Journal of Electrical and Electronic Engineering
2024, Vol. 12, No. 4, pp. 71-83
https://doi.org/10.11648/j.jeee.20241204.12

Research Article

UNNIGSA: A Unified Neural Network Approach for Enhanced Stutter Detection and Gait Recognition Analysis

Ravikiran Reddappa Reddy*, Santhosh Kumar Gangadharaih

Electronics and Communication Engineering, East West College of Engineering, Bangalore, India

Abstract

Stuttering, also known as stammering, is a speech disorder characterized by involuntary disruptions or disfluencies in a person's flow of speech. These disfluencies may include repetitions of sounds, syllables, or words; prolongations of sounds; and interruptions in speech known as blocks. This paper introduces Unified Neural Network for Integrated Gait and Speech Analysis (UNNIGSA), methodology that synergizes stutter detection (SD) and gait recognition through a unified neural network architecture. UNNIGSA is engineered to address two distinct yet interrelated challenges: the accurate detection of stuttering for enhanced beneficial interventions and the precise identification of individuals based on gait analysis. The system integrates a global attention mechanism to meticulously highlight salient features within speech patterns, thereby improving the accuracy of stutter classification and offering a potential leap forward in speech therapy practices. Additionally, UNNIGSA incorporates novel data processing techniques to manage the class imbalance prevalent in stuttering speech datasets, resulting in significantly enhanced performance over existing models. The methodology also extends the functionality of automatic speech recognition (ASR) systems, fostering greater inclusivity for individuals with speech disorders and enabling their more seamless interaction with virtual assistant technologies. Overall, UNNIGSA sets a new standard in the domains of speech disorder treatment and biometric identification, offering innovative solutions to long-standing challenges and paving the way for more inclusive and secure applications.

Keywords

UNNIGSA- Unified Neural Network for Integrated Gait and Speech Analysis, ASR- Automatic Speech Recognition, SD-Stutter Detection, PWS-People Who Stutter, SLP-Speech-Language Pathologists, ST-Speech therapists

1. Introduction

Speech disorders, also referred to as speech impairments, encompass the incapacity of an individual to generate the customary speech sounds necessary for proficient interpersonal communication [1]. Lisping, cluttering, stuttering, dysarthria, apraxia, and others are symptoms of these diseases. Stuttering, referred to as stammering or disfluency is a speech disorder that manifests through the occurrence of involuntary pauses, repetitions, and prolongations of sounds, syllables, sentences, or phrases. Stuttering is a neuro-developmental speech disorder that occurs due to the rapid development of brain connections involved in language, speech, and affective activities by the presence of extended interruptions in the typical progression of speech. In addition to experiencing difficulties with speech, individuals frequently display atyp-

*Corresponding author: ravikiran1182.ece@sjcit.ac.in (Ravikiran Reddappa Reddy)

Received: 20 July 2024; **Accepted:** 2 September 2024; **Published:** 26 September 2024

ical behaviors, including head swaying, lip tremors, rapid eye blinks, and distinctive lip shapes. Fluency is defined as the capacity to communicate in a seamless and effortless manner, while maintaining a consistent pace [2]. Fluency in conversation requires a combination of expertise in the topic being discussed and proficiency in the linguistic skills associated with the spoken language. The precise coordination of the supraglottis, respiratory system, and larynx is essential for maintaining fluency from a physiological perspective [3]. Speech difficulties, such as stuttering, may arise if any of the specified conditions are not met. Communication disruptions can manifest in various ways, such as the occurrence of repetitions, interjections, periods of silence or complete pauses, incomplete phrases, and revisions.

The manifestations can occur in multiple ways, the process of communication typically comprises of predominantly fluent components, which are occasionally interspersed with disfluent elements. Normal disfluencies have the function of enabling speakers to prepare or make corrections in real time, thereby improving the quality of their speech output. Disfluencies, such as stammering, frequently present difficulties for speakers in effectively organizing their words [4]. Individuals who encounter fluency challenges, commonly referred to as People Who Stutter (PWS), encounter a transient inability to proficiently convey their intended message, in contrast to individuals without such impairments. Stuttering can be classified into two broad categories.

Stuttering is a prevalent speech disorder that frequently manifests during the developmental phase. Stuttering is commonly observed as the most prevalent speech disfluency during the period of ongoing language acquisition, typically occurring between the ages of two and seven. Based on recent research findings, it has been determined that developmental stuttering is a complex matter influenced by a variety of genetic and neurological factors [5].

Neurogenic stuttering is a condition characterized by the manifestation of stuttering due to different forms of brain injuries, including those induced by strokes. The occurrence of incoherent speech can be attributed to the dysfunction present in various brain regions responsible for the speech production process [6].

The stuttering phenomenon is known for its intricate and mysterious characteristics. Stuttering is a speech disorder that can be affected by multiple factors, such as stress, delayed early development, and abnormalities in speech motor control. A significant correlation has been observed between stress, anxiety, and the occurrence of stuttering. Disfluencies are frequently observed in individuals who encounter stress or anxiety, speak at a fast pace, or engage in simultaneous dual cognitive tasks. In contrast, individuals diagnosed with Parkinson's disease who participate in synchronized singing or employ modified auditory feedback during speech demonstrate a decreased frequency of speech interruptions. Stuttering is believed to be caused by the central nervous system's inability to generate the necessary motor instruction patterns for fluent speech, as suggested by Smith and Weber in their recent study [1] on the multifactorial dynamic pathways theory.

Stuttering is a manifestation of impaired sensorimotor mechanisms involved in the production of speech. The linguistic and emotional factors have a significant impact on individuals with Parkinson's disease throughout their lifespan. During a typical stuttering assessment, speech-language pathologists (SLPs) or speech therapists (STs) perform a manual evaluation of the individual who stutters (PWS)'s speech or analyze recorded samples. The evaluation of stuttering severity generally focuses on measuring the proportion of disfluent words or duration in relation to the total number of words or duration [7]. In the context of standard speech therapy sessions, individuals with stuttering tendencies (PWS) are instructed to actively observe and monitor their own speech in order to effectively modify their speech patterns [7]. According to research findings, it has been established that early intervention can result in recovery rates ranging from 60% to 80% following speech treatment. The assessment of stuttering severity and the measurement of therapy outcomes is a complex and resource-intensive undertaking. Additionally, this process is often subjected to the influence of subjective perspectives held by speech-language pathologists. Individuals with a stutter, commonly referred to as People Who Stutter (PWS), frequently necessitate routine appointments with a Speech-Language Pathologist (SLP) as a result of the characteristics of their condition.

