Formal Verification of an Enhanced Deep Learning Model: Unveiling Computational Effectiveness in Speech Recognition

Toktam Zoughi [a,*], Mahmood Deypir [b]

a. Department of computer engineering, Shariaty College, National University of Skills (NUS), Tehran, Iran

b. Faculty of Computer Engineering, Shahid Sattari Aeronautical University of Science and Technology, Tehran, Iran

*. Corresponding author.

E-mail addresses: t.zoughi@shariaty.ac.ir (T. Zoughi);

mdeypir@ssau.ac.ir (M. Deypir)

Abstract—Automatic speech recognition (ASR) assumes a crucial function in various domains, improving search engines, aiding healthcare with medical reporting and diagnosis, enhancing service delivery, and facilitating effective communication in service providers. This paper introduces FNNRA (Flexible Neural Network with Recursive Architecture), a novel method aimed at addressing overfitting issues in environments with limited training datasets in the field of automatic speech recognition (ASR). FNNRA utilizes a sophisticated architecture to extract and analyze important data features while maintaining data integrity through deep network layers. Theoretical and practical evaluations demonstrate FNNRA's ability to handle speaker variations, effectively train with constrained data sets, and broaden its relevance past speech understanding. The method is evaluated on established datasets like CallHome, TIMIT, and FarsDAT, showcasing its adaptability and efficacy across different data contexts. Comparative analysis with leading speech recognition methods reveals FNNRA's superior performance, achieving significant reductions in phoneme recognition errors by approximately 7.88%. This research sets a strong foundation for future advancements in the field and underscores FNNRA's potential in enhancing recognition systems, warranting further investigation.

Keywords **— Recursive Architecture of deep model, Flexible Neural Network, Data adaptability, Multimodal processing, Neural networks.**

1. INTRODUCTION

Automated speech identification (ASR) systems strive to transform human vocal signals into sequences of terms or phonetic units [1-3], playing a crucial role in transcribing spoken language into written text. This study introduces a robust acoustic model that emphasizes phoneme recognition [4-6]. ASR acts as an entry point for a variety of clever and proficient systems, impacting

numerous fields, including enhancing search engines with voice recognition, aiding medical reporting and disease diagnosis in the healthcare industry, improving service delivery, and facilitating communication in service providers by directing callers to the appropriate operators.

ASR has numerous applications, including speech interfaces for mobile devices, telephone routing, automatic dictation for tasks like letter writing or radiology reporting, managing home systems remotely, setting up Interactive Voice Response (IVR) systems, and labeling speech and video content for online searches [7, 8]. Reliable ASR systems are indispensable across all these scenarios [9-11]. Certain research investigations yield valuable results in the domain of ASR [12-13]. While ASR technology is extensively utilized in a variety of contexts [14-17], its performance can be significantly enhanced under challenging circumstances [18-21] to achieve a level of accuracy akin to human recognition [22-24]. This discrepancy may be attributed to two primary factors: variations in phoneme duration and spectral alterations.

The initial classification pertains to variations in the length of phoneme expressions within a unit of speech, which stem from diverse factors [25-26]. Individuals exhibit variations in the rate at which they articulate phonemes, with speech velocity differing from person to person. Moreover, an individual's speaking pace can vary across different contexts. As a result of these diverse speech speeds, phoneme durations may vary in length. Furthermore, the length of phoneme expressions is also influenced by contextual factors. The schematic representation in Fig. 1 delineates the sequential stages of the speech recognition procedure. Initially, a speech signal is captured, after which the most pertinent features are extracted. Subsequently, an acoustic model and a language model are employed to derive textual information from the signal.

The second classification of alterations arises from variations in the spectrum of the speech signal. Variations in the speech channel's attributes, the lengths of vocal tracts, accents, individual speaker characteristics, and the gender of speakers contribute to alterations in the spectrum of speech signals. Consequently, automatic speech recognition encounters a significant obstacle stemming from the fluctuations in speech signals seen both across different speakers and within a single speaker's speech. Enhancing the accuracy of the audio model can be achieved by incorporating models that account for the duration of phoneme expression and spectral transformations within the speech signal.

On the contrary, the utilization of deep neural networks in acoustic modeling involves the initial presentation of input data to the foremost layer. However, with the network's progression, subsequent layers begin to lose track of this input data. Consequently, the flow of information within the intricate model becomes constrained, impeding the efficient extraction of insights from the input data. By acknowledging and incorporating these fluctuations within the model, a reduction in error rates is achieved, rendering our model resilient to such fluctuations.

In this study, our objective was to address the previously mentioned challenges by introducing a novel recursive neural network that provides increased adaptability. Past research has extensively delved into approaches for adapting to speakers and managing

varying lengths of utterances [17, 27-29]. The aim of our research is to establish a framework capable of adapting to changes in phonemic length and compensating for alterations in spectral characteristics. In lieu of presenting an independent model for speech length, our strategy implicitly contemplates variations in speech duration and spectral attributes. Previous endeavors to tackle these hurdles have primarily concentrated on amplifying convolutional network structures. This study reveals an original neural network structure referred to as the Flexible Neural Network (FNN), which captures localized alterations in the time-frequency domain through adaptable-length windows applied to the spectrogram of speech signals. Notably, both temporal and frequency aspects are taken into account. The deep neural network we suggest showcases enhanced effectiveness in contrast to other models [30]. Our objective is to exploit the functionalities of residual neural networks to tackle challenges linked to overfitting discerned in deep neural networks constrained by scarce resources. Furthermore, our intention is to craft a potent acoustic model for languages possessing restricted speech data, like Persian. Within our suggested structure, we adjust the advancement of input data across the network. Residual models assume a critical role in analyzing and dissecting insights from data in deep layers. We have effectively addressed the challenge of integrating flexible windows and efficiently transmitting input data to deeper layers in complex models. This manuscript also introduces a novel deep neural network, denoted as the Flexible Neural Network with Recursive Architecture, which expands upon the concept of the flexible neural network. Our model enhances the design of the residual network by skillfully guiding weight parameters, resulting in more accurate predictions, particularly in profoundly deep models. This innovative model can exploit input data across all layers, thereby boosting the effectiveness of the speech recognition system. Evaluation outcomes suggest that the flexible neural network with recursive architecture provides an effective model that surpasses the deep residual network and outperforms recent approaches. In Section II, we explore the relevant literature. Section III elaborates on the FNN and examines Multiple residual networks featuring weighted links. Section IV details the experimental outcomes derived from the TIMIT, FarsDAT, and CallHome databases. Finally, Section V concludes by summarizing the key discoveries and contributions of the paper.

2. Related Works

The length of phoneme expressions harbors valuable linguistic information. Most systems for speech recognition fail to integrate models for phoneme or word utterance duration. Previous empirical results have shown that integrating models for phoneme expression duration enhances speech recognition [27-29]. Typically, two approaches exist for modeling utterance length. The initial method involves the portrayal of the time span of phonetic components through Hidden Markov Models. Conversely, the subsequent approach integrates the demonstration of the duration of phonetic elements via a distinct model alongside the acoustic model.

Within the primary classification, adjustments have been enacted on the configuration of the Hidden Markov model [29]. These techniques encompass substantial computational intricacy resulting in extended training durations and minimal enhancements in model efficiency. A majority of speech recognition systems make use of Hidden Markov Models (HMMs) to exemplify the duration of phonetic components. These methods often delineate the duration of phonetic elements through gamma or logarithmic-normal distributions. Techniques rooted in HMMs like Continuous Semi-Markov Models (CSMM) and Extended State Hidden Markov Models (ESHMM) generate a series of observations to depict the duration of phonetic components in each state [29, 31]. ESHMM documents alterations in phonetic component duration through the probability density function of a state [27]. In ESHMM, each HMM state is substituted with multiple alternate HMMs. Consequently, within each ESHMM state, multiple HMMs exist for modeling the duration of phonetic components. Certain methodologies like Conditional Random Fields (CRF) have been constructed based on conditional random fields [28]. CRF identifies features linked to speech duration to represent varied vowel lengths for diverse speakers.

In a different array of solutions, the duration of phonetic components is examined distinctively from the acoustic model. Several of these approaches employ feedforward neural networks to anticipate the probability density function of phonetic element duration [29]. Certain researchers assess the conditional likelihood of each phonetic element duration using a feedforward neural network [32].

Nonetheless, a robust model like Long Short-Term Memory (LSTM) surpasses the aforementioned methods in terms of efficiency as an acoustic model. The LSTM model can illustrate the duration of phonetic components in its hidden layers [32-33]. The fusion of HMM and LSTM yields commendable outcomes when compared to other deep neural networks for speech recognition [34]. Furthermore, LSTM-RNN-based acoustic models implicitly signify the duration of phonetic components, negating the necessity for multi-state HMMs [33].

