Speaker Identification in Crowd Speech Audio using Convolutional Neural Networks

Ghadeer Qasim Ali[1], Husam Ali Abdulmohsin[2,*]

[1]Computer Science Department, College of Science, University of Baghdad

Emails: Ghadeer.Ali2201m@sc.uobaghdad.edu.iq; Husam.a@sc.uobaghdad.edu.iq

Abstract

Crowd speaker identification is the most advanced technology in the field of audio identification and personal user experience which researchers have extensively focused on, but still, science hasn't been able to achieve high results in crowed identification. This work aims to design and implement a novel crowd speech identification method that can identify speakers in a multi speaker environment, (two, three, four and five speakers). This work will be implemented through two phases. The training phase is the Convolutional Neural Network (CNN) training and testing phase. Through this phase, the training will be implemented on data generated via the Combinatorial Cartesian Product approach. This approach uses two primary processes, the Computation of the Cartesian product process and combinatorial selection process. The second phase is the prediction phase. The aim of this phase is to check the CNN trained in the first phase, through testing it on new crowed audios than the data that the CNN was trained on in the first phase, these new crowded audios exist in the Ghadeer-Speech-Crowd-Corpus (GSCC) dataset, which is a new database designed through this work. Compared to the state-of-the-art speaker identification in multi speaker environment approaches, the results are impressive, with a recognition rate of 99.5% for audio with three speakers, 98.5% for music with four speakers, and 96.4% for audio with five speakers.

Keywords: Crowded speech identification; Combinatorial Cartesian Product; GSCC Dataset

1. Introduction

Speaker identification technology is essential for voice-activated devices, security, and transcription. This approach identifies people based on their vocal features. However, the work becomes more complicated when several speakers are talking at the same time, resulting in overlapping audio signals. Traditional speaker identification methods, which rely on manual criteria and statistical models, fail to accurately identify speakers in complex auditory situations. Recent years have witnessed a significant boom in research focusing on the use of speakers' voices for multiple purposes, because of rapid technological advances in the fields of artificial intelligence, machine learning, and voice recognition. The human voice is one of the most prominent biometric features, as it carries with it unique information that distinguishes each individual from another, making its use a fertile field for research and development [1-4].

One of the earliest methods used was based on statistical models such as Hidden Markov Models (HMM). These models are used to represent temporal sequences of sound and to recognize the acoustic patterns that characterize each speaker [5]. With the development of technology and increased computing power, voice recognition techniques have become more accurate and effective, entering into the use of deep learning techniques and neural networks that have greatly improved the ability of systems to recognize speakers in multiple and diverse environments [6].The primary contribution of this study lies in the Combinatorial Cartesian Product Approach and application of a Convolutional Neural Network (CNN) for the task of speaker recognition in mixed audio environments where multiple speakers are speaking simultaneously.

DOI: https://doi.org/10.54216/FPA.160208
Received: December 21, 2023 Revised: February 25, 2024 Accepted: June 02, 2024

This approach is novel in its combination of advanced feature extraction techniques and deep learning for multi-speaker audio classification. By extracting Mel Frequency Cepstral Coefficients (MFCCs), chroma features, and Mel spectrograms, the method captures comprehensive audio signal characteristics that are crucial for distinguishing between different speakers. The proposed CNN architecture effectively learns and differentiates these features, achieving an unprecedented accuracy of 99.9%, 99.5%, 98.5% and 96.4% for 2, 3, 4 and 5 mixing speakers. This significant result demonstrates the potential of CNNs in handling complex audio recognition tasks, thus advancing the field of audio signal processing and providing a robust solution for real-world applications where speaker overlap is common.

Section 2 describe related work; Section 3 describes the Dataset is used in our work and the approach to generate the mixing of audio. In section 4 we describe the methodology, feature extraction that are used to build this model, Model Architecture and Training are described. The section 5 represents the result of the proposed method. Finally, the conclusion in section 6.

2. Related work

To have a general idea of the approach used from previous studies to identification speaker in crowd environments. In 2018, [7] develop a multi-target speaker detection and identification system through the fusion of Probabilistic Linear Discriminant Analysis (PLDA) and Deep Neural Network (DNN).The baseline PLDA approach achieved a Top-1 detector Equal Error Rate (EER) accuracy of 7.38% after 40 iterations, while the fusion of PLDA and DNN improved system performance, enhancing discrimination of blacklist and background speakers. The DNN model showed varying accuracies, with the smallest Top-S detector EER on the development set and the highest on the test set. The Top-S detector EER reached an accuracy of 1.78% with a specific parameter after 40 iterations.