Stuttering therapy sessions are characterized by their rigorous nature and extended duration, often lasting several months or even years [8]. Individuals who experience stuttering, also known as people who stutter (PWS), often encounter financial obstacles when seeking speech therapy services. This is primarily due to the considerable expenses associated with these sessions, as well as the need for confidentiality. Consequently, they are unable to procure the essential treatment. Hence, the development of interactive automated stuttering detection systems (SD) is of utmost importance. The smooth pronunciation of speech is crucial for ensuring the efficient operation of automatic speech recognition (ASR) systems. Nevertheless, these systems exhibit limitations in accurately identifying speech that is impacted by stuttering. As a result, individuals with speech impairments may face challenges when attempting to utilize virtual assistant applications like Apple Siri, Alexa, and other comparable platforms [9]. The SD card is a valuable resource for individuals with limited financial means, as it enables modifications to be made to the functionality of the virtual assistant. To achieve consistent and objective measurement of stuttering on an hourly basis, it is essential to employ automated stuttering identification systems (ASIS). The development of intelligent and interactive self-diagnostic (SD) and therapeutic tools has been made possible by recent advancements in natural language processing, machine learning, and deep learning. Despite having multiple applications, ASIS has not garnered

substantial attention in this field.

Motivation and Contribution

The motivation behind developing a stuttering detection and gait recognition systems lies in addressing the profound communication barriers faced by individuals with speech disorders and enhancing biometric security measures. For people who stutter, overcoming the challenges of social interaction and self-expression is vital for personal and professional growth. Automated systems promise a more objective and accessible assessment of speech fluency, potentially revolutionizing therapy outcomes. In terms of gait recognition accurately identifying individuals across varying conditions, ensuring robust and reliable security systems. This research is focused by the pursuit of technological inclusivity, aiming to create tools that enable seamless communication and secure identification for all.

1) This paper proposes a UNNIGSA methodology which further introduces an innovative neural network architecture for processing both spatial and temporal data, enhancing stutter detection accuracy and gait recognition across varying angles.
2) The system incorporates a global attention mechanism, significantly refining feature selection for stutter classification, leading to improved speech therapy outcomes.
3) It presents a robust solution for the inclusion of individuals with speech disorders in using virtual assistant technologies, advancing the inclusivity of automatic speech recognition systems.
4) UNNIGSA demonstrates the effective handling of class imbalance in speech data, offering a substantial improvement in stuttering identification performance over traditional models.

2. Related Work

Stuttering is a speech disability that arises from neuro-developmental factors. It is characterized by the presence of uncontrollable utterances, known as interjections, as well as core behaviors such as blocks, repeats, and prolongations. Stuttering is primarily attributed to a failure in speech sensorimotor function. The detection of stuttering (SD) presents a considerable challenge due to its intricate characteristics. The prompt and accurate recognition of speech patterns in individuals with stuttering, also known as people who stutter (PWS), can support speech therapists in efficiently monitoring and addressing their speech issues. Individuals with Prader-Willi Syndrome often display speech disfluency, which is characterized by stuttering [10]. Additionally, they may experience notable instability in small quantities. In order to tackle the problem of class imbalance in the SD domain, a multi-branching (MB) approach is employed. This approach involves assigning weights to the classes within the overall loss function. This functionality allows for the efficient resolution of the current issue. The application of this methodology leads to a significant improvement in the classification performance of stuttering on the SEP-28 k dataset in comparison to the baseline model, StutterNet.

According to reference the research of H. Geng et al., wearable sensors are employed to monitor daily activities and lifestyle preferences [11]. Accurate segmentation of recorded activities is achieved by leveraging machine learning algorithms that are built for this task. One crucial aspect of the process focuses on the establishment of priorities for the identification and analysis of vital tasks. In the reference Z. Wang, S. Hou, M. Zhang, X. Liu, C. Cao and Y. Huang, the design of the segmentation system allows it to effectively recognize and distinguish six main actions: walking, climbing stairs, descending stairs, sitting, standing, and lying down. The aforementioned operations are accurately identified and differentiated from a continuous signal. The first step in the procedure focuses on the partitioning of continuous signals. The next step in the process focuses on the identification of transitional activities. The Greedy Gaussian Segmentation (GGS) approach was utilized by the researchers to detect discontinuities that arise during action transitions. The XGBoost model was employed to forecast the human activity for every frame. The identification of transitional acts was not addressed by the writers [12].

The study's authors have presented a new technique for the segmentation and categorization of continuous sequences of actions. The sequence can be classified into three primary categories: transitional activities, static acts, and dynamic actions. This work presents a novel algorithm for the detection of gait cycles using biometric data [13]. The main goal of this method is to precisely identify and classify every dynamic activity that occurs in a predetermined sequence. This was facilitated through the analysis of the characteristics of the gait signal during dynamic activities. The main goal of the analysis was to evaluate the distribution of fluctuation points in two different types of activities, while considering the difference in signal conversion frequency between transitional and static actions [14].

The GaitParsing framework was created as a specialized approach for the analysis of human semantic parsing and gait recognition. The primary objective is to perform a comprehensive analysis of the human body and classify it into distinct anatomical components with well-defined and comprehensive structures. In order to meet this requirement, a dual-branch feature extraction network is utilized. The network has been intentionally designed to effectively manage both individual body parts and whole-body gait in an efficient manner. The application of the self-occlusion frame evaluation technique offers a practical method for evaluating self-occlusion in a gait sequence. The primary goal of this assessment is to optimize the efficiency of gait frames that demonstrate substantial variations [15].

In the design and manufacture of a biomedical device intended for regular patient use is designed. The primary objective of this device is to identify occurrences of stuttering and capture pertinent information during regular conversa-

tions. The recorded conversations will subsequently undergo review by speech specialists at a later point in time. The biomedical discovery holds promise for aiding caregivers and medical professionals in supporting individuals with stuttering, helping them overcome this behavior and enabling them to actively engage in society. This article explores the potential applications and advancements of the device, as well as the feasibility study conducted and the prototype developed. The objective of this biomedical advancement is to offer insights into the different criteria associated with stuttering that necessitate evaluation. The assessment enabled by this innovative development is expected to improve the effectiveness of therapy provided by healthcare professionals [16].

The objective of this study is to explore a computer-based methodology for identifying and examining repetitive patterns of stuttering in the Chinese language. Proposed improvement options are derived from research findings pertaining to the characteristics of repetitions observed in Chinese stuttering speech. The initial stage of the process focuses on the utilization of multi-span looping forced alignment decoding networks for the detection of recurring syllables in Chinese speech that demonstrate stammering. Furthermore, the decoding networks are enhanced by integrating a branch penalty factor, which serves to alleviate errors that may arise due to the intricate nature of these networks. The purpose of this factor is to modify the decoding trajectory by means of a recursive search. In order to improve the reliability of the detection results, we have conducted a reassessment of the stutters that were initially identified using confidence computation [17].