3. MATERIAL AND METHODS

The precise identification of phonemes, each showcasing unique variations, presents a notable hurdle in the realm of speech recognition. Deep neural networks frequently encounter difficulties in categorizing phonemes with differing speech lengths into matching categories. Furthermore, training extremely deep neural networks with scarce data can trigger concerns related to overfitting [34-37]. To confront these obstacles, various remedies have been suggested in existing literature, encompassing techniques like early cessation, exclusion, data expansion, regularization, Residual Networks (Res-Net), and preliminary training [38].

Early cessation is applied throughout the training phase to avert overfitting. Training ceases upon the development set's error escalation, with the development set being entirely distinct from the test and training data, devoid of any overlap. It is utilized to

fine-tune the model's parameters. Exclusion is another method that randomly nullifies certain neuron activities during training, promoting generalization and thwarting overfitting. Neurons are systematically eliminated during training, resulting in a network with an altered configuration. Data augmentation involves artificially enhancing the training data by modifying image scale, angle, and rotation, thereby amplifying the training data volume and diminishing overfitting [39]. Moreover, some strategies introduce minor weight modifications to the network's weights at every training epoch to prevent extreme weight values and alleviate overfitting. Residual blocks and preliminary training [40] are also deployed to tackle the overfitting predicament. Nevertheless, when employing profoundly deep Res-Net with constrained training data, the utilization of input information might not be optimal [41-42].

Recently, the amalgamation of Convolutional Neural Networks, Residual Neural Networks, and HMMs has achieved considerable triumph in speech handling [43]. These complex neural networks effectively outline detailed speech characteristics. In this study, our suggested method seeks to exploit and enhance deep neural networks as their effectiveness still significantly trails human recognition levels. Nevertheless, the proposed strategy abstains from presenting an independent model for phonetic element duration. Conversely, it implicitly outlines the duration of the phonetic unit.

To achieve these aims, we employ a creative approach known as the FNN, which combines diverse window sizes onto the spectrogram of the speech signal. Figure 2 illustrates the process of taking the raw speech signal as input and deriving Mel-frequency cepstral coefficients (MFCC) as characteristics, which act as input for the proposed ANNRA system. Our original method produces likelihoods for HMM conditions, where each condition represents a phoneme.

Observing that the automated speech recognition showcases notable discrepancies both amidst speakers and within an individual speaker's speech, largely attributed to fluctuations in the speech signal, we acknowledge the need for a novel strategy. Contemporary deep models often face constraints in information dissemination. To tackle these hurdles, we introduce an original acoustic modeling configuration titled flexible neural network with recursive architecture. Within this suggested technique, we regulate the progression of input specifics across the Res-Nets employing statistical perceptions. Furthermore, this mechanism ought to possess the ability to establish a flexible model adept at distinctly discerning a phonetic unit along with its individual duration and spectral characteristics. Enhanced data extraction from input data in automated speech recognition approaches yields superior outcomes and mitigates the possibility of overfitting.

The configuration of the residual neural network and recursive architecture share resemblances, yet they diverge in the subsequent areas. Initially, the adaptable neural network and weight distribution permit the utilization of diverse window dimensions within the network. Every frame symbolizes the span of an individual phonetic element within the time-frequency spectrum. Hence, if a phonetic unit endures an extended duration, it necessitates a correspondingly elongated frame in comparison to a unit with swift articulation, and vice versa. Subsequently, residual neural networks incorporate supplementary connections in addition to the

prevailing connections in the residual neural network. Consequently, the newly introduced model presents augmented dimensions in contrast to the original residual network. Furthermore, whereas each layer in the residual neural network assimilates insights from two or more antecedent layers, the proposed recursive structure assimilates insights from two layers preceding itself in conjunction with the input layer. Moreover, all links in recursive architectures possess weights that are distanced from the residual neural network. In the intermediary layers of the residual neural network, input specifics dwindle due to decreased connections. Conversely, in scenarios involving multi-path neural networks, input data methodically propagate to subsequent layers within the network. This regulated data flow indicates an escalation in the aforementioned training data within specific algorithms.

In Figure 3, a Flexible Neural Network with a Recursive Architecture (FNNRA), also known as FNNRA, is depicted. This framework, illustrated in Figure 3, comprises an FNN layer and 6 RA layers. In this visual representation, the initial layer interconnects with other layers. Consequently, particular stages within the framework establish connections with the initial layer's output through direct links. In this context, for clarity, the outcome of the first intermediate layer is denoted as h^1. Another hurdle in the residual layers of the network arises when each intermediary layer, represented as h^l for all $l \in 2L+1$, merges its output with two preceding layers h^{l-2} without accumulating weights. Typically, in a residual neural network, one layer possesses a direct link while the subsequent layer lacks such a connection. Let's assume the current layer as h^l without a direct connection and utilize a link originating from the two prior layers h^{l-2} for the lower layer. Consequently, the ensuing layer tends to yield higher output values than the ongoing layer, resulting in increased error margins. By introducing weights to the connections, an approximate 3% reduction in absolute PER is observed. In the proposed framework, each layer h^l is accompanied by a link to the initial layer h^1 and connections to two preceding layers h^{l-2}. These paired connections, termed previous information, hold significant importance. Each of these connections embeds valuable data that aids in delineating decision boundaries in the neural network. Input data from layer h^1 is effectively channeled through these connections to the deeper layers. By integrating these direct links, an intermediary layer can access details from both historical and current layers concurrently. This comprehensive data is pivotal for a more accurate analysis of input data in deeper layers.

In Figure 3, the proposed methodology is visually represented. The FA system, with its weighted connections spanning the framework, is distinctive, facilitating broad knowledge assimilation and generalization in lower layers while the upper layers focus on acquiring specific and distinct insights. This data propagation enhances evaluation and distinction capabilities in various settings. In the proposed method, past data is adjusted with a scaling factor of $1-\gamma$, while current data is adjusted with a scaling factor γ. The parameter α regulates the data transfer from the input layer across the system. As the process progresses, the parameter α decreases in tandem with the value of β. Consequently, a direct adjustment in the parameter β impacts the value of α. Through

the parameter α, data from layer h^{l-2} is merged with data from layer h^1. Experimental results suggest that optimal outcomes are achieved when $\beta = 0.3$. This parameter value of the scaling factor 0.3 showcases its influence on the h^1 layer as data is transmitted to deeper layers.

3.1. process of the proposed approach

The training methods for the proposed model include succinctly summarized as follows. The initial layer, referred to as the FNN, adapts to phoneme expression length variations by utilizing a convolution operator. The resulting output from the convolutional sublayer is then subjected to a maximum function and connected to a pooling sublayer. The maximum value within each window is transferred to the first pooling sublayer, followed by another maximum function on the output. The Flexible Neural Network with Recursive Algorithm (FNNRA) combines windows of various sizes under the second layer to ensure robustness against variations in phoneme duration, vocal tract length, and speaker characteristics. Batch Normalization (BN) is used during training to enhance model performance the adjusted rectified linear unit (ReLU) activation function is employed subsequent to the pooling process. The Recursive Algorithm (RA) calculates the result of the initial layer, with BN and ReLU coming after the convolution process at this phase. The fully linked layers beyond the RA systems consolidate characteristics from various frequency ranges. The SoftMax mechanism functions as the ultimate layer, determining the conditions of the Hidden Markov model. The FNNRA educates the complete intricate neural network through the backpropagation technique. Viterbi decoding is used to determine the phoneme sequence. The FNNRA method utilizes Gaussian Mixture Model-based Speaker Adaptive (SGMM) models to compute speaker information. These training methods are depicted in Figure 4 along with accompanying information.

The displayed arrangement of the FNNRA configuration can be delineated in the subsequent manner (refer to Figure 4): The primary layer encompasses RA and an FNN, comprising a convolutional sub-level and two composite sub-levels. Succeeding layers employ hidden RA layers, culminating in the establishment of three FNN panes spanning the spectrum derived from 11 speech frames. The outcomes from these panes are relayed to the RA layers. The hidden RA layers are crafted from non-pooling convolutional layers and incorporate the suggested direct connections. These layers employ convolution operations, batch normalization, and ReLU activation. ReLU activation introduces non-linear aspects to the layer outputs. The ultimate layer in the RA methodology utilizes the SoftMax activation mechanism. After the RA layers, three entirely linked layers are employed. The end layer gauges the probabilities of HMM states. The fusion of FNN and RA facilitates the application of an intricate model with restricted data.

In our proposed layout, the primary layer encompasses a Convolutional Neural Network (CNN) with compatible windows. This configuration empowers the model to adeptly manage disparities and speaker modifications. The fusion of the FNN and Recursive Algorithm (RA) is collectively denoted as FNNRA. Through this integration, the model can leverage the adaptability of FNN and

the recurrent essence of RA. To integrate broad speaker insights into the model, the FNNRA approach employs the subspace Gaussian mixture model methodology. This integration enhances the model's ability to accurately recognize speech across diverse datasets by considering speaker-specific characteristics. Before conducting experimental evaluations and comparisons, it is essential to mathematically prove the convergence and generalization properties of FNNRA [44-46]. This ensures that the model is capable of effective learning and adaptation to various inputs, leading to improved performance in speech recognition tasks. It is indicated that the FNNRA network converges if it is deep [30]. This demonstrates that with increasing network depth, it is expected that the model will converge and enhance its performance in speech recognition tasks. With the deepening of the network, the model gains more capability in capturing patterns and complex representations in speech data, consequently improving recognition accuracy and overall performance.

