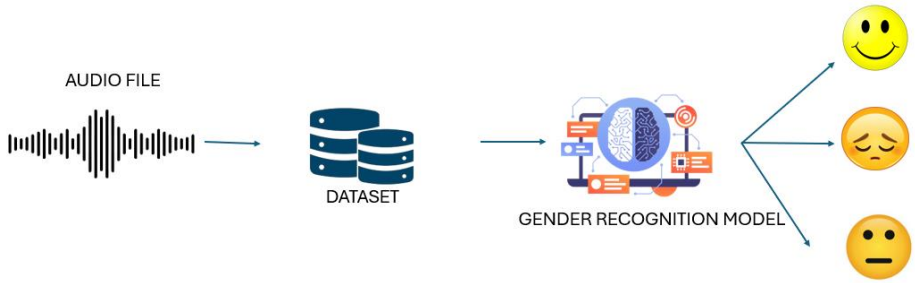

Figure 1: Emotion Recognition Model.

We created two test scripts, one for **recognizing the voice from the input** and displaying the speaker's emotions on the screen and one that **takes the data from the audio files** provided and displays his emotional state. To achieve high performance for this model, the hyperparameters of the classifiers and regressors must be modified and optimized. For this, we applied two dedicated algorithms like **Grid Search** and **Random Search**, obtaining the most suitable parameters for our model. The following image shows a comparison between the performances of our emotion detection model and other existing models.

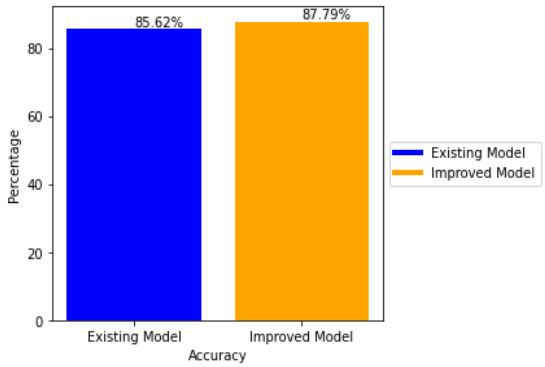

Figure 2: Comparison of Performance Metrics: Improved vs. Existing Emotion Model.

To determine the **gender of speakers**, we developed a new program using a large dataset, specifically **Mozilla's Common Voice** [5].

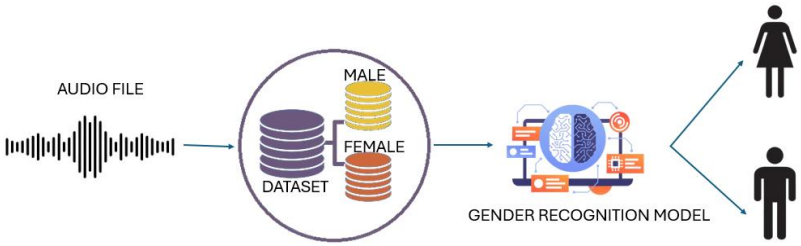

Figure 3: Gender Recognition Model.

Initially, we filtered out invalid samples and selected those that met our criteria within the gender framework. Each vocal sample was converted into a fixed-length vector, ensuring a balanced representation of both male and female samples. We then constructed a model using a customizable function to enhance its performance, which outputs the predicted gender and associated probabilities. the neural network used is a feed-forward network with five dense layers. To find out the necessary number of epochs for training this model, we used **early stopping** with a **dropout rate of 5 epochs.** So, the model training will stop after a smaller number of epochs than the one set at the beginning. Training involved multiple epochs to optimize the model, followed by evaluation to assess accuracy and losses as it can be seen in the figure below. As we expected, advancing in the number of epochs determined the accuracy increasing and the losses decreasing.

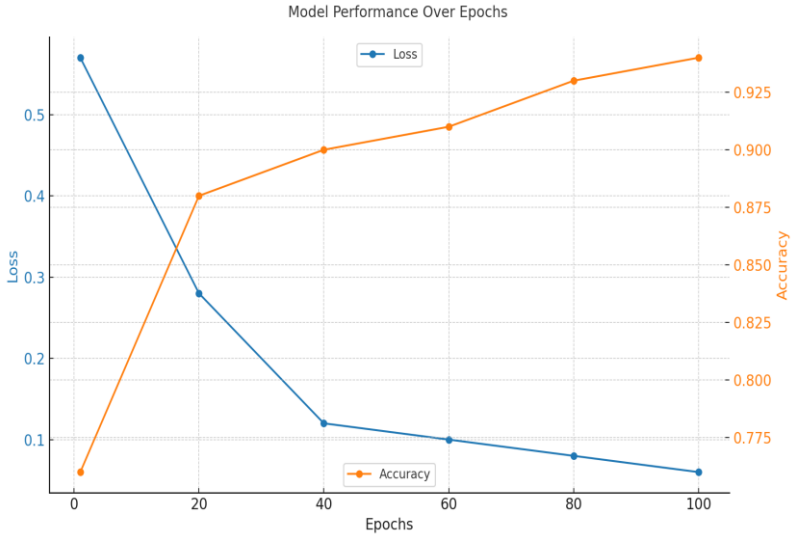

Figure 4: Performance Metrics of Gender Model.

4. Hardware Aspects

Most existing voice activity detection projects utilize a microphone to capture sound signals from input, where detection triggers the activation of an LED. These projects employ machine learning to train models that respond to specific commands such as "LIGHT ON," "LIGHT OFF," and "NOISE" [12]. Data sets can be generated using available open-source tools to facilitate model training. There are several types of microphones that can be used for Speaker Recognition, but the most common are a **directional microphone with adjustable directivity feature** or an **adjustable directivity directional microphone** [11]. Depending on the characteristic we want to get for the new hardware model, we can choose between an Arduino and a Raspberry Pi board.