Individuals with Prader-Willi Syndrome often display speech disfluency, which is characterized by the presence of stuttering. Additionally, they may encounter notable instability in small quantities. In order to mitigate the problem of class imbalance in the SD domain, a multi-branching (MB) approach is employed. This approach involves assigning weights to the classes in the overall loss function. This functionality allows for the efficient resolution of the current issue. The application of this methodology leads to a significant improvement in the classification performance of stuttering on the SEP-28k dataset, in comparison to the StutterNet model used as a reference. This study aims to assess the efficacy of integrating data augmentation into a multi-branched training scheme as a potential remedy for the scarcity of available data [18].

The primary aim of this study is to establish a foundation for a Brain-Computer Interface (BCI) system that can assist individuals with Profound and Multiple Learning Disabilities (PMLD) in identifying and assessing their primary emotional states. The objective emotions of individuals with PWS were captured in this study using an EEG-based mind wave sensor. The emotions of the intended audience were subsequently determined by analyzing the electroencephalogram (EEG) data associated with emotions. The results indicate that EEG-based sensors offer a viable and economical approach for monitoring the brain activity of the target population. Emotional challenges have been observed in individuals with Prader-Willi Syndrome based on analysis of their electroencephalogram (EEG) data [19].

3. Proposed Methodology

The proposed UNNIGSA (Unified Neural Network for Integrated Gait and Speech Analysis) network takes in a time-sequenced input, which could represent frames of motion capture data. TF for the number of time frames, C for the number of connections, and for the ch_{inp}number of input channels, $TF* C*ch_{inp}$. This input is further passed on to the network which converts the input to spectrogram to handle problems associated with speech processing. These inputs are further passed to a SE-Residual Network to analyze the frame level understanding. As each stuttering type has different spatio temporal properties present within the spectrograms. Finally, an attention mechanism is used at the final layer which focuses on the features responsible for stutter classification.

Architecture

In this section, the design of a novel neural network architecture is used for processing tasks that necessitate a dual focus on spatial and temporal data particularly for motion capture data. This architecture is characterized by its multifaceted approach to data processing, encompassing a range of operations from adaptive temporal pooling to loss function application.

Input Data Processing: The network architecture is initiated by accepting time-sequenced input, potentially representing frames of motion capture data. This input, designated as TF for time frames, C for the number of connections, and ch_{inp} for number of input channels that undergoes an initial adaptive temporal pooling. This step is vital in reducing temporal resolution while preserving critical information, thus enabling the effective capture of temporal features.

Temporal Filtering Operations: Subsequently, the architecture bifurcates into two distinct processing paths, each to handle specific aspects of temporal data: The left branch of the network encompasses a series of pooling and convolutional layers, terminating in fully connected layers. This configuration is based at extracting and refining temporal features, further enhanced by feature scaling and reshaping operations. Conversely, the right branch, diverges in its application, focusing on feature scaling using temporal filters to capture diverse scales of temporal dynamics.

$$fil_u = ⍰(⍰\left(tem_{pool\left(tem_{conv(z_{output})}\right)}, Y_3\right), Y_4) \quad (1)$$

fil_u represents the filter with u size, tem_{pool} denotes temporal pooling and tem_{conv} denotes temporal convolution,

Y_3 reduces the channel dimension and Y_4 recovers the channel dimension.

Feature Aggregation and Topology Learning: The convergence of these two paths leads to a softmax function, indicative of a classification or decision-making stage based on the extracted features. Concurrently, the network undertakes a process of feature aggregation coupled with topology learning. This involves the integration of both fixed and learned topologies, incorporating the graph neural network components. Such a configuration is key in refining the features or decision-making process by leveraging multiple relational structures within the data. To generate a topology, the maximum index value formulated is denoted by:

$$in_{Top} = argmax\, u_{Top} \tag{2}$$

in_{Top} depicts the topological view and the adjacent topology is chosen by Top_{set} from in_{Top} is shown as below:

$$Top_1 = Top_{set}[in_{Top}] \tag{3}$$

Top_1 reflects properties of the corresponding which is however not sufficient to denote all variations henceforth to extend the characteristics, by considering the data linear weights by combining the topologies in the Top_{set}, this is determined as

$$Top_2 = \sum_{i=1}^{num_v} u_{top}^i\, Top_v^i \tag{4}$$

num_v denotes number of views, u_{top}^i denotes the normalized value and Top_v^i denotes the learnable topology. Top_3 is a fixed topology like the regular graph which extracts general features. In the end Top_1 Top_2 and Top_3 is combined with variables t_1, t_2 and t_3 to obtain Top_{adap_top}:

$$Top_{adap\_top} = t_1 Top_1 + t_2 Top_2 + t_3 Top_3 \tag{5}$$

Top_{adap_top} is employed to dynamically link body joints in graph convolutions, which incorporates joint information over a long range in addition to correlating joints in the immediate neighborhood.

Output Processing and Loss Functions: The network's output processing is characterized by the generation of a $TopList$ known as topology list, likely an array of topological structures or feature sets derived from the input data. This process employs a softmax function, possibly for normalization or probabilistic classification purposes. In parallel, a loss function denoted as N_{view} potentially addresses view-specific features or learning aspects in a multi-view learning scenario. In the final stages, the network utilizes loss functions, namely triplet loss and circle loss. These functions play a pivotal role in the training phase, optimizing the network to develop robust feature representations suitable for classification or clustering tasks. Final Output is Z_{output}, representing the network's final output. This output may constitute the classification result consists of a set of encoded features, ready for subsequent processing or decision-making tasks.

$$Z_{output} = \mu_{1\,N_{tri}} + \mu_{1\,N_{circle}} + \mu_{1\,N_{DE}^{view}} \tag{6}$$

This proposed neural network architecture presents a comprehensive and multifaceted approach to handling complex temporal dynamics, integrating a variety of advanced techniques from feature extraction to graph-based learning, ultimately leading to a robust system capable of spatial and temporal data analysis.