In the subsequent segment, we unveil empirical evidence to demonstrate the efficiency of the FNNRA technique. Additionally, we will delve into how different configurations of FNNRA influence the ultimate efficacy of automated speech recognition. By appraising performances on CallHome, TIMIT, and FARSDAT datasets, particularly for terms and phonemes, we discern heightened precision utilizing the proposed tactic. The results underscore the supremacy of the recommended strategy compared to other models for tackling acoustics.

4. EXPERIMENTAL EVALUATIONS

To conduct thorough experiments, it was essential to assess the effectiveness of the suggested FNNRA technique across three distinct speech datasets: CallHome, TIMIT, and FarsDAT. The assessment predominantly centered on pinpointing error margins for both lexical terms and phonetic elements. Experimental configurations were established, and progress was made by meticulously scrutinizing the outcomes derived from these datasets. The FNNRA approach leverages a spectrum of speech signals as its primary input. Feature vectors were generated through 25-millisecond speech frames and 10-millisecond sliding windows. This arrangement scrutinized a series of 11 consecutive speech frames within each segment. These frames underwent a transformation employing a Fourier transform-oriented filter array, restructured with a logarithmic energy coefficient arranged in the Mel scale. The resultant characteristic vectors, coupled with their primary and secondary temporal differentials, were subsequently inputted into the system. All tests were conducted on a computer equipped with a 2.88GHz processor, 8 processing units, and 32GB of random-access memory. The graphics processing unit (GPU) utilized in this system is the GeForce GTX 780, boasting a memory capacity of 6144 megabytes and 2304 CUDA processing cores.

4.1. Examining the Impact of different window sizes

In the first experimental trial, both CNN and FNN algorithms were examined. These algorithms were combined with the HMM model, referred to as CNN-HMM and FNN-HMM, respectively. Each algorithm educates a neural network comprising a pair of

convolutional layers and three fully linked layers. The CNN-HMM algorithm utilizes the deep convolutional neural network described in references [43-44] as the baseline method. In this experiment, as depicted in Fig. 5, the horizontal axis labeled with "3x3" represents the window size in the CNN-HMM method. However, in the FNN-HMM method, three windows of different dimensions are considered. The size of the first window matches the number on the axis (e.g., 3x3), while the sizes of the other two windows are obtained by adding two and four to that number. Hence, for the FNN method, three windows with sizes of 3x3, 5x5, and 7x7 are taken into account. However, for simplicity in visualization, only the size of the smallest window, 3x3, is shown. Consequently, in Fig. 5, the horizontal axis labeled with "5x5" for the FNN-HMM method indicates three windows with sizes of 5x5, 7x7, and 9x9, whereas the "5x5" label on the axis for the CNN-HMM method represents a convolutional neural network with a window size of 5x5.

Fig. 5 examines the influence of increasing window sizes on phoneme recognition error. The x-axis symbolizes the size of the window, with the y-axis denoting the error in recognizing phonemes in the TIMIT dataset. A lower phoneme recognition error signifies a more desirable outcome, indicating the algorithm's accuracy in assigning the correct label to each data point. As observed in Fig. 5, considering the variations in the proposed algorithm has enhanced its efficiency. The findings suggest that our technique outperforms the CNN-HMM approach in nearly all instances on the TIMIT dataset. The improved results can be attributed to the structure of the proposed model, which considers three different window lengths, accommodating phonemes with diverse lengths and spectral characteristics. As shown in Fig. 5, a window size of 5x5 yields better results compared to other window sizes. This is because a very small window size (e.g., 3x3) fails to capture the necessary information for a phoneme, while an excessively large window size increases the likelihood of irrelevant information entering the system, leading to a higher phoneme recognition error.

4.2. Investigating the Influence of the Quantity of Fully Connected Layers

In the subsequent experiment, we delve into the consequences of augmenting the quantity of fully linked layers. Both CNN and FNN algorithms undergo evaluation within this trial. The neural networks of both algorithms entail dual convolutional layers, with the quantity of fully connected layers fluctuating. Within the FNN-HMM technique, window dimensions of 5x5, 7x7, and 9x9 are employed, while the CNN-HMM technique employs a window dimension of 5x5. The horizontal axis in Figure 6 illustrates the count of fully connected layers, whereas the vertical axis signifies the phoneme detection error rate. For instance, the digit 2 on the horizontal axis denotes that the proposed blueprint integrates 2 fully connected layers. Figure 6 scrutinizes the repercussions of enhancing the quantity of fully linked layers on phoneme detection errors. As illustrated in Fig. 6, better results are achieved, and the recognition error decreases as the depth of the feedforward neural network with fully connected layers increases. The improved results can be attributed to the structure of the proposed model, as increased depth enhances the ability to hierarchically process and analyze data in higher layers. Lower layers contain general information, whereas higher layers capture discriminative

information. As observed in Fig. 6, increasing the number of hidden layers enhances recognition accuracy.

4.3. Examining the Impact of γ Values

In this study, we explore the impact of varying γ values. The proposed approach involves scaling prior data by the factor $1-\gamma$ and current data by the factor γ. Figure 7 showcases how the γ parameter influences phoneme recognition errors. The graph displays $(\gamma, 1-\gamma)$ values along the horizontal axis and phoneme recognition errors in the TIMIT dataset along the vertical axis. Multiple $(\gamma, 1-\gamma)$ values are tested in this analysis.

As showcased in Figure 7, peak efficiency is attained when the model seamlessly merges 50% of historical data with 50% of present data. This outcome is realized when the parameters $(\gamma, 1-\gamma)$ are set to (0.5 and 0.5). This study underscores the significance of historical data and the incorporation of input data within the network's intermediate layers. Additionally, the illustration emphasizes the advantage of integrating input data into the model's flow. While input data can traverse other connections within the model, merging it via a distinct connection with the remainder of the model offers enhanced control over the input data. These attributes facilitate a more thorough analysis of input characteristics within deeper layers.

4.4. Examining the Impact of Hyperparameter α

Here, we explore the influence of modifying the hyperparameter α on the FNNRA system. The outcomes of these investigations are showcased on both the evaluation group and development group of the TIMIT dataset. The influence of α on the efficiency of speech recognition is scrutinized. The α parameter dwindles at every phase of the procedure with a *constant* value. By modifying this $constant$ parameter, the value of α undergoes alterations. Hence, in this investigation, we showcase how this $constant$ factor impacts the network's efficacy in speech understanding. Initially, the AWCNN is implemented in the first layer of the proposed design. Various window sizes—5×5, 7×7, and 9×9—are analyzed within the AWCNN framework. Subsequent to the AWCNN phase, 18 MRes layers with 3×3 windows are integrated. To optimize outcomes based on previous experiments $(\gamma, 1-\gamma)$ values are configured to (0.5 and 0.5).

Figure 8 delineates the influence of the $constant$ parameter (on the x-axis) on phoneme recognition inaccuracies (on the y-axis) for the FNNRA structure with 19 convolution layers. One line signifies evaluation on the test dataset, while the other represents evaluation on the development dataset.

The α parameter regulates the transmission of input data within the network. Data from the $(l-2)$ th layer is merged with data from the initial layer utilizing the α parameter. As shown in Figure 8, the most favorable outcome materializes at $Const = 0.3$. At Const=0.3, the impact of data from the h^1 layer diminishes by a factor of 0.3 as it progresses to deeper layers. At this *constant*

value, the third layer solely receives its entire preceding data from h^1. Within the fifth layer, 70% of its past data is derived from h^1 and 30% from h^{l-2}. The seventh layer acquires 40% of its prior data from h^1 and 60% from h^{l-2}. The ninth layer incorporates only 10% of its past data from h^1 and 90% from h^{l-2}. Beyond the eleventh layer, the Const variable becomes negative, indicating zero value enforcement behavior by the algorithm. Layers past the ninth behave akin to a traditional residual learning network where they do not receive data from h^1 [45]. By augmenting the $constants$, lower layers acquire a greater share of data from h^1, whereas deeper layers do not receive data from h^1.For instance, at $Const = 0.4$, layers beyond the seventh do not receive data from h^1. Shallower layers necessitate more data input, whereas in this instance, the data input is unsuitable for shallower layers. Hence, opting for $Const = 0.4$ leads to an escalation in error rates. At $Const = 0.05$, data from h^1 extends up to the 51st layer. Nevertheless, the higher layers do not necessitate data from h^1, which contains raw data, as they require more discriminative data. Fig. 8 illustrates that choosing smaller α values also diminish algorithm accuracy. Consequently, by selecting suitable α values, the FNNRA method excels due to its capability to learn both general and discriminative traits. These attributes result in increased error rates when opting for $Const = 0.05$.