Characteristic	Directional Speech Detection Microphone	Adjustable Directivity Directional Microphone
Directivity Pattern	Kidney, suitable for feedback reduction	Bullet, wide kidney, kidney, super kidney, hyper kidney, figure of eight
Frequency Characteristic	Compensated below 25 Hz, suitable for general applications	Compensated below 16 Hz (-12 dB), more precise and efficient for low-frequency control
Output Types	Jack (9 V battery) or XLR	Jack (9 V battery) or XLR
Versatility, Applications and	Limited to kidney directivity characteristic and output options, suitable for feedback reduction in	Much more versatile due to adjustable directivity and frequency characteristics, suitable

Adjustment	performance or recording and limited to minimal microphone sensitivity adjustments	for use in studio recordings and adjustable to compensate for specific acoustic effects

Table 3: Comparison between Directional Speech Detection Microphone and Adjustable Direction Directional Microphone (with information from [11]).

We have developed a **voice recognition system** specifically trained for use in smart home automation. In this context, a unidirectional microphone that detects noise is particularly useful at night (as identified by a photo resistive sensor), enabling the system to turn on an LED to illuminate the way to the house. These sensors are connected to the pins of an Arduino UNO board.

Table 3 presents the most important characteristics of the two microphones discussed and can be considered a **landmark** in terms of a suitable choice in correct, efficient and complete speech detection.

5. Applications for Speech Recognition and Speech Diarization

We developed several applications to test the applicability of the models in real life through use cases. The most important applications involve:

- **Speaker Diarization using a graphical interface**: the user selects a desired audio file, and our improved model makes the segmentation, showing after three seconds of processing the segments where was recognized voice in the audio file.

- **Karaoke**: the user reads out loud the words printed on screen and if a word is recognized, it is highlighted else nothing happens.

- **Speaker Transcription**: the user starts speaking and the words he says are saved in a *.txt file* (he can speak in any language because our model is able to detect almost all languages due to the library used for training – KAIST dataset [4]). After saving the file, we trained a new model to add punctuation in the file and its accuracy is about 87%.

Conclusions

As previously noted, this study demonstrated that speech activity identification is complex, influenced by factors such as input parameters, training datasets, and vocal signal characteristics. Evaluating models on diverse, realistic datasets is crucial for optimal performance and developing more accurate models.

Our analysis showed that the SpeechBrain model, trained over 100 epochs, outperforms other VAD models in precision. This underscores the importance of continuous evaluation and potential improvements in speech processing. For emotion and gender detection, we developed two programs using public data and advanced preprocessing.

Emotion recognition employed complex signal transformations and SVC algorithms, while gender identification used audio format transformations and Grid Search for optimization. Hardware implementations of these models can benefit fields like medicine, security, and customer service.

Acknowledgments

The work presented in this paper was supported by the Core Program within the National Research Development and Innovation Plan 2022-2027, financed by Ministry of Research, Innovation and Digitalization of Romania, project no 23380601. This work is partially supported by the Research Grant no. 94/11.10.2023 Modern Distributed Platform for Educational Applications in Cloud Edge Continuum Environments GNAC-ARUT-2023.

We would also like to thank the reviewers for their time and expertise, constructive comments and valuable insight.

R E F E R E N C E S

[1] Deonise, Costin-Alexandru, Taisia-Maria Coconu, Traian Rebedea, and Florin Pop. "Improved Speech Activity Detection Model Using Convolutional Neural Networks." In *2023 22nd RoEduNet Conference: Networking in Education and Research (RoEduNet)*, pp. 1-8. IEEE, 2023.

[2] Anguera, Xavier, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. "Speaker diarization: A review of recent research." *IEEE Transactions on audio, speech, and language processing* 20, no. 2 (2012): 356-370.

[3] Al-Hadithy, Thaer M., and Mondher Frikha. "A Real-Time Speaker Diarization System Based On Convolutional Neural Networks Architectures." In *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1-9. IEEE, 2023.

[4] Seo, Deokjin, Heung-Seon Oh, and Yuchul Jung. "Wav2kws: Transfer learning from speech representations for keyword spotting." *IEEE Access* 9 (2021): 80682-80691.

[5] Fadheli, A. (2019). *Gender recognition using voice*. In GitHub repository (1.0.0) [Computer software]. GitHub. https://github.com/x4nth055/gender-recognition-by-voice.

[6] Fadheli, A. (2019). *Speech Emotion Recognition*. In GitHub repository (1.0.0) [Computer software]. GitHub. https://github.com/x4nth055/emotion-recognition-using-speech.

[7] Ravanelli, Mirco, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan et al. "SpeechBrain: A general-purpose speech toolkit." *arXiv preprint arXiv:2106.04624* (2021).

[8] Eric Mikulin. Picovoice. In GitHub repository (3.0) [Computer software]. GitHub. https://github.com/Picovoice/picovoice.

[9] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[10] Doukhan, David, Eliott Lechapt, Marc Evrard, and Jean Carrive. "Ina's mirex 2018 music and speech detection system." *Music Information Retrieval Evaluation eXchange (MIREX 2018)* (2018).

[11] Hairol Shah, M.Z. Rashid, Mohd Fairus Abdollah, Muhammad Nizam Kamarudin, C.K. Lin, and Z. Kamis. Biometric voice recognition in security system. *Indian Journal of Science and Technology*, 7:104–112, 02 2014.

[12] Ibrahim, Dogan. *PIC microcontroller projects in C: Basic to advanced*. Newnes, 2014.

[13] Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492-1500. 2017.

[14] Radosavovic, Ilija, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. "Designing network design spaces." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428-10436. 2020.