In 2018, [8] designed system to distinguish between single-speaker and multi-speaker audio files, and then identify the speaker(s) accurately. The method used involved pre-processing of audio files to extract features, including reduction, silence removal, framing, windowing, and DCT calculation. Feature extraction was done using the Mel Frequency Cepstral Coefficients (MFCC) technique. The system was trained using neural networks and the Error Back Propagation Training Algorithm (EBPA). From 25 recording of multi speaker 22 recordings identified correctly.

In 2019, [9] presented a novel heterogeneous-input multi-channel acoustic model (AM) for distant multi-talker speech recognition. Initially, a single-channel AM is trained, followed by training a multi-channel AM with a randomly initialized multi-channel input branch. The model incorporates a complemental speech enhancement (SE) module to enhance the AM performance.
The research was instrumental in the development of the Hitachi/JHU CHiME-5 system that achieved the second-best result in the CHiME-5 competition. The research evaluated the proposed AM method on the AMI Meeting Corpus and achieved significant outcomes. For the AMI Corpus, it achieved a word error rate (WER) of 30.12% for the development set and 32.33% for the evaluation set, marking the best results reported to the authors' knowledge.

In 2020, [10] two Overlapping Speaker Identification (OSID) systems were proposed: a two-stage OSID system (T-OSID) and a single-stage OSID system (S-OSID). The two-stage OSID system first determines the number of simultaneous speakers and then identifies the speaker(s), while the single-stage system directly performs speaker identification using a single classifier, which is more computationally efficient. Experimental results showed that the two-stage OSID system outperformed the single-stage system in terms of identification accuracy. Additionally, both OSID systems based on one-dimensional convolutional neural networks (1DCNN) performed better than systems based on multilayer perceptron (MLP) and Gaussian mixture models (GMMs). The proposed 1DCNN-based two-stage OSID system achieved high accuracy of 98.55% for clean audio data with up to five simultaneous speakers. Even in challenging conditions involving background noises and high overlapping energy ratios, the system-maintained accuracies above 90%. In 2020, [11] different spectral features were evaluated, and it was found that pyknogram features outperformed other commonly used speech features, to investigate the detection of overlapping speech on short segments of 25 ms using Convolutional Neural Networks. The result achieved to 84% and an F-score of 88% in predicting overlapping speech on a dataset of mixed speech based on the GRID dataset.

In 2021, [12] The method used involved proposing a simple switching algorithm between observed and enhanced speech based on the estimated signal-to-interference ratio and signal-to-noise ratio. This switching mechanism was designed to improve recognition performance when processing artifacts from speech enhancement were

DOI: https://doi.org/10.54216/FPA.160208
Received: December 21, 2023 Revised: February 25, 2024 Accepted: June 02, 2024

detrimental to ASR. The proposed switching mechanism showed improved recognition performance compared to processing artifacts that were detrimental to ASR, indicating enhanced accuracy under specific conditions. In 2021, [13] proposed novel methods for multi-person speaker identification and diarization, enhancing the original model by modifying its normalization method and achieving state-of-the-art accuracy in speaker diarization. The approach used involves utilizing compositional embedding, which extend single-speaker embedding through a composition function to infer the set of speakers speaking within an input audio. In a speaker diarization experiment on the AMI Headset Mix corpus, the state-of-the-art accuracy achieved is a Diarization Error Rate (DER) of 22.93%, this result is slightly better than the previous best result of 23.82%, showcasing the effectiveness of the proposed compositional embedding models in handling overlapping speech from multiple speakers. The compositional embedding include a function to separate speech from different speakers and a composition function to infer the set of speakers within the input audio.

In 2022, [14] proposes a learning-based switching method that directly estimates whether ASR will benefit more from an enhanced or observed signal to address the performance degradation of automatic speech recognition (ASR) caused by processing artifacts generated by speech enhancement (SE) technologies, especially in overlapping speech scenarios. It introduces soft-switching, where a weighted sum of the enhanced and observed signals is used for ASR input based on the output posteriors of the switching model, leading to improved ASR performance and a relative reduction in character error rate of up to 23% compared to conventional methods.