Table 1. *Algorithm for stutter detection.*

Input	TF for the number of time frames, C for the number of connections, and for the ch_{inp} number of input channels, $Input$= TF* C*ch_{inp}
Step 1	Pre-processing the input: *For each Convert the input into an STFT spectrogram within a specific frame length*
Step 2	Initialize the SE-Residual model: *Set up SE-Residual Network blocks + convolution+ batch normalization along with ReLU activation layers* *Include SE-Residual network blocks and the residual connections.*
Step 3	Process each spectrogram through SE-Residual model *For each spectrogram:* *Feed through the SE-Residual network to learn frame-level spectral representations*
Step 4	Initialize a Z network to learn temporal relationships: *two consecutive layers with hidden units for each is added* *Apply a dropout to each layer to prevent overfitting*
Step 5	Pass the frame-level representations through the network: *For each set of frame-level features from SE-Residual Network*

Input is fed through the Z network to capture temporal relationships

Apply a global attention mechanism to the network outputs:

Step 6 *For the output of the second Z layer:*

Utilise a global attention mechanism to focus on salient features for stutter classification

Generate the final classification output:

Step 7 *Combine the attention mechanism output with the Z output using a tanh activation function to produce the binary classification result*

Implement the UNNIGSA model in a deep learning framework:

Step 8 *Construct the network using layers defined in steps 2 to 7*

Compile the model with $RMSProp$ optimizer and a binary cross-entropy loss function

Training with audio spectrograms:

For a specified number of epochs:

Step 9 *For each batch of spectrograms:*

Train the model on the batch

If the loss falls below a predefined threshold, perform early stopping

Validate the proposed model:

For each input in the validation set:

Step 10 *Predict the presence of stutter using the trained model*

Evaluate the model performance based on classification accuracy

End

output *Binary classification of stutter presence: N_{tri} and N_{circle}*

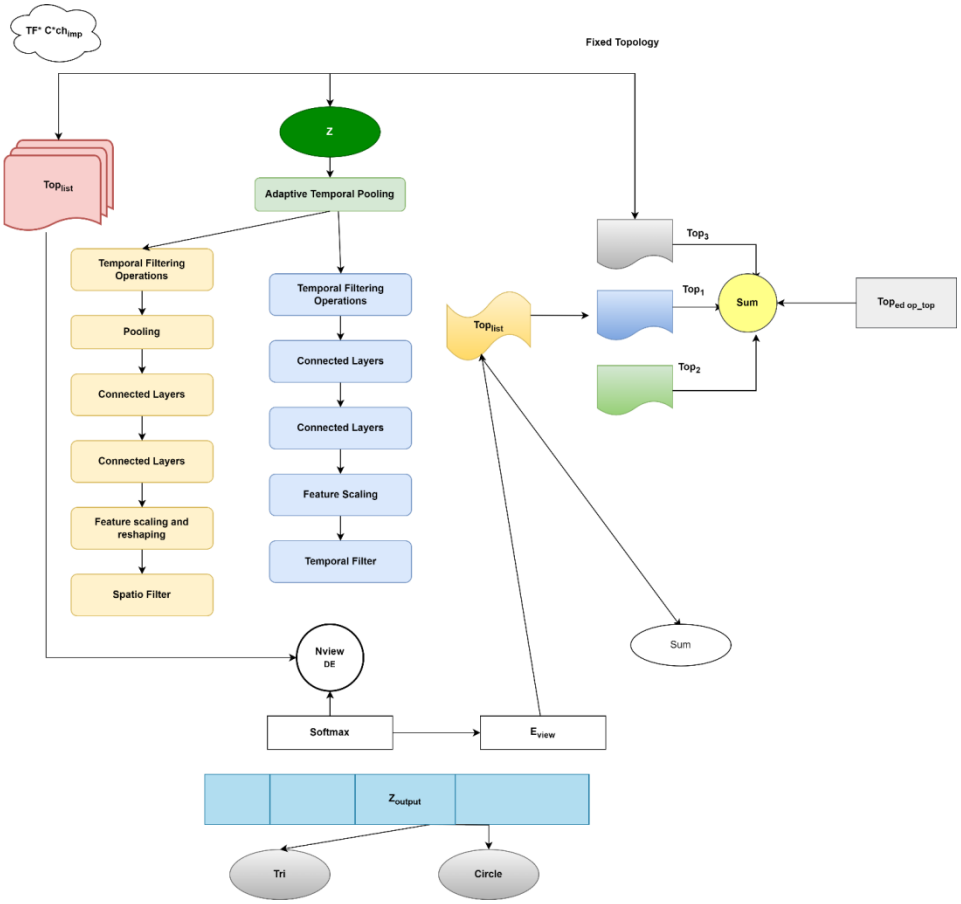

Figure 1. *Proposed architecture.*

The algorithm describes the process for implementing and training an end-to-end neural network, designed for the binary classification of stutter presence in the input. The network utilizes a combination of advanced deep learning techniques, including SE-Residual Networks and Z networks, augmented with a global attention mechanism. Initially, the algorithm processes input, converting them into Short-Time Fourier Transform (STFT) spectrograms within a specified frame length. This transformation is crucial for capturing both the spectral and temporal aspects of the audio data. The SE-Residual Network, comprising multiple blocks with convolution layers, batch normalization, and ReLU activation, is then initialized. These blocks are equipped with SE units and residual connections to effectively learn frame-level spectral representations from each spectrogram. To capture the temporal relationships in the data, the algorithm employs a Z network,, with two consecutive layers and a dropout mechanism to prevent overfitting. The frame-level features extracted by the SE-Residual Network are fed into this Z network.

The algorithm further enhances its capability by applying a global attention mechanism to the outputs of the Z network's final layer. This mechanism focuses on the most relevant features for stutter classification, thereby improving the model's accuracy. The final classification output is generated by combining the outputs from the attention mechanism and the Z network using a tanh activation function. This output is intended for binary classification, distinguishing between the presence and absence of stuttering. Z is then implemented within a deep learning framework, compiled with an RMSProp optimizer and binary cross-entropy loss function. The model undergoes training with audio spectrograms over a specified number of epochs, incorporating an early stopping mechanism if the loss falls below a predetermined threshold. Finally, the trained Z model is validated using a separate set of input to predict stutter presence and evaluate the model's performance based on classification accuracy. The algorithm's output is a binary classification indicating stutter presence, contributing to advancements in speech processing and offering potential benefits in therapeutic and diagnostic settings for speech disorders.

4. Performance Evaluation

The gait recognition datasets from CASIA-B (Institute of Automation, Chinese Academy of Sciences) are included in the performance evaluation. A range of measures are employed to evaluate the efficacy of gait recognition techniques at varying viewing angles. Using the Sep-28k (Stuttering Events in Podcasts) dataset, a wide range of techniques, including class imbalance handling approaches, are assessed in the speech emotion identification domain. Metrics like P, R, In, F1-Score, B, and Total Accuracy are taken into consideration. The results are displayed as graphs and tables, and these assessments are compared with the PS methods for handling the particular difficulties presented by gait and speech recognition tasks.