In the forthcoming experiment, we delve into the quantity of parameters essential for diverse networks. Columns 2 and 3 within Table. 1 respectively present the network identifier and the layer count. To ensure equitable comparison across all networks, we stipulate a total of ten hidden layers. Column 4 in Table. 1 illustrates the computed sum of adaptable parameters for each network. The initial layer accepts the speech input signal, hence devoid of adaptable parameters. The parameter count for each convolutional layer is denoted by $l \times h \times f \times k$, with h and l indicating the layers and hidden units correspondingly. Moreover, $f \times k$ represents the window magnitude on the speech spectrogram. Every sub-layer includes a bias, yet biases are disregarded for all layers. In pooling sub-layers, numerous neighbors are substituted by their maximal or mean value, thus no adaptable parameters exist in the pooling sub-layer. For each layer with complete connections, each input unit links to a hidden unit with an individual weight. Consequently, the parameter count for the initial layer is $n \times h$, where n denotes the input quantity and h signifies the hidden units. The parameter count for the remaining layers with complete connections is $h \times h = h^2$, where h signifies the hidden units. This procedure recurs for the other hidden layers. Hence, the parameter counts for a network with l hidden layers is $n \times h + l \times h^2$.

Within Table. 1's fifth column, the precise calculation of parameter quantities for each network is detailed. Throughout most experiments detailed in this paper, 7×7 window sizes have predominated. Hence, $f = k = 7$ is the established norm. With $h = 1024$ hidden units utilized in the tests, the corresponding value in the sixth column of the table is derived. Notably, the

parameter count in the proposed FNNRA technique aligns closely with that of the residual neural network model. Solely fully deep neural networks exhibit a notable surplus of parameters compared to alternative approaches. This divergence stems from the absence of convolutional layers in deep neural networks. Consequently, except for deep neural networks, which necessitate prolonged training durations due to their augmented parameter count, the parameter quantities and training durations across other networks remain relatively consistent.

4.5. Results of phoneme detection in the FARSDAT dataset

The efficiency of the novel method was assessed through experimental tests on a Persian dataset called FARSDAT. As depicted in Table 2, the FARSDAT dataset encompasses 304 speakers, encompassing both genders, with a total of 4738 vocalizations. Following the precedent of prior investigations [31], the datasets were segregated into training, validation, and test sets, comprising 3994 utterances for training, 457 for validation, and 287 earmarked for testing. This segment delves into the appraisal of different methodologies using the evaluation and validation datasets drawn from FARSDAT. Figure 9 delivers an extensive assessment of the proposed FNNRA-HMM technique alongside other sophisticated strategies on the FARSDAT dataset. This illustration scrutinizes an array of deep networks and integrates endorsed strategies, all of which integrate the HMM framework.

In the FNNRA approach, a batch size of 256 is employed with Stochastic Gradient Descent (SGD). The initial learning rate is established at 0.08, with a reduction factor of 0.5 triggered when the discrepancy in errors between successive stages in the development set falls below 0.02, alongside a momentum coefficient of 0.9. Across the three datasets detailed in this segment, the values for γ and $(1-\gamma)$ are precisely fixed at 0.5 each, while β is configured to 0.3.

Specifically, Figure 9 showcases the outcomes of the FNNRA-HMM mechanism featuring 19 convolutional layers in contrast to alternative models. Nonetheless, this analysis concentrates solely on contrasting the outcomes of various approaches onto the test collection. The initial and second columns in Figure 9 (b) portray the influence of employing CNN and FNN methodologies with a setup embracing 10 convolutional and dense layers. These bars underscore that utilizing the FNN approach correlates with a decline in Word Error Rate (WER). Moreover, the third and fifth columns from the top in Figure 9 (b) reveal that the innovative FNNRA method attains a substantial 7.88% reduction in error rate when juxtaposed with the DNN-HMM technique. Furthermore, an evaluation between Res and FNNRA networks was carried out. Figure 9 demonstrates that the FNNRA-HMM approach showcases a remarkable 5.53% decrease in error rate relative to the Res-HMM strategy, comprising 19 convolutional layers.

4.6. Phone recognition results on TIMIT

At the outset, we undertake a comparison of error rates among diverse algorithms utilized in tasks pertaining to speech recognition. The TIMIT dataset holds widespread recognition within the realms of speech and acoustic exploration, standing as a yardstick for appraising automated speech recognition systems. Table 3 furnishes an outline of the fundamental attributes of the TIMIT

repository. It encompasses an exhaustive compilation of 6,300 phrases articulated by 630 distinct speakers. The captured sounds are archived in a 16-bit structure and sampled at a frequency of 16 kHz. Each speaker presents 10 sentences sourced from 8 disparate linguistic regions spread across the United States. Notably, the dataset incorporates two sentences shared by all speakers. For assessment purposes, the complete TIMIT dataset is employed. Our scrutiny is specifically directed towards 24 speakers from the primary test assemblage, encompassing 192 unique sentences segregate from the development set. The training corpus incorporates 3,696 sentences vocalized by 462 speakers, all enunciating the two shared sentences (SA1 and SA2). Initially, within the TIMIT dataset, 61 phonetic units are correlated to 48 categories for acoustic modeling. Nevertheless, for the sake of phonetic identification error documentation, these categories are further condensed to 39 classes. Phonetic recognition is executed through the utilization of the Viterbi algorithm coupled with a bigram language model.

In order to assess the efficacy of the proposed approach, we juxtapose it with cutting-edge methods in the realm of speech recognition. The evaluation includes the presentation of phoneme recognition errors for both the existing approaches and the proposed method. This comparison allows for a comprehensive analysis of the performance and improvements offered by the FNNRA approach in the context of the TIMIT dataset.

In this section, we present an evaluation of diverse approaches in conjunction with the suggested model. We executed the Res-HMM method utilizing diverse configurations. It should be noted that this model can similarly be applied with an increased count of convolutional layers, for instance, 51 or 101. The error rate of sound detection versus various $constant$ values for the FNNRA method in the TIMIT dataset. Experimental findings demonstrate a notable enhancement in the model's efficiency and precision by incorporating data not only from the current layer but also from the two preceding layers. We also assess the CNN-HMM method with and without the inclusion of RA and Res networks. Moreover, investigations are carried out on the FNN-HMM model both with and without the integration of Res and RA components. It is essential to underscore that all these models are deep and encompass HMM within, ensuring a just and rational comparison among them.

Figure 10 illustrates a comparative analysis between the FNNRA-HMM methodology proposed in this research and the most intricate methods in the current literature, particularly focusing on their performance within the TIMIT dataset. This assessment encompasses FNN-HMM and CNN-HMM techniques, which also embrace a similar HMM-oriented strategy. Additionally, we explore the amalgamation of Res and RA networks with these techniques. The amalgamated approach of FNN with RA is termed FNNRA, while the hybrid methodologies of FNN-HMM and CNN-HMM with Res are denoted as FNN-Res-HMM and CNN-Res-HMM, respectively.

In Figure 3, the similarity of layers and connections within the proposed strategy to the residual network is evident. However, the bypass links between these two systems differ. Notably, the crucial layer in the FNN-Res-HMM technique showcases an FNN that sets it apart from the other layers. This FNN evaluates three distinct window dimensions (3x3, 5x5, and 7x7) in the initial layer

and functions as input for the subsequent layer. Beyond the initial layer, the subsequent layers of the residual network use a window size of 3x3. By taking into account the diverse lengths of speech articulation and the amalgamation of three windows, the FNN model heightens its adaptability to changes in speech length and temporal frequency fluctuations.

Within Figure 10(b), the primary column displays the initial phoneme error rate (PER) for the GMM-HMM method computed through maximum likelihood estimation [46]. In this evaluation, a 3-state left-to-right HMM with 40 Gaussian components achieves a PER of 26.43%. The eighth column from the top represents the proposed FNN-HMM methodology, yielding a PER of 20.88% in the TIMIT test set. In contrast to the GMM-HMM approach, the FNN-HMM method diminishes the absolute PER by over 5.55%.

The exhibited columns from 1 to 12 offer a comparison between sophisticated methodologies and the proposed approach within the TIMIT dataset. The results underscore the unwavering superiority of the suggested strategy over most of these methodologies. Particularly in column 11, the proposed technique with 19 convolutional layers and 3 concatenated layers showcases significant advancements. The FNNRA-HMM network arrangement simplifies effective data management and information transmission to deeper layers. Empirical assessments have evidenced that integrating RA results in a boost of more than 5.15%. Bars 10 and 11 highlight that FNNRA methodologies with 19 convolutional layers outperform Res-nets. With the escalation in layers count in Res-nets, data information fails to disseminate throughout the entire profound model, diminishing its efficacy. This issue is tackled through both pre-training and the proposed FNNRA technique.