In 2023, [15] proposed a unified modelling framework for multi-talker overlapped speech recognition and diarization tasks. This approach focuses on efficiently handling the challenges posed by multi-talker scenarios in speech techniques. The approach involves incorporating a diarization branch into a frozen, well-trained single-talker ASR model using a Sidecar separator. This method enables the unified modelling of ASR and diarization tasks with minimal additional parameters, leading to better performance in both tasks. The proposed framework leverages the inter-dependence between ASR and diarization tasks to achieve improved results in handling two- and three-speaker overlapped speech recognition tasks while maintaining a cost-effective solution.

3. Dataset

In this work we created a new dataset called GSCC "Ghadeer-Speech-Crowed-Corpus". The dataset was recorded by Iraqi Arab citizens living in different parts in Iraq, covering more than one Iraqi accent. The recording process took over the course of three months. 210 individuals participated in recording this dataset. The dataset proposed is fully balanced with respect to gender and recordings (the English recordings are equal to the Arabic recordings). The dataset is a mono dataset, and contains 15,626 audio samples recorded in 44100Hz sample rate, and 16-bit Bit depth, and the bit rate is 705.6kb/s. GSCC contains two types of recordings, solo and crowed, each stored in a separate folder. The crowed speech folder contains 4 subfolders, each subfolder for a different number of crowed. The crowed recordings, included 2, 3, 4, and 5 individuals. Each crowed subfolder, contain 2 subfolders, one for each language, one for Arabic and one for English. For each language we have 56 groups that speak 9 sentences. So, the total is 504 audios. The second folder is for solo speech that contains 210 folders, each folder represents a speaker. The speaker folder contains 2 folders, one for each language. Each speaker has 9 audios for each language.

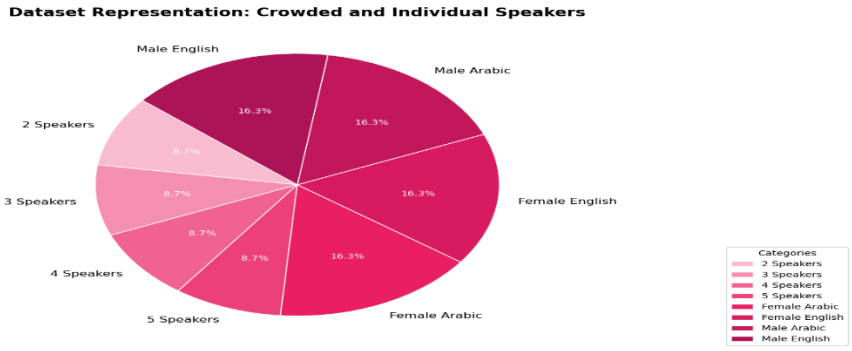

Figure 1. (GSCC) dataset description

4. Methodology (Combinatorial Cartesian Product Approach)

This section describes the structure of the crowd speaker identification using Combinatorial Cartesian Product Approach (CCPA).

DOI: https://doi.org/10.54216/FPA.160208
Received: December 21, 2023 Revised: February 25, 2024 Accepted: June 02, 2024

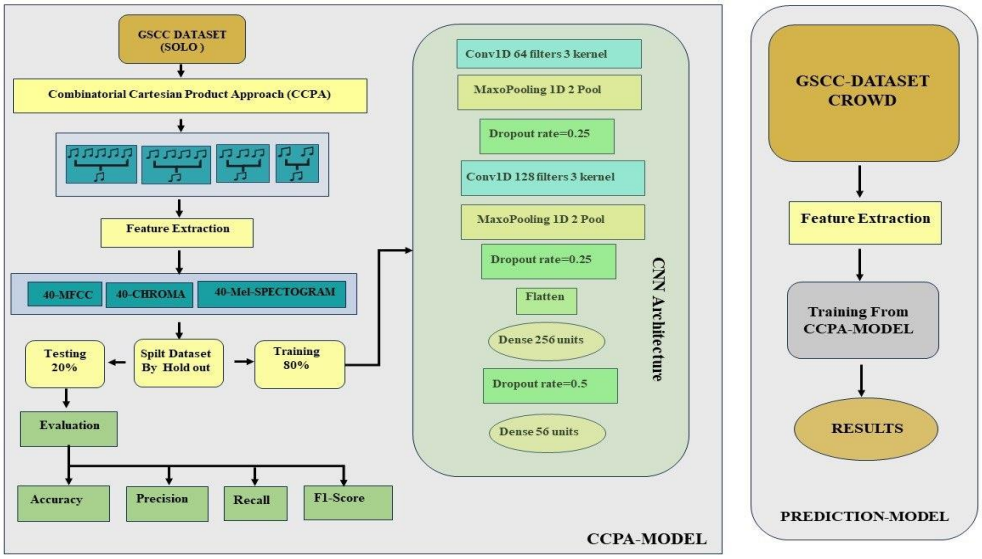

Figure 2. Crowed Speaker Identification structure.