4.1. Dataset Details

CASIA-B [20]: The experiment described in this paper utilizes the CASIA-B dataset for both the training and testing phases. In 2005, a comprehensive multi-view gait database known as CASIA-B was compiled and made accessible to the public. The study involved a cohort of 124 individuals, and gait data was collected using 11 different views. The viewpoints were constructed using angles ranging from 0 to 180 degrees, with a consistent 18-degree interval between each perspective. The CASIA-B system classifies walking into three distinct categories: regular walking (NM), walking while holding a bag (BG), and walking while wearing a coat or jacket (CL). The frame rate of the CASIA-B dataset is 25 frames per second (FPS). The dataset consists of a total of 798,000 testing frames and 854,000 training frames. The gait recognition methods employed in this study were trained and evaluated using the CASIA-B dataset.

SEP-28k [4]: The SEP-28k stuttering dataset involved the identification and inclusion of a total of 385 podcasts in the selection process. The length of the initial podcast recordings exhibits variability. Each section is delineated by a consistent duration of three seconds. A total of 28,177 segments were collected. The SEP-28k dataset consists of two distinct label categories: stuttering and non-stuttering. Speech characteristics such as nonsensical speech, unclear speech, absence of speech, poor audio quality, and the presence of music are not considered indicative of stuttering and thus do not hold relevance in the context of our study. Speech disorders, such as stuttering, encompass a variety of speech behaviors, which consist of blocks, prolongations, repetitions, interjections, and fluent segments. The dataset consists of the following types of speech: 10.35 hours of uninterrupted speech, 3.34 hours of interruptions, 2.74 hours of recitation, 1.48 hours of extension, 1.75 hours of block, and an additional 1.48 hours of extension. After the completion of the tagging procedure, each voice fragment, which has a duration of three seconds, undergoes down sampling. The frequency of the voice fragment is reduced to 16 kHz through the down sampling procedure.

4.2. Experimental Analysis

The Figure 2 presents a comparative analysis of gait recognition methods across various viewing angles. The PS method emerges as the most robust, with an mean accuracy of 98.8%, indicating its efficient performance and consistency across all angles. GaitTAKE (ES) [21] also shows high consistency, with a notable mean accuracy of 96.2%. Methods like GaitGL [22] and MT3D [23] perform well, with means above 94%, demonstrating their effectiveness in gait recognition tasks. The Gaitset [23] method, while generally effec-

tive, shows the greatest variation in performance, especially at more extreme angles such as 0 ° and 180 °, with its lowest score being 80.2% at 180 °. The marked decrease in performance at these angles for Gaitset [24] suggests that it is more sensitive to angle changes compared to other methods. Conversely, the PS method not only maintains high accuracy across all angles but also achieves the highest individual score of 98.89% at 36 °, underlining its potential for practical application where angle variance is a concern.

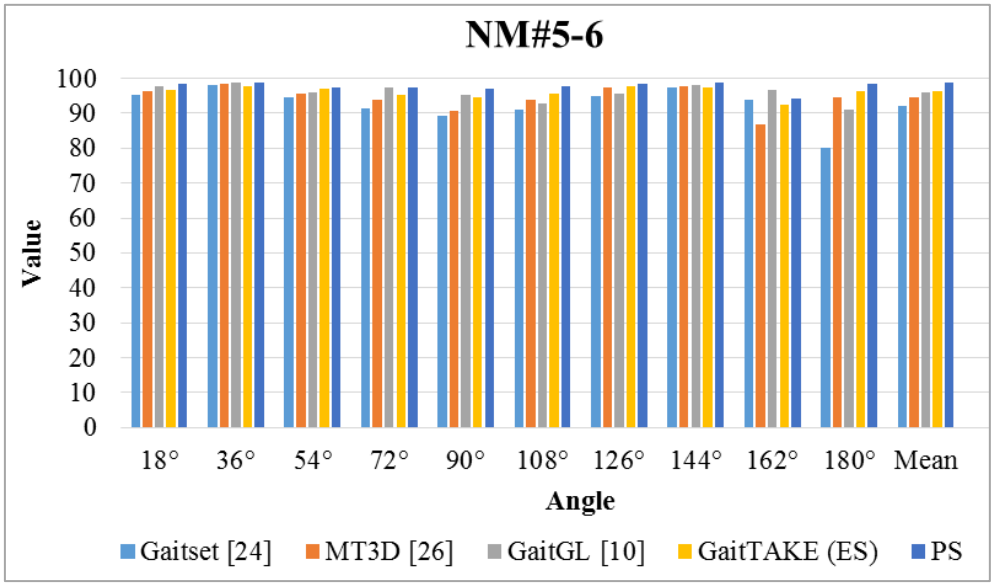

Figure 2. *Accuracy comparison of state-of-art techniques with PS for NM#5-6.*

Figure 3 showcases the performance of various gait recognition methods across multiple angles, revealing that the PS method outperforms others with a high average accuracy of 96.89%. This method not only achieves better results overall but also demonstrates remarkable consistency across all viewing angles, suggesting its robustness for practical applications. GaitTAKE [26] also exhibits strong performance with a 94.3% average, indicating its efficacy. GaitGL [9] and MT3D [25] present moderate results with averages of 92.1% and 89.8% respectively, while Gaitset [23] trails with an 84.3% average, showing notable performance dips at extreme angles. The data indicates that some methods are sensitive to angle variations, particularly at 0 ° and 180 °, the PS method maintains a high level of accuracy throughout, highlighting its potential as the most reliable method for gait recognition across varied conditions.