A distinct trial was executed to explore the exact influence of the FNNRA method on diverse frameworks. Bars 9 and 11 exhibit a notable 3.43% decrease in absolute inaccuracies, accomplished with the recommended FNNRA technique in comparison to the corresponding FNN-Res-net configuration. Derived from the discoveries, it can be deduced that the employment of the FNNRA technique incorporating 19 convolutional layers notably results in enhanced effectiveness by diminishing error frequencies within the examination dataset. Integrating shortcut links from layer h^1 to the present layer empowers the network to preserve both current and preceding input details in deeper layers. Nevertheless, in extremely deep layers, the parameter γ influences the necessity of retaining input details. Consequently, the amalgamation of input shortcut connections is inferred to enhance model efficacy.

The suggested FNNRA-HMM approach consistently attains a decrease in WER when juxtaposed with alternate methods across diverse trials. These revelations regarding the elevated precision of this technique predominantly underscore the simplified immediate conveyance of input specifics via the proposed dynamic links. Additionally, the weighted links within this network play a role in augmenting precision in contrast to intricate methods. Bars 11 and 12 demonstrate a substantial 4.435% decrease in absolute inaccuracies, accomplished by the proposed FNNRA technique when harmonized with wav2vec-U possessing a similar configuration.

4.7. Word recognition on CallHome

The dataset from CallHome functioned as the principal source of data in our trials to assess the efficacy of our speech recognition system. These records encapsulate English telephone dialogues spanning a broad spectrum of subjects, speakers, and speech modulations. On the flip side, the CallHome dataset encompasses spontaneous phone discussions covering more conventional and genuine speech manners. The datasets present a variety of speech segments for the meticulous evaluation of our models across different speech settings. In our experiments, we strictly followed a detailed preprocessing sequence for the CallHome dataset, encompassing standard steps like normalization, segmentation, and feature extraction. Acoustic characteristics such as MFCCs were derived from speech signals using frame-based extraction methods. These attributes were utilized as inputs for our models during both the training and testing phases.

We trained the proposed FNNRA approach alongside other similar techniques employing robust optimization algorithms like Stochastic Gradient Descent (SGD) or Adam. The model hyperparameters were carefully fine-tuned using techniques such as cross-validation and network exploration. Figure 11 delineates the comparison between our proposed methodology termed "FNNRA" and various intricate methods in speech recognition on the CallHome dataset. The objective of this evaluation was to gauge the performance and competitiveness of our strategy by assessing its efficacy against these methodologies. Objective metrics like WER were applied to gauge and juxtapose the effectiveness of diverse techniques on the CallHome test sets. Following an exhaustive scrutiny of the outcomes, it emerged that our proposed approach (emphasized by bar 6 from the top) attained commendable outcomes in contrast to other techniques. While certain methodologies like CNN-BLSTM and LSTM-ResNet outperformed, our approach notably shone in situations with restricted linguistic resources.

These empirical assessments furnish a profound insight into the robust points of our proposed FNNRA methodology in contrast advanced techniques, charting a course for advancements in speech recognition research.

5. CONCLUSION AND DISCUSSION

In brief, this investigation presents FNNRA, a novel technique addressing overfitting in situations with restricted training information. The distinctive trait of FNNRA lies in its capacity to effectively navigate fluctuations in speech signals and deliver efficient training with minimal data. The versatility of this approach is showcased through its initial blueprint, integrating up to 19 convolutional layers for precise extraction and examination of vital data while conserving it in profound layers. Both theoretical scrutiny and practical authentication affirm the efficacy of FNNRA across diverse data contexts. Assessments on standardized archives validate its competitive or superior efficiency when juxtaposed with contemporary advanced strategies in speech

recognition. Particularly, FNNRA displays promise in diminishing errors in word identification, establishing a basis for forthcoming strides in detection systems.

The discussion accentuates critical hurdles and perspectives unearthed in the exploration. Scalability and computational efficacy emerge as crucial factors, notably when grappling with more extensive and intricate datasets characterized by an array of audio attributes and acoustic terrains. Subsequent research to ensure the resilience of the approach when applied to expansive datasets remains imperative. Furthermore, delving into the utilization of FNNRA in varied data configurations and a spectrum of practical applications, like biometric authentication or medical image scrutiny, charts avenues for future revelations. Enhancing the depth of convolutional layers, analyzing alterations in connectivity structures, and managing trade-offs among computational resources, time efficiency, and precision stand out as pivotal focal points for forthcoming studies. While FNNRA showcases substantial progress in the accuracy of speech recognition and the adaptability to data, it also unveils numerous avenues for enhancing methodologies and broadening our insights beyond conventional speech and image datasets.

Engagement in this study contributes to propelling deep learning strategies in the auditory and visual domains, fortifying detection technologies. FNNRA's ingenuity lies in its adept handling of sparse training data, its resilience against overfitting, and its exhibition of competitive or superior performance across diverse data contexts, setting it apart from prevailing speech recognition methodologies.

FUNDING DETAILS

This research has been conducted without a budget.

DISCLOSURE STATEMENT

The authors of the report have disclosed that there are no potential conflicts of interest.

REFERENCES

[1] Li, J. “Recent advances in end-to-end automatic speech recognition”, APSIPA Transactions on Signal and Information Processing, 11(1), (2022). DOI:10.1017/ATSIP.2013.99.

[2] Ding, N., Jiaxin, G., Jing W., et. al. "Speech recognition in echoic environments and the effect of aging and hearing impairment”, Hearing Research pp. 108725, (2023). DOI: 10.1016/j.heares.2023.108725.

[3] Liu, A. H., Wei-Ning H., Michael A., et. al. "Towards end-to-end unsupervised speech recognition", IEEE Spoken Language Technology Workshop (SLT), pp. 221-228 (2022). DOI: 10.1109/SLT54892.2023.10023187

[4] Thomas, B., Samuel, K., and Salah K. "Efficient adapter transfer of self-supervised speech models for automatic speech recognition”, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7102-7106 (2022). DOI: https://doi.org/10.48550/arXiv.2202.03218 .

[5] Gupta, A., Puneet, G., and Esa R. "Fatalread-Fooling visual speech recognition models: put words on lips", Applied Intelligence, pp. 1-16 (2022). DOI: https://doi.org/10.1007/s10489-021-02846-w.

[6] Pawar, A. B., Pranav G., Mangesh G., et. al. "Challenges for hate speech recognition system: approach based on solution", International Conference on Sustainable Computing and Data Communication Systems, pp. 699-704 (2022). DOI: 10.1109/ICSCDS53736.2022.9760739

[7] Peng, Y., Siddharth D., Ian L., et. al. "Branchformer: parallel mlp-attention architectures to capture local and global context for speech recognition and understanding", International Conference on Machine Learning, pp. 17627-17643 (2022). DOI: https://doi.org/10.48550/arXiv.2207.02971.

[8] Yuvaraj, S., Abhishek B., William, P., et. al. "Speech recognition based robotic arm writing", Proceedings of International Conference on Communication and Artificial Intelligence: ICCAI 2021, pp. 23-33 (2022). DOI: https://doi.org/10.1007/978-981.

[9] Aditya, J., Kulkarni, G., and Shah, V. "Natural language processing", International Journal of Computer Sciences and Engineering, 6(1), pp. 352-357 (2018). DOI:10.26438/ijcse/v6i1.161167.

[10] Kumar, P., Saini, R., Roy, P.P., et. al. "Envisioned speech recognition using EEG sensors", Personal and Ubiquitous Computing, 22(1), pp. 185-199 (2018). DOI: https://doi.org/10.1007/s00779-017-1083-4.

[11] Chen, Z., Droppo, J., Li, J., et. al. "Progressive joint modeling in unsupervised single-channel overlapped speech recognition", IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 26(1), pp. 184-196 (2018). DOI: https://doi.org/10.48550/arXiv.1707.07048.

[12] Zeyer, A., Doetsch, P., Voigtlaender, P., et. al. "A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2462–2466 (2017). DOI: https://doi.org/10.48550/arXiv.1606.06871.

[13] Chan, W., Jaitly, N., Le, Q., et. al. "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964 (2016). DOI: 10.1109/ICASSP.2016.7472621.

[14] Kumar, L. Ashok, D., Renuka, K., et. al. "Deep learning based assistive technology on audio visual speech recognition for hearing impaired", International Journal of Cognitive Computing in Engineering 3, pp. 24-30 (2022). DOI:10.1016/j.ijcce.2022.01.003.

[15] Pingchuan, M., Haliassos, A., Lopez, A., et. al. "Auto-AVSR: audio-visual speech recognition with automatic labels", International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5 (2023). DOI: https://doi.org/10.1109/ICASSP49357.2023.10096889.