4.1 Combinatorial Cartesian Product Approach (CCPA)

When combining objects from various subsets, there are a number of unique combinations that can be produced. This can be calculated methodically using the Combinatorial Cartesian Product Approach. This method is very helpful in situations where you have to combine several sets of objects to create new ones, like mixing audio tracks from different subfolders. Computation of the Cartesian product and combinatorial selection are the two primary processes in the method.

Step 1: Combinatorial Selection
The first step involves selecting subsets from a larger set. We start with a total of 210 subfolders, each containing audio files. Our objective is to choose multiple subfolders at once in order to blend audio files. The number of ways to choose subsets of subfolders from the total is calculated using the combination formula:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Where n is the total number of subfolders, and k is the number of subfolders to choose.

Step 2: Cartesian product of Files
Once the subfolders are selected, the next step is to identify every possible combination of audio files from these subfolders. There are nine audio files in each subfolder; we must select one file from each subfolder. The Cartesian product is used to get the total number of combinations for a set of chosen subfolders:

$$9^k$$

Where k is the number of subfolders selected.

Total Output Calculation
The final step is to calculate the total number of the audio files output, this done by multiplying the number of ways to choose the subfolders by the number of ways to choose the audio files within those subfolders:

$$\binom{n}{k} \times 9^k$$

Given 210 subfolders, we can calculate the total number of output files for different values of k:

1. Mixing 2 Subfolders at a Time:

$$\binom{210}{2} \times 9^2 = \frac{210!}{2!\,(210-2)!} \times 81 = \frac{210 \times 209}{2} \times 81 = 1{,}777{,}545$$

2. Mixing 3 Subfolders at a Time:

$$\binom{210}{3} \times 9^3 = \frac{210!}{3!\,(210-3)!} \times 729 = \frac{210 \times 209 \times 208}{6} \times 729 = 1{,}080{,}266{,}440$$

DOI: https://doi.org/10.54216/FPA.160208
Received: December 21, 2023 Revised: February 25, 2024 Accepted: June 02, 2024

3. Mixing 4 Subfolders at a Time:

$$\binom{210}{4} \times 9^4 = \frac{210!}{4!\,(210-4)!} \times 6561 = \frac{210 \times 209 \times 208 \times 207}{24} \times 6561 = 127{,}596{,}522{,}510$$

4. Mixing 5 Subfolders at a Time:

$$\binom{210}{5} \times 9^5 = \frac{210!}{5!\,(210-5)!} \times 59049 = \frac{210 \times 209 \times 208 \times 207 \times 206}{120} \times 59049 = 9{,}525{,}040{,}727{,}940$$

The Cartesian product and combinatorial selection are combined in the Combinatorial Cartesian Product Approach to quickly determine the total number of possible combinations. When there are several groups of objects and choosing one item from each group is required for each combination, this method is especially helpful. By calculating the total number of mixed audio files that can be created by combining files from various subfolders, the calculations above demonstrate how this approach can be applied to different numbers of subfolder selections, accurately determining the total number of mixed audio files that can be produced by mixing files from different subfolders.

4.2 Features Extraction

MFCCs are an extremely effective instrument, When it comes to compactly encoding the main qualities of audio signals in a manner that is consistent with human auditory perception, Because of this, they are essential in a wide range of audio processing and recognition applications [4] [16]
In this study, we used three audio file features to train a CNN to identify speakers. Frequency domain features are retrieved using the Librosa library, which includes:

1. Mel-frequency Cepstral Coefficients (MFCCs)

MFCCs aim to mimic the human auditory system. The Mel scale used in MFCCs is a perceptual scale that is closely related to the response of the human ear from linearly spaced frequency bands. This makes MFCCs very good at capturing the complexities of human speech. Using MFCCs, speech processing systems can effectively capture and analyze important aspects of the human voice, making them a powerful tool for many signal processing applications.