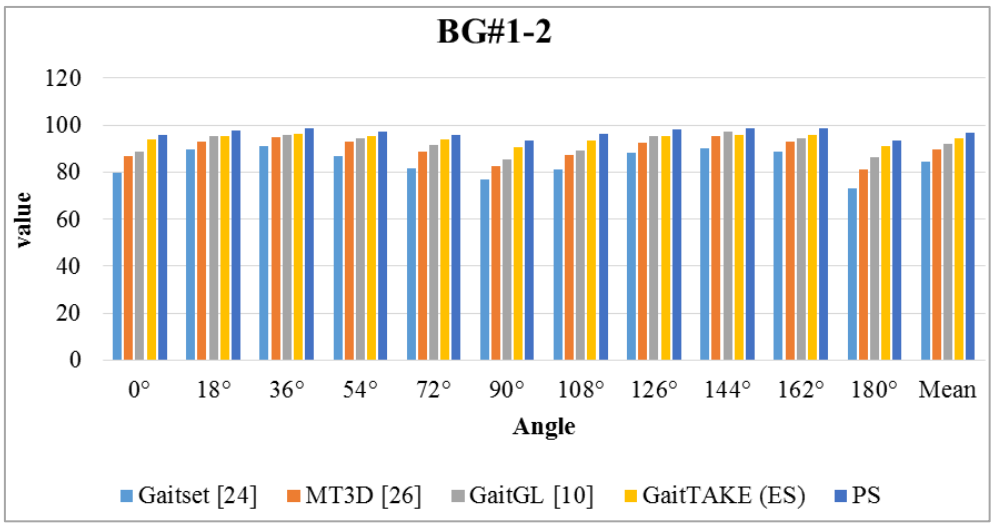

Figure 3. *Accuracy comparison of state-of-art techniques with PS for BG#1-2.*

The Figure 4 assesses the performance of five gait recognition models across a range of angles. The PS model stands out with the highest mean accuracy of 90.87%, suggesting it is the most effective and consistent across varying angles. GaitTAKE (ES) [26] also performs well, with a notable mean accuracy of 86.5%, indicating robustness. GaitGL [27] follows with a mean of 78.3%, while MT3D [25] has a lower mean accuracy of 75.6%. Gaitset [23] exhibits the least effective performance with a mean of 62.5%, including a significant drop at 240 °. The higher performance of the PS model, particularly at 45 ° and 60 °, where it achieves over 95% accuracy, underscores its potential for applications where accuracy across diverse angles is crucial. Conversely, the marked decline in Gaitset [23] at 240 ° to 45.9% reveals a potential weakness at more extreme angles.

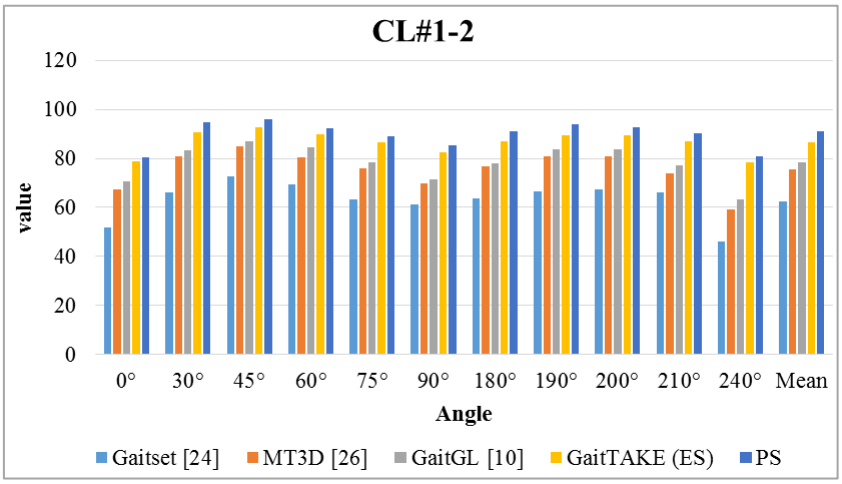

Figure 4. *7 Accuracy comparison of state-of-art techniques with PS for CL#1-2 setting.*

Analyzing the Figure 5, the PS method distinctly outperforms the various configurations of MC StutterNet across all metrics—R, P, B, In, and F. It achieves the highest scores in each category, most notably in B (15.84) and In (70.34), indicating a high F value (80.85), which is a measure of a test's accuracy. The MC StutterNet configurations show varying levels of performance, with MC StutterNet + No scoring lowest in R (28.54) and P (32.39) but highest in the F-value (80.42) among the MC StutterNet methods. This suggests that while MC StutterNet + No achieves a high overall accuracy. MC StutterNet + A4 shows strong performance in P (39.75), while MC StutterNet (Clean) has the highest score in R (33.36) among the MC StutterNets. However, all MC StutterNet configurations are outmatched by PS, which showcases superior overall performance, suggesting it may be the more robust method for applications.

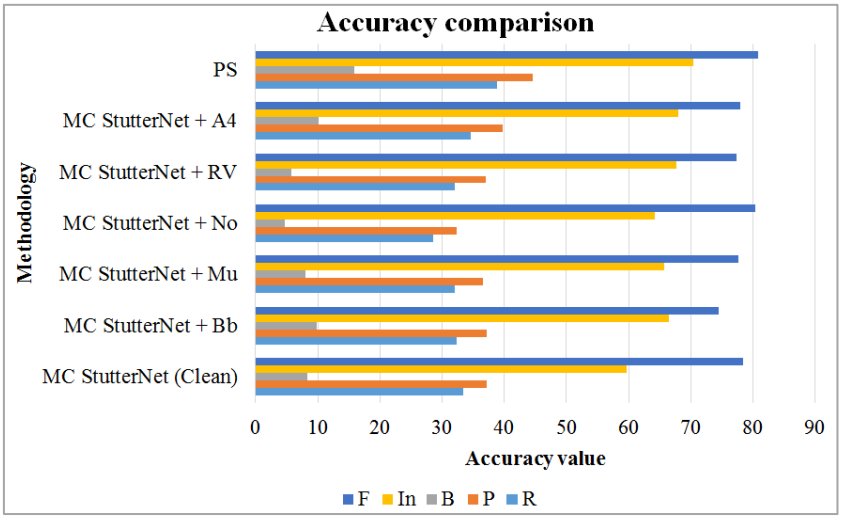

Figure 5. *Accuracy comparison.*

From the Figure 6, the PS method stands out with the highest Total Accuracy score of 66.312, indicating it surpasses the various MC StutterNet configurations in this metric. The MC StutterNet methods exhibit a narrower range of TA scores, from 58.71 to 61.72, with MC StutterNet + A4 having the highest TA score within the group. MC StutterNet + Bb has the lowest TA score, suggesting it may be less effective compared to the other configurations. Overall, the PS method shows a significantly higher TA score, which could suggest a more robust or efficient approach in the context of this analysis. The relative closeness of the TA scores among the MC StutterNet configurations indicates that while there are variations in their effectiveness, they are relatively less compared to the performance leap observed with the PS method.

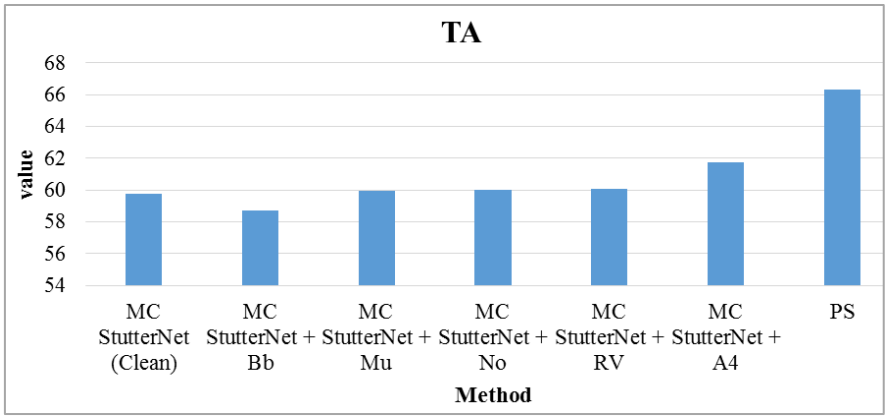

Figure 6. *Total Accuracy comparison for various state-of-art techniques comparison with PS.*

In Figure 7, the PS method significantly surpasses all MC StutterNet configurations with an F1 score of 49.78%. This score is indicative of a high balance between precision and recall, suggesting that PS is the most effective method for the metric evaluated. Among the MC StutterNet variations, MC StutterNet + A4 leads with an F1 score of 46%, while MC StutterNet + No lags behind at 42.8%. The other configurations MC StutterNet (Clean), MC StutterNet + Bb, MC StutterNet + Mu, and MC StutterNet + RV are clustered within a narrow range from 43.86% to 44.4%, showing relatively consistent performance. Despite this consistency, they all fall short compared to the PS method, which appears to offer a efficient approach, as reflected in its higher F1 score.