[16] Qiya, S., Sun, B., and Shutao, Li. "Multimodal sparse transformer network for audio-visual speech recognition", IEEE Transactions on Neural Networks and Learning Systems, (2022). DOI: 10.1109/TNNLS.2022.3163771.

[17] Hanan, A., Ullah, A., Ram, S., et. al. "Unsupervised automatic speech recognition: a review", Speech Communication, (2022). DOI: https://doi.org/10.1016/j.specom.2022.02.005.

[18] Chiu, C. C., Qin, J., Zhang, Y., et. al. "Self-supervised learning with random-projection quantizer for speech recognition", International Conference on Machine Learning, pp. 3915-3924 (2022). DOI: https://doi.org/10.48550/arXiv.2202.01855.

[19] Ambuj, M., Majumder N., Bharadwaj R, et. al. "A review of deep learning techniques for speech processing", Information Fusion, pp. 101-869 (2023). DOI: https://doi.org/10.48550/arXiv.2305.00359.

[20] Li, B., Sainath, T. N., Weiss, R. J., et. al. "Neural network adaptive beamforming for robust multichannel speech recognition", INTERSPEECH, pp. 1976–1980 (2016). DOI: http://dx.doi.org/10.21437/Interspeech.2016-173.

[21] Yiming, W., Li, J., Wang, H., et. al. “Wav2vec-switch: contrastive learning from original-noisy speech pairs for robust speech recognition”, International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7097-7101 (2022), IEEE. DOI: 10.1109/ICASSP43922.2022.9746929

[22] Ali, M. H., Jaber, M. M., Abd, S. K., et. al. “Harris hawks sparse auto-encoder networks for automatic speech recognition system”, Applied Sciences, 12(3), pp. 10-91 (2022). DOI: https://doi.org/10.3390/app12031091.

[23] William, P., Gade, R., Chaudhari, R. R., et. al. “Machine learning based automatic hate speech recognition system”, International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 315-318 (2022). DOI: 10.1109/ICSCDS53736.2022.9760959.

[24] Bharathi, B., Chakravarthi, B., R., Subalalitha C., et. al. “Findings of the shared task on speech recognition for vulnerable individuals in tamil”, Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 339-345 (2022). DOI:10.18653/v1/2022.ltedi-1.52.

[25] Heymann, J., Drude, L., and Haeb-Umbach, R. “Wide residual blstm network with discriminative speaker adaptation for robust speech recognition”, CHiME workshop, pp. 12-17 (2016).

[26] Leggetter, C. J., and Woodland, P. C. “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models”, Computer Speech & Language, 9(2), pp. 171–185 (1995). DOI: https://doi.org/10.1006/csla.1995.0010.

[27] Ramesh, P., and Wilpon, J. G. “Modeling state durations in hidden markov models for automatic speech recognition”, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, pp. 381–384 (1992). DOI: 10.1109/ICASSP.1992.225892.

[28] Justine, P. N, Kao, T., Zweig, G. “Discriminative duration modeling for speech recognition with segmental conditional random fields”, ICASSP, pp. 4476-4479 (2011). DOI: 10.1109/ICASSP.2011.5947348.

[29] Yu, S. Z. “Hidden semi-markov models”, Artificial Intelligence, 174(2), pp. 215–243 (2010). DOI: https://doi.org/10.6084/m9.figshare.

[30] Zoughi, T. Deypir, M. “Mathematical analysis of amres: unlocking enhanced recognition across audio-visual domains”, International Journal of Information Technology, pp. 1-20 (2024). DOI: https://doi.org/10.1007/s41870-024-01739-8.

[31] BabaAli, B. “A state-of-the-art framework for persian speech recognition”, Signal and Data Processing, 13(3), pp. 1–13 (2016). DOI:10.1109/IADCC.2009.4808998.

[32] Hadian, H., Povey, D., Sameti, H., et. al. “Phone duration modeling for lvcsr using neural networks”, Interspeech, pp. 20–24 (2017). DOI: http://dx.doi.org/10.21437/Interspeech.2017-1680.

[33] Senior, A., Sak, H., and Shafran, I. “Context dependent phone models for lstm rnn acoustic modelling”, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 19–24 (2015). DOI: 10.1109/ICASSP.2015.7178839.

[34] Li, D., and Platt, J. “Ensemble deep learning for speech recognition”, Interspeech. (2014). DOI: 10.21437/Interspeech.2014-433.

[35] Muhammad, F., Ohi, A. Q., Hamid, M., et. al. “A study on the challenges and opportunities of speech recognition for Bengali language”, Artificial Intelligence Review, pp. 1-25 (2022). DOI:10.1007/s10462-021-10083-3.

[36] Reitmaier, T., Wallington, E., Kalarikalayil Raju, D., et. al. “Opportunities and challenges of automatic speech recognition systems for low-resource language speakers”, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1-17 (2022). DOI: 10.1145/3491102.3517639.

[37] Rosa, D. L., Braaten, J., Kummervold, P., et. al. “Boosting norwegian automatic speech recognition”, Proceedings of the 24th Nordic Conference on Computational Linguistics, pp. 555-564 (2023). DOI: https://doi.org/10.48550/arXiv.2307.01672.

[38] Yu, Z., Daniel S., Han, P., et. al. “Bigssl: exploring the frontier of large-scale semi-supervised learning for automatic speech recognition”, IEEE Journal of Selected Topics in Signal Processing, 16(6), pp. 1519-1532 (2022). DOI: 10.1109/JSTSP.2022.3182537

[39] Sainath, T. N., Kingsbury, B., Soltau, H., et. al. “Optimization techniques to improve training speed of deep neural networks for large speech tasks”, IEEE Transactions on Audio, Speech and Language Processing, 21(11), pp. 2267–2276 (2013). DOI: 10.1109/TASL.2013.2284378.

[40] Felix, W., Kim, K., Pan, J., et. al. “Performance-efficiency trade-offs in unsupervised pre-training for speech recognition”, ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7667-7671 (2022). DOI: https://doi.org/10.48550/arXiv.2109.06870.

[41] Abdinabi, M., Khujayarov, I., Djuraev, O., et. al. “Automatic speech recognition method based on deep learning approaches for Uzbek language”, Sensors, 22(10), pp. 36-83 (2022). DOI: https://doi.org/10.3390/s22103683.

[42] Binbin, Z., Hang, L., Guo, P., et. al. “Wenetspeech: a 10000+ hours multi-domain mandarin corpus for speech recognition”, ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6182-6186 (2022). DOI: https://doi.org/10.48550/arXiv.2110.03370.

[43] Sakshi, D., Kumar, S. S., Albagory, Y., et. al. “Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network”, Applied Sciences, 12(12), pp. 62-23 (2022). DOI: https://doi.org/10.3390/app12126223.

[44] Hardt, M., and Ma, T. “Identity matters in deep learning”, International Conference on Learning Representations, pp. 131-139 (2017). DOI: https://doi.org/10.48550/arXiv.1611.04231.

[45] Saxe, A., McClelland, J. L., and Ganguli, S. “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks”, International Conference on Learning Representations, pp. 18-24 (2014). DOI: https://doi.org/10.48550/arXiv.1312.6120.

[46] Afsharpour, P., Zoughi, T., Deypir, M., et. al. “Robust deep learning method for fruit decay detection and plant identification: enhancing food security and quality control”, Frontiers in Plant Science, 15, pp. 136-6395 (2024). DOI: https://doi.org/10.3389/fpls.2024.1366395.

[47] Petrov, S., Pauls, A., and Klein, D. “Learning structured models for phone recognition”, Proceedings of the Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 897–905 (2007).

[48] Bengio, Y. “Learning deep architectures for ai”, Foundations and Trends in Machine Learning, 2(1), pp. 1–127 (2009).

[49] Domingos, P. “The role of occam's razor in knowledge discovery”, Data Mining and Knowledge Discovery, 3, pp. 409–425 (1999). https://doi.org/10.1023/A:100986892989.

[50] Huang, L., Xu, J., Sun, J., et. al. “An improved residual lstm architecture for acoustic modeling”, International Conference on Computer and Communication Systems (ICCCS), pp. 101–105 (2017). DOI: 10.1109/CCOMS.2017.8075276

[51] Glorot, X. and Bengio, Y. “Understanding the difficulty of training deep feedforward neural networks”, AISTATS, 9, pp. 249–256 (2010).

[52] Martens, J., & Sutskever, I. “Training deep and recurrent networks with hessian-free optimization”, Neural Networks: Tricks of the Trade: Second Edition, pp. 479-535 (2012). DOI: https://doi.org/10.1007/978-3-642-35289-8_27.

[53] Celikyilmaz, A., Sarikaya, R., Hakkani-Tur, D., et. al. “A new pre-training method for training deep learning models with application to spoken language understanding”, Interspeech, pp. 3255–3259 (2016). DOI: http://dx.doi.org/10.21437/Interspeech.2016-512.

[54] Lee, H., Grosse, R., Ranganath, R., et. al. “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations”, International Conference on Machine Learning, 21, pp. 1–8 (2009). DOI: https://doi.org/10.1145/1553374.15534.

[55] rahman Mohamed, A., Dahl, G., and Hinton, G., “Acoustic modeling using deep belief networks”, IEEE Transactions on Audio, Speech, and Language Processing, 20(1), pp. 14–22 (2012). DOI: 10.1109/TASL.2011.2109382.

[56] Taylor, G. W., Fergus, R., LeCun, Y., et. al. “Convolutional learning of spatio-temporal features”, Lecture Notes in Computer Science, 16(6), pp. 140–153 (2010). DOI: https://doi.org/10.1007/978-3-642-15567-3_11.

Toktam Zoughi received her Ph.D. in computer engineering (artificial intelligence) from the Amirkabir University of Technology, Tehran, Iran, in 2019 and M.S.C in computer engineering (artificial intelligence) from Shiraz University, shiraz, Iran, in 2010. Since January 2020, she has been an assistant professor at Department of Electrical and Computer Engineering, Shariaty College, Technical and Vocational University (TVU), Tehran, Iran. Her research interests mainly focus on deep learning, machine learning, speech processing, image processing, and NLP. Her email address is: t.zoughi@shariaty.ac.ir.

Mahmood Deypir received his Ph.D. degree in 2011 and M.Sc degree in 2006 both from Shiraz University. He is interested in researching areas such as Data Mining and Pattern Recognition, and Network Security. He has published a number of papers in ISI journals and international conferences. His email address is: mdeypir@ssau.ac.ir.

FIGURE CAPTIONS

Figure 1. Block Diagram of the Speech Recognition Process.

Figure 2. The architecture of the proposed Flexible Neural Network with Recursive Architecture (FNNRA).

Figure 3. The suggested approach employs certain layers from the FNNRA and completely linked layers. The results from the fully connected layers represent probabilities related to the HMM state.

Figure 4. Training Steps Flowchart for the Proposed FNNRA Model.

Figure 5. Exploring the Impact of Window Size Increase on Phoneme Recognition Error in CNN and FNN Algorithms.

Figure 6. Comparative analysis of CNN and FNN algorithms, emphasizing the impact of increasing Fully Connected Layer depth.

Figure 7. The error rate of sound detection in relation to various γ values for the FNNRA technique for the TIMIT dataset.

Figure 8. The error rate of sound detection in relation to various $constant$ values for the FNNRA technique within the TIMIT dataset.

Figure 9. Phoneme error rate (PER) comparison of different proposed method structures on the FarsDAT dataset.

Figure 10. Contrasting different configurations of the proposed approach with sophisticated methods on the TIMIT dataset.

Figure 11. Contrast among varied configurations of the suggested technique and cutting-edge methodologies in the CallHome dataset.

TABLE CAPTIONS

Table 1. Comparison of Parameter Counts between the Proposed Model and Other Models.

Table 2. FARSDAT Dataset information [31]

Table 3. Information about the TIMIT dataset

FIGURES

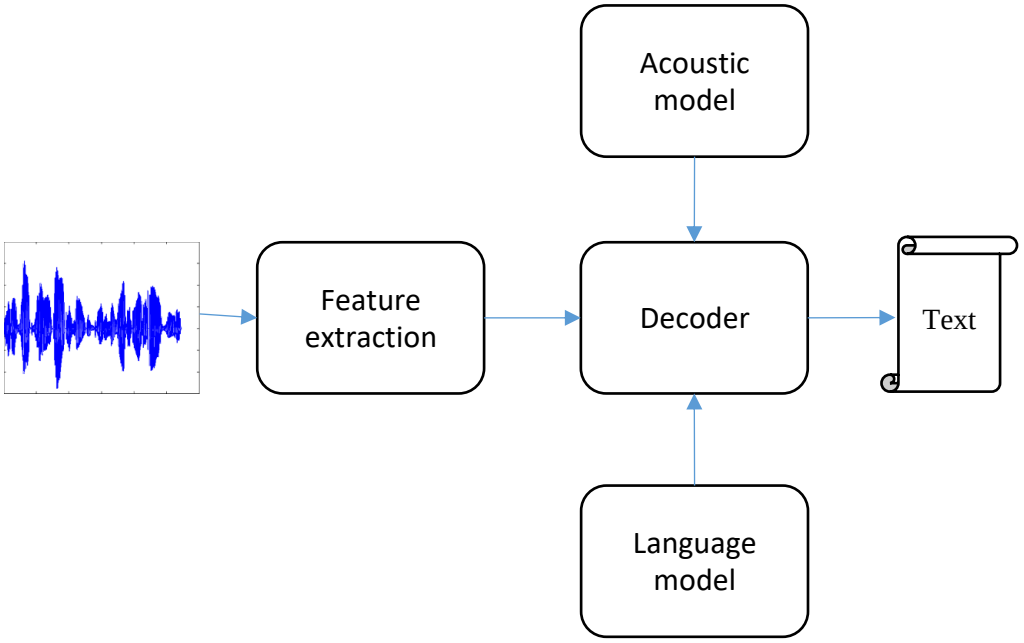


Figure 12. Block Diagram of the Speech Recognition Process.

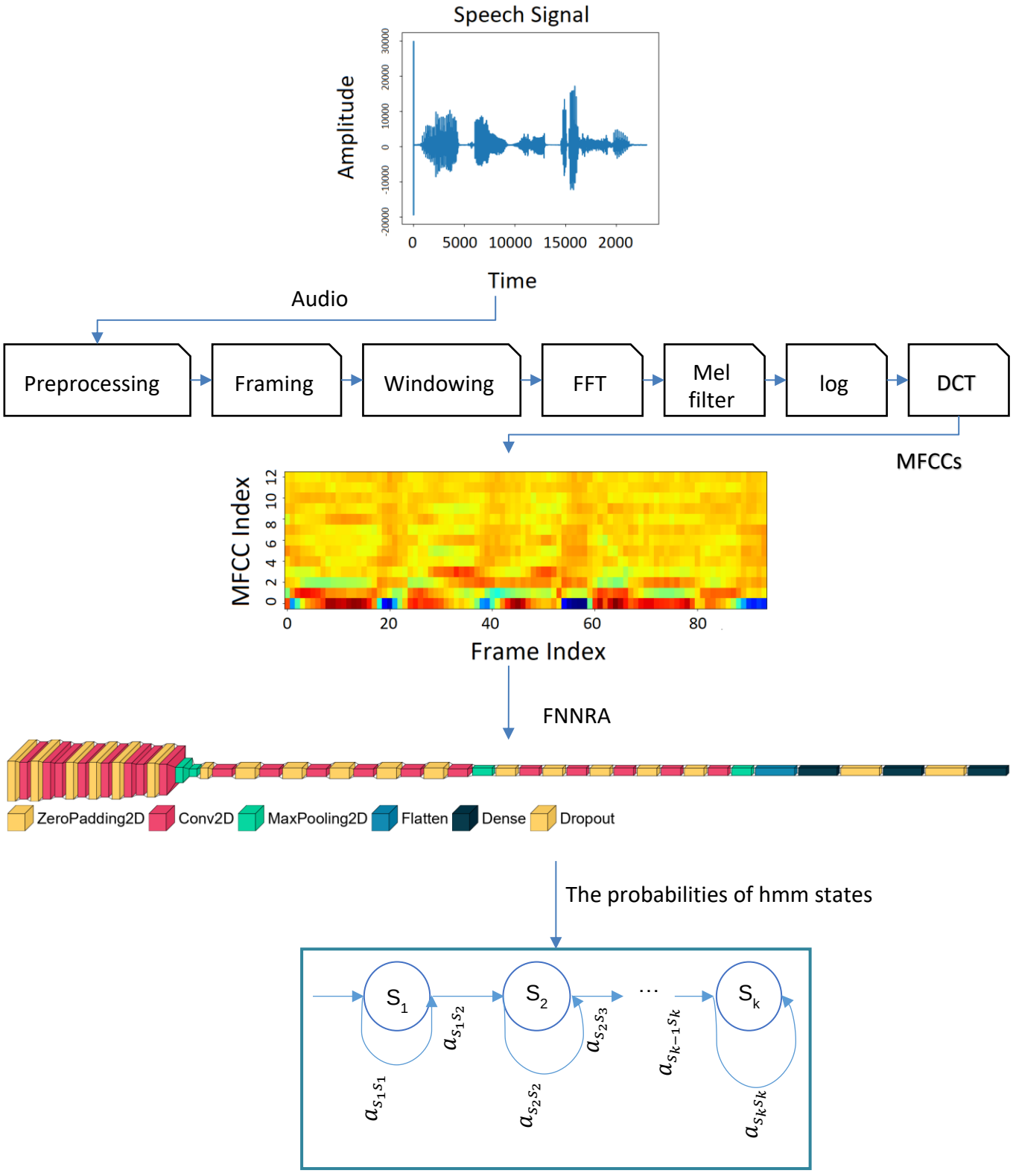


Figure 13. The architecture of the proposed Flexible Neural Network with Recursive Architecture (FNNRA).