2. Chroma Features

each speaker has a unique pattern of pitch and intonation that can be captured by chroma features. The chroma captured harmonic that have relates to the pitch and intonation. Chroma features are combined with other features, such as MFCCs, to provide a complete feature set for each speaker.

3. Mel-spectrogram

A Mel spectrogram are detailed representations of how the energy of an audio signal is distributed across different frequencies over time. This level of accuracy helps capture the distinct spectral features of a speaker's voice.

Extract these three features allows the model of CNN to captured the most important features from the audio speech. This feature increased the robustness of identification system and the accuracy was very high even when the audio signal was complex.

4.3 Model Architecture and Training:

4.3.1 Model Architecture:

The architecture of the CNN is specifically designed to handle the complexities of audio data:

- Convolutional Layers: The model starts with two convolutional layers. The first layer includes 64 filters and the second layer has 128 filters, each with a kernel size of 3. These layers use convolution to extract local temporal information from audio data, capturing fine details of the speaker's speech.

- RelU (rectified linear unit) activation function: is used in each convolutional layer, introducing nonlinearity and facilitating learning of complex patterns.

DOI: https://doi.org/10.54216/FPA.160208
Received: December 21, 2023 Revised: February 25, 2024 Accepted: June 02, 2024

- After each convolutional layer, max pooling layers with a pool size of 2 are applied. These layers reduce feature maps to reduce computational complexity and focus on essential features.

- Dropout Layers: Max-pooling and Dense layers are followed by dropout layers at 0.25 and 0.5. Dropout is a regularization approach that ensures the model performs effectively when generalizing to new data by randomly changing a portion of the input units to zero during training. This prevents overfitting.

- Flatten Layer: This layer is suitable for the dense layers since it transforms the 2D feature maps into a 1D feature vector.

- Dense Layers: The model consists of two dense layers: one fully connected layer with 256 units and a ReLU activation function, and another dense layer with a Softmax activation function and units equal to the number of speaker classes. The Softmax activation function assigns the input to one of the speaker's classes by generating a probability distribution across the classes.

4.3.2 Training Process:

The training process involves several critical steps:

- Data Preparation: The dataset is divided into training and testing sets using an 80:20 ratio. The testing set is used to evaluation the model's performance on unseen data, whereas the training set is used to fit the model.

- Input Reshaping: To make the input data work with the CNN, it is reshaped. To match the predicted input form of the convolutional layers, each audio feature vector is modified to have a shape of (number of features, 1).

- Label Encoding and One-Hot Encoding: To make the speaker labels compatible with the category cross-entropy loss function, they are first encoded using a label encoder and then transformed into a one-hot encoded format.

- Model Compilation: The Adam optimizer, an adaptive learning rate optimization technique well-suited for deep neural network training, is used to assemble the model. For multi-class classification tasks, categorical cross-entropy is the appropriate loss function. Accuracy is the main metric used to assess the model as well.
- Training: The model is trained for 300 epochs using batch sizes of 32 for two speakers, 64 for three to four speakers, and 128 for five speakers. The model uses gradient descent and backpropagation to learn how to minimize loss function. During training. Validation data is used during training to assess model performance on unobserved data.

5. Results

The Convolutional Neural Network (CNN) model achieve a high performance to identify a speaker in crowed speech environment. In the beginning, we trained the CNN model using only 13 features of (MFCCs) but the result was very bad to our system to identify speaker from other. So, to solve this problem we add more features chroma and Mel-spectograms. This addition of feature led to improvement the accuracy of result, in addition to added more feature we expand the number of all feature are used to 40 this step led to amazing results.
For a crowed speech audio contain 2 speakers the model achieved a high accuracy reached 99.9 with batch size 32. this high accuracy to prove the ability of the model to identify the speaker even when overlapping their voice when tested with three speakers mixed the model keep the high accuracy 99.5 with 64 batch size the accuracy slightly decreased to 98.5 when four speakers tested also with batch size 64 and this refer to increase the complexity to identify speakers finally 'the model achieved 96.4% with five speakers and 128 batch size. the accuracy still very high despite the complexity to identify the speakers as showed in figure 3. After applying the Convolutional Neural Network (CNN) to the Audio generating by CCPA, the results were highly impressive when apply it on the real data within the dataset used as show in Figure 2 in prediction model. The model demonstrated significant accuracy and performance improvements, indicating its robustness and efficacy in handling different types of data. This success further validates the effectiveness of using CNNs for complex data analysis tasks.