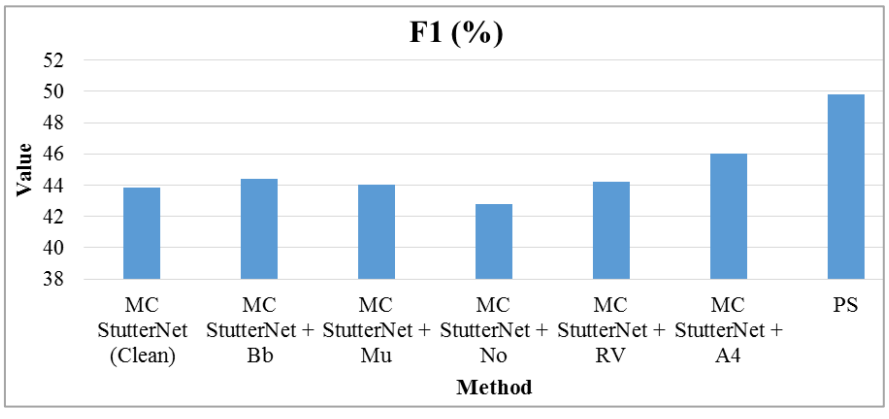

Figure 7. *F1-score comparison for various state-of-art techniques comparison with PS.*

5. Discussion

The experimental inquiry conducted a comprehensive comparison of different methodologies and found that the PS methodology outperformed its competitors in both gait identification and stuttering detection. The PS methodology consistently outperforms other approaches such as GaitTAKE, GaitGL, MT3D, and Gaitset in terms of gait identification accuracy. The efficacy of these approaches may vary due to changes in viewing angle, as they are sensitive to such varia-

tions. The resilience of the PS approach is highly notable, particularly when considering high angles. The PS approach outperforms previous MC StutterNet setups in terms of stuttering detection, achieving higher accuracy and F1 scores. The Particle Swarm (PS) technique has demonstrated its reliability in real-world scenarios, particularly in hard and niche fields. The ability to accurately identify gait and detect stuttering is indicative of its capabilities.

6. Conclusions

Stuttering can affect individuals of all ages, but it most commonly arises in young children during the developmental stages of speech and language acquisition. In conclusion, StrideSpeak represents a pivotal advancement in the fields of speech pathology and biometric recognition. By leveraging a sophisticated neural network architecture, StrideSpeak offers a dual-faceted approach that not only refines stutter detection but also enhances gait recognition, thereby addressing two critical domains with a single, integrated solution. The introduction of a global attention mechanism marks a significant leap in identifying and classifying stuttering more accurately, potentially transforming speech therapy practices and patient outcomes. Furthermore, StrideSpeak's handling of class imbalances within stuttering speech datasets culminates in a robust and reliable system that surpasses traditional models in performance. By extending the inclusivity and functionality of ASR systems, StrideSpeak also breaks new ground in enabling individuals with speech disorders to engage more fully with technology, reducing the societal and technological barriers they face. This research not only contributes to the scientific and medical communities but also has profound implications for enhancing the everyday lives of individuals with speech disorders, reinforcing the role of technology as a tool for empowerment and inclusivity. This work signifies a step forward in bridging the gap between the healthcare and technology sectors for combining the gait phenome with speech to detect stuttering, ultimately benefiting a wide range of applications and individuals.

Abbreviations

SD	Stuttering Detection
PWS	Persons Who Stutter
MFCC	Mel-Frequency Cepstral Coefficients
AGT	Adaptive Graph Topology
ConvNet	Convolution Network
SLP	Speech-Language Pathologist
EEG	Electroencephalogram
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
fNIRS	Functional Near-Infrared Spectroscopy
DBaaS	Database as a Service
AWS	Amazon Web Services
EDA	Electrodermal Activity
TDNN	Time-Delay Neural Network
VRET	Virtual Reality Exposure Therapy
BCI	Brain- Computer Interface
ASIS	Automated Stuttering Identification System
LS SVM	Least Square Support Vector Machine
EGG	Electroglottogram
PS	Particle Swarm
UNNIGSA	Unified Neural Network for Integrated Gait and Speech Analysis

Acknowledgments

I would like to express our sincere gratitude to all those who have supported and contributed to this research project. Primarily, I extend our heartfelt thanks to our guide for his unwavering guidance, invaluable insights, and encouragement throughout the research process.

Author Contributions

Ravikiran Reddappa Reddy: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software

Santhosh Kumar Gangadhariah: Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

Funding

No funding is raised for this research.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix

The supplementary material can be accessed at https://link.springer.com/chapter/10.1007/978-981-19-3148-2_61

References

[1] S. A. Sheikh, M. Sahidullah, F. Hirsch and S. Ouni, "Advancing Stuttering Detection via Data Augmentation, Class-Balanced Loss and Multi-Contextual Deep Learning," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 5, pp. 2553-2564, May 2023, https://doi.org/10.1109/JBHI.2023.3248281

[2] R. Hosseini, B. Walsh, F. Tian and S. Wang, "An fNIRS-Based Feature Learning and Classification Framework to Distinguish Hemodynamic Patterns in Children Who Stutter," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 6, pp. 1254-1263, June 2018, https://doi.org/10.1109/TNSRE.2018.2829083

[3] A. -K. Al-Banna, E. Edirisinghe and H. Fang, "Stuttering Detection Using Atrous Convolutional Neural Networks," 2022 13th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2022, pp. 252-256, https://doi.org/10.1109/ICICS55353.2022.9811183

[4] C. Lea, V. Mitra, A. Joshi, S. Kajarekar and J. P. Bigham, "SEP-28k: A Dataset for Stuttering Event Detection from Podcasts with People Who Stutter," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6798-6802, https://doi.org/10.1109/ICASSP39728.2021.9413520

[5] B. Alhalabi, J. Taylor, H. A. Sanghvi and A. S. Pandya, "A Proposed Framework for Stutter Detection: Implementation on Embedded Systems," 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 2022, pp. 829-833, https://doi.org/10.1109/AIC55036.2022.9848966