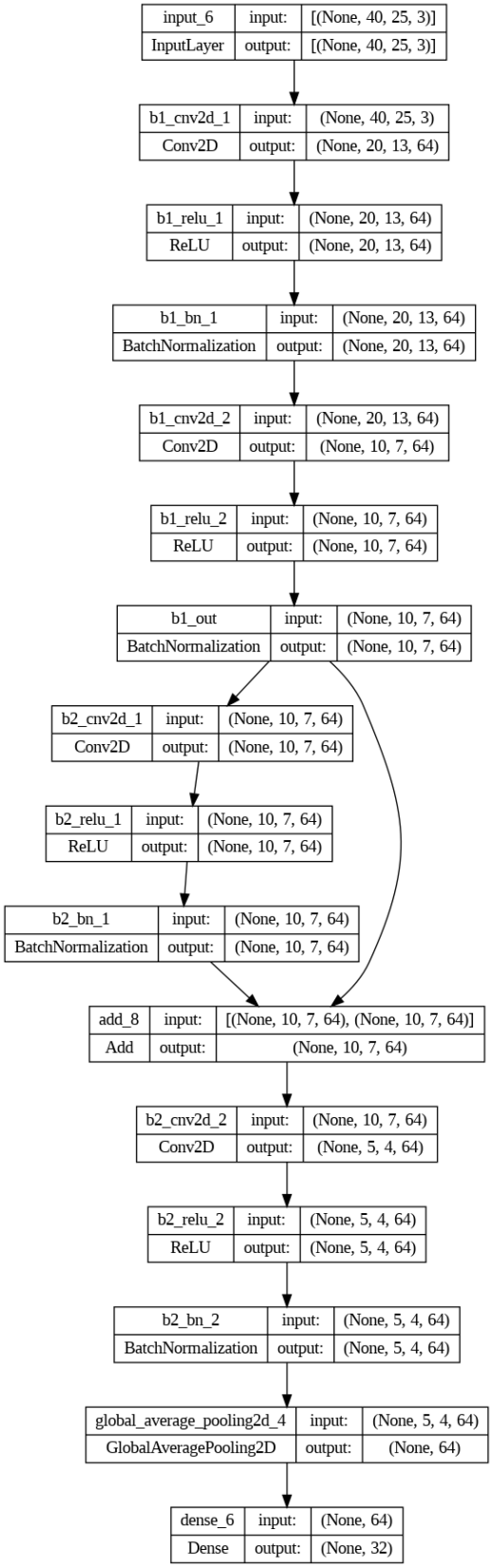


Figure 14. The suggested approach employs certain layers from the FNNRA and completely linked layers. The results from the fully connected layers represent probabilities related to the HMM state.

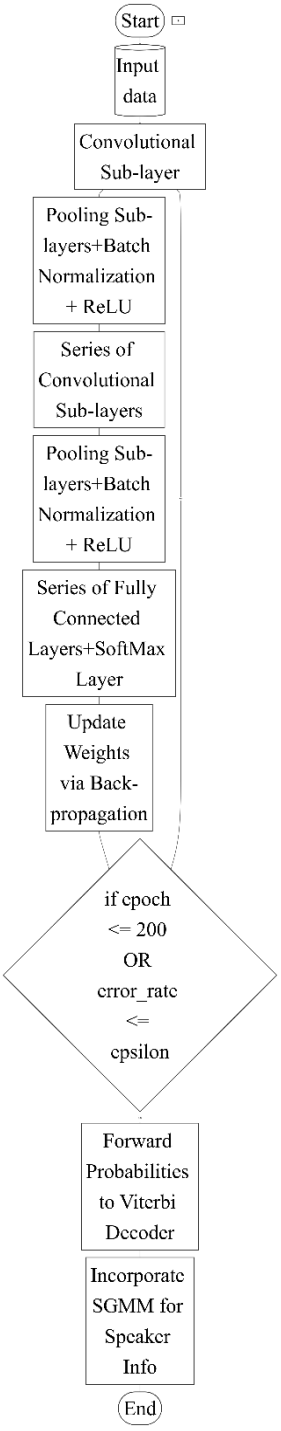


Figure 15. Training Steps Flowchart for the Proposed FNNRA Model.

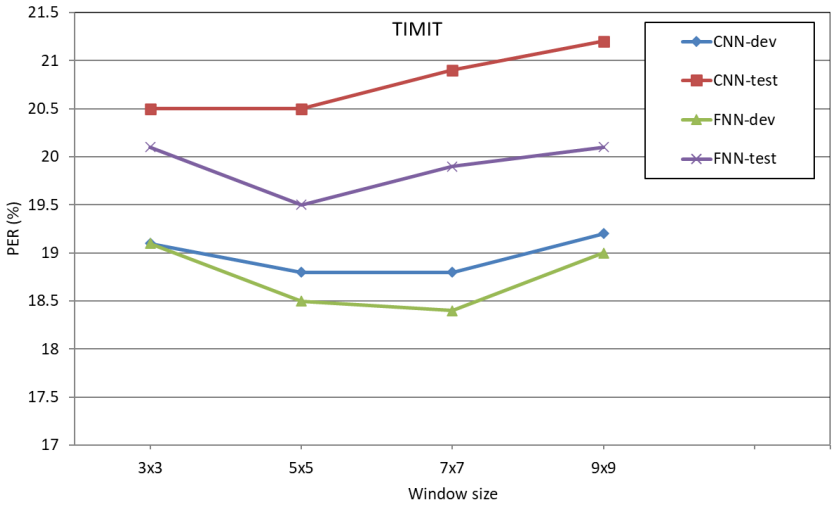


Figure 16. Exploring the Impact of Window Size Increase on Phoneme Recognition Error in CNN and FNN Algorithms.

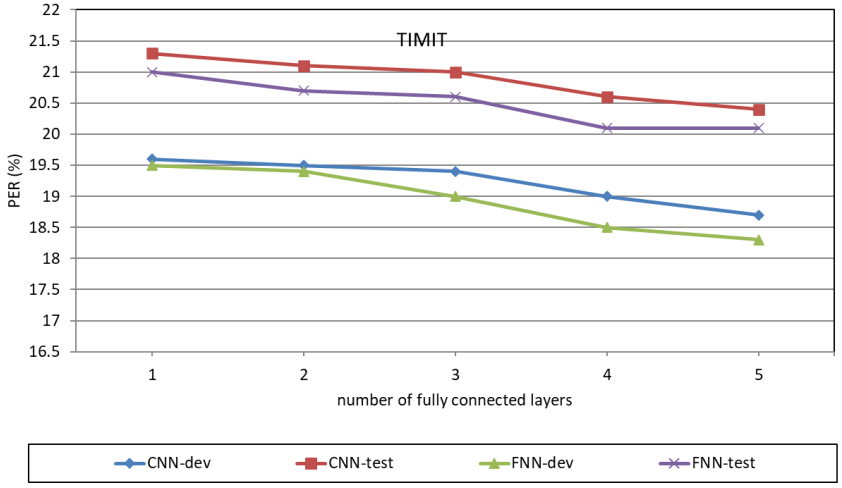


Figure 17. Comparative analysis of CNN and FNN algorithms, emphasizing the impact of increasing Fully Connected Layer depth.

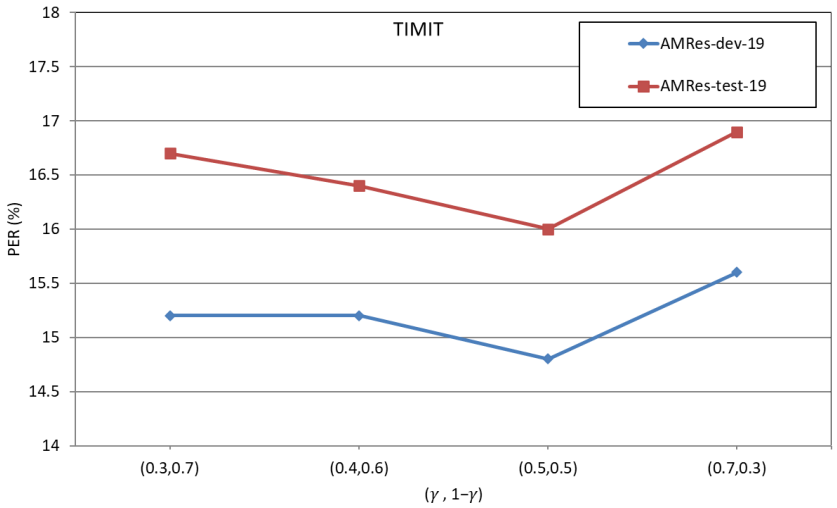


Figure 18. The error rate of sound detection in relation to various γ values for the FNNRA technique for the TIMIT dataset.