This model has proven its robustness and strength also through its high precision, recall, and f1 scores. The high precision refers that the model identified true positive effectively and without effects by false positive. High recall

DOI: https://doi.org/10.54216/FPA.160208
Received: December 21, 2023 Revised: February 25, 2024 Accepted: June 02, 2024

indicates that the model accurately detects most instances of each speaker. The model achieves a high F1-score, which measures precision and recall, indicating balanced performance in accurately detecting speakers and avoiding missing the result show in the table 1.

Table 1: The result of the model.

NO.Speaker	Accuracy	F1score	Recall	Precision
2	99.9%	0.991	0.991	0.995
3	99.5%	0.992	0.992	0.992
4	99.5%	0.977	0.982	0.975
5	99.4%	0.962	0.962	0.962

The exceptional performance of the model has major practical implications. Suggest such a system can be effectively deployed in demanding real-world applications. Accurate identification of the speaker, such as security systems, transcription services, and Communications platforms. These applications benefit greatly from the ability to be precise Distinguish between multiple speakers in a single audio recording, enhancing functionality and user experience. The impressive accuracy of the CNN model is 99.9% for two speakers, and 99.5% for three speakers, 98.2% for four speakers, and 96.4% for five speakers' effectiveness and durability. Comprehensive feature extraction, well designed the model's structure and rigorous evaluation metrics contribute to this success. Highlighting the potential of deep learning in complex sound processing and collocation tasks the route is for practical applications that require precise speaker identification.

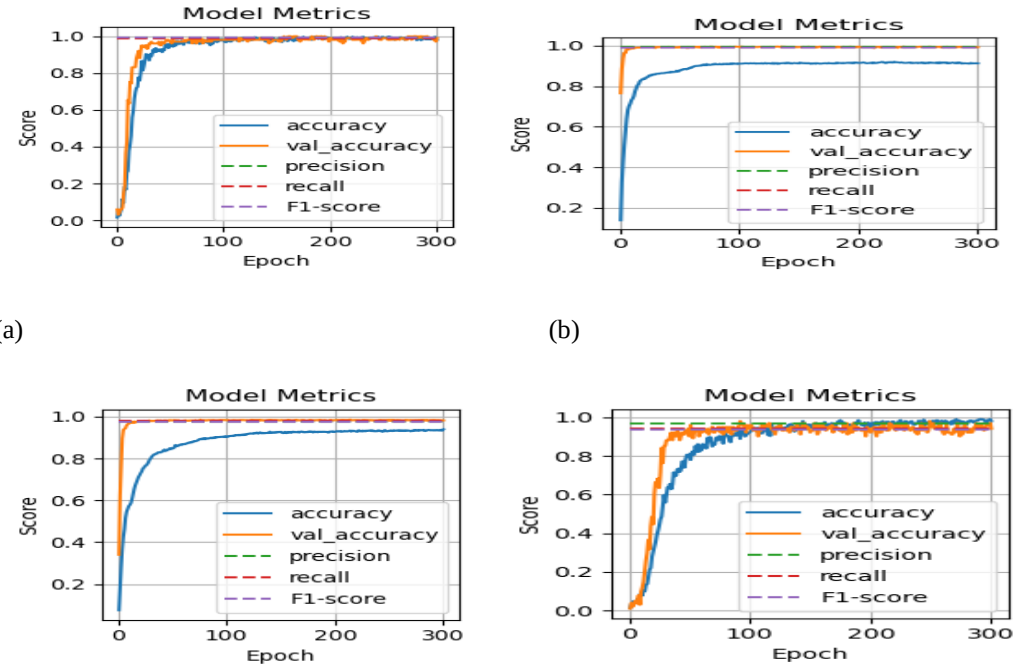

Figure 3. (a) result of 2 mixing speaker, (b) result of 3 mixing speaker, (c) result of 4 mixing speaker, (d) result of 5 mixing speaker.