[6] J. Zhang, B. Dong and Y. Yan, "A Computer-Assist Algorithm to Detect Repetitive Stuttering Automatically," 2013 International Conference on Asian Language Processing, Urumqi, China, 2013, pp. 249-252, https://doi.org/10.1109/IALP.2013.32

[7] S. A. Waheed and P. S. Abdul Khader, "IoT based approach for detection of dominating emotions in persons who stutter," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020, pp. 14-18, https://doi.org/10.1109/I-SMAC49090.2020.9243392

[8] T. Kourkounakis, A. Hajavi and A. Etemad, "FluentNet: End-to-End Detection of Stuttered Speech Disfluencies With Deep Learning," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2986-2999, 2021, https://doi.org/10.1109/TASLP.2021.3110146

[9] S. A. Sheikh, M. Sahidullah, F. Hirsch and S. Ouni, "Robust Stuttering Detection via Multi-task and Adversarial Learning," 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 190-194, https://doi.org/10.23919/EUSIPCO55093.2022.9909644

[10] K. Li et al., "Applying multivariate segmentation methods to human activity recognition from wearable sensors' data," JMIR mHealth uHealth, vol. 7, no. 2, Feb. 2019, Art. no. e11201, https://doi.org/10.2196/11201

[11] H. Geng, Z. Huan, J. Liang, Z. Hou, S. Lv and Y. Wang, "Segmentation and Recognition Model for Complex Action Sequences," in IEEE Sensors Journal, vol. 22, no. 5, pp. 4347-4358, 1 March1, 2022, https://doi.org/10.1109/JSEN.2022.3144157

[12] Z. Wang, S. Hou, M. Zhang, X. Liu, C. Cao and Y. Huang, "GaitParsing: Human Semantic Parsing for Gait Recognition," in IEEE Transactions on Multimedia, vol. 26, pp. 4736-4748, 19 October 2023, https://doi.org/10.1109/TMM.2023.3325962

[13] A. N. Tarekegn, M. Sajjad, F. A. Cheikh, M. Ullah and K. Muhammad, "Efficient Human Gait Activity Recognition based on Sensor Fusion and Intelligent Stacking Framework," in IEEE Sensors Journal, vol. 23, Issue. 22, pp. 28355-28369, 02 October 2023, https://doi.org/10.1109/JSEN.2023.3319353

[14] A. Smith and C. Weber, "How stuttering develops: The multifactorial dynamic pathways theory," JSLHR, vol. 60, no. 9, pp. 2483–2505, 2017, https://doi.org/10.1044/2017_JSLHR-S-16-0343

[15] V. Mitra et al., "Analysis and tuning of a voice assistant system for dysfluent speech," in Proc. Interspeech2021, 2021, pp. 4848–4852, https://doi.org/10.48550/arXiv.2106.11759

[16] L. Verde, G. De Pietro and G. Sannino, "Voice disorder identification by using machine learning techniques," IEEE Access, vol. 6, pp. 16246–16255, 2018, https://doi.org/10.48550/arXiv.2106.11759

[17] N. P. Narendra and Paavo Alku. 2019. Dysarthric speech classification from coded telephone speech using glottal features. Speech Commun. 110, C (Jul 2019), 47–55. https://doi.org/10.1016/j.specom.2019.04.003

[18] C. Quan, K. Ren and Z. Luo, "A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech," in IEEE Access, vol. 9, pp. 10239-10252, 2021, https://doi.org/10.1109/ACCESS.2021.3051432

[19] S. Alharbi et al., "A lightly supervised approach to detect stuttering in children's speech," in Proc. Interspeech2018, pp. 3433–3437, https://doi.org/10.21437/Interspeech.2018-2155

[20] Shiqi Yu, Daoliang Tan and Tieniu Tan, "A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition," 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 2006, pp. 441-444, https://doi.org/10.1109/ICPR.2006.67

[21] Peng, Y., Ma, K., Zhang, Y. et al. Learning rich features for gait recognition by integrating skeletons and silhouettes. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-15483-x

[22] B. Lin, S. Zhang and X. Yu, "Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 14628-14636, https://doi.org/10.1109/ICCV48922.2021.01438

[23] C. Fan et al., "GaitPart: Temporal Part-Based Model for Gait Recognition," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 14213-14221, https://doi.org/10.1109/CVPR42600.2020.01423

[24] H. -M. Hsu, Y. Wang, C. -Y. Yang, J. -N. Hwang, H. L. U. Thuc and K. -J. Kim, "Learning Temporal Attention Based Keypoint-Guided Embedding for Gait Recognition," in IEEE Journal of Selected Topics in Signal Processing, vol. 17, no. 3, pp. 689-698, May 2023, https://doi.org/10.1109/JSTSP.2023.3271827

[25] Beibei Lin, Shunli Zhang, and Feng Bao. 2020. Gait Recognition with Multiple-Temporal-Scale 3D Convolutional Neural Network. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 3054–3062. https://doi.org/10.1145/3394171.3413861

[26] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in Proc. AAAI Conf. Artif. Intell., 2019, pp. 8126–8133, https://doi.org/10.48550/arXiv.1811.06186

[27] Rubén San-Segundo, Jaime Lorenzo-Trueba, Beatriz Martínez-González, and José M. Pardo. 2016. Segmenting human activities based on HMMs using smartphone inertial sensors. Pervasive Mob. Comput. 30, C (August 2016), 84–96. https://doi.org/10.1016/j.pmcj.2016.01.004

Biography

Ravikiran Reddappa Reddy, earned his Bachelors of Engineering (BE) degree in ECE from VTU, Belagavi in 2009. He has obtained his master's degree in M.Tech (ECE) from VTU in 2012. And currently he is a research scholar at VTU, Belagavi doing his Ph.D in Electronics and Communication Engineering and also working as Assistant Professor in SJC Institute of Technology, Chickballapur, Karnataka. He has attended many workshops and induction programs conducted by various universities. His areas of interest are Image Processing and Signal Processing.

Santhosh Kumar Gangadhariah, is a Professor and Principal in the Electronics and Communication Engineering Department at East West College of Engineering, Bangalore with an experience of 15 years in Teaching. He is qualified in Bachelor and Master Degrees in Electronics an Degrees in Electronics and Communication Engineering and and Ph.D in Electronics and Communication Engineering in the area of Image Processing. His areas of interest are Image processing and Signal Processing.

Research Field

Ravikiran Reddappa Reddy: Image Processing, Signal Processing, Controls and Systems, Communication Systems, Wireless Communications

Santhosh Kumar Gangadhariah: Image Processing, Signal Processing, Controls and Systems, Communication Systems, Embedded systems.