6. Conclusion

Identifying speaker through their voices is a difficult process, especially if there is overlapping of several speakers' voices, but with the advent of deep learning, it has become possible to achieve this goal. Our work differed from other works in terms of the results it achieved and the number of people whose voices overlapped in one audio clip. The proposed method was taking all possible probability of mixing speaker voices by using the Combinatorial

DOI: https://doi.org/10.54216/FPA.160208
Received: December 21, 2023 Revised: February 25, 2024 Accepted: June 02, 2024

Cartesian Product Approach. After taking all probability extract features which are represented by MFCC, Chroma, and Mel-spectrogram. Using the convolutional neural networks (CNN) to train our model with a complex architecture led to high accuracy 99.9% for identify speaker in audio contain 2 speakers, 99.5% for audio with three speakers, 98.5% for mixing with four speakers, and 97% for audio with five speakers.

References

[1] A. Shakat, K. I. Arif, S. Hasan, Y. Dawood, and M. A. Mohammed, "YouTube keyword search engine using speech recognition," *Iraqi J. Sci.*, vol. 2021, pp. 167–173, 2021, doi: 10.24996/ijs.2021.SI.1.23.

[2] Samyuktha, S. Kavitha, D. Kshaya, V. Shalini, P. Ramya, R. "A Survey on Cyber Security Meets Artificial Intelligence: AI– Driven Cyber Security," Journal of Journal of Cognitive Human-Computer Interaction, vol. 2, no. 2, pp. 50-55, 2022. **DOI:** https://doi.org/10.54216/JCHCI.020202

[3] anthi, V. Kumar, A. "Enhancing Healthcare Monitoring through the Integration of IoT Networks and Machine Learning," Journal of International Journal of Wireless and Ad Hoc Communication, vol. 7, no. 1, pp. 28-39, 2023. **DOI:** https://doi.org/10.54216/IJWAC.070103

[4] H. A. Abdulmohsin, B. Al-Khateeb, S. S. Hasan, and R. Dwivedi, "Automatic illness prediction system through speech," *Comput. Electr. Eng.*, vol. 102, p. 108224, 2022.

[5] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Inf. Fusion*, p. 101869, 2023.

[6] A. Tripathi, H. Lu, and H. Sak, "End-to-end multi-talker overlapping speech recognition." Google Patents, Dec. 06, 2022.

[7] N. M. Jakovljevic, T. V Delic, S. V Etinski, D. M. Miskovic, and T. G. Loncar-Turukalo, "A multi-target speaker detection and identification system based on combination of plda and dnn," in *2018 26th Telecommunications Forum (TELFOR)*, IEEE, 2018, pp. 1–4.

[8] M. Thakker, S. Vyas, P. Ved, and S. Shanthi Therese, "Speaker identification in a multi-speaker environment," in *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2016, Volume 2*, Springer, 2018, pp. 239–244.

[9] N. Kanda, Y. Fujita, S. Horiguchi, R. Ikeshita, K. Nagamatsu, and S. Watanabe, "Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6630–6634.

[10] V.-T. Tran and W.-H. Tsai, "Speaker identification in multi-talker overlapping speech using neural networks," *IEEE Access*, vol. 8, pp. 134868–134879, 2020.

[11] M. Yousefi and J. H. L. Hansen, "Frame-based overlapping speech detection using convolutional neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6744–6748.

[12] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, T. Moriya, and N. Kamo, "Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition," *arXiv Prepr. arXiv2106.00949*, 2021.

[13] Z. Li and J. Whitehill, "Compositional embedding models for speaker identification and diarization with simultaneous speech from 2+ speakers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7163–7167.

[14] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, N. Kamo, and T. Moriya, "Learning to Enhance or Not: Neural Network-Based Switching of Enhanced and Observed Signals for Overlapping Speech Recognition," *arXiv preprint arXiv: 2201.03881*, 2022. [Online]. Available: https://arxiv.org/abs/2201.03881

[15] L. Meng, J. Kang, M. Cui, H. Wu, X. Wu, and H. Meng, "Unified Modeling of Multi-Talker Overlapped Speech Recognition and Diarization with a Sidecar Separator," *arXiv Prepr. arXiv2305.16263*, 2023.

[16] H. A. Abdulmohsin, "Automatic Health Speech Prediction System Using Support Vector Machine," in *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021*, Springer, 2022, pp. 165–175.

DOI: https://doi.org/10.54216/FPA.160208
Received: December 21, 2023 Revised: February 25, 2024 Accepted: June 02, 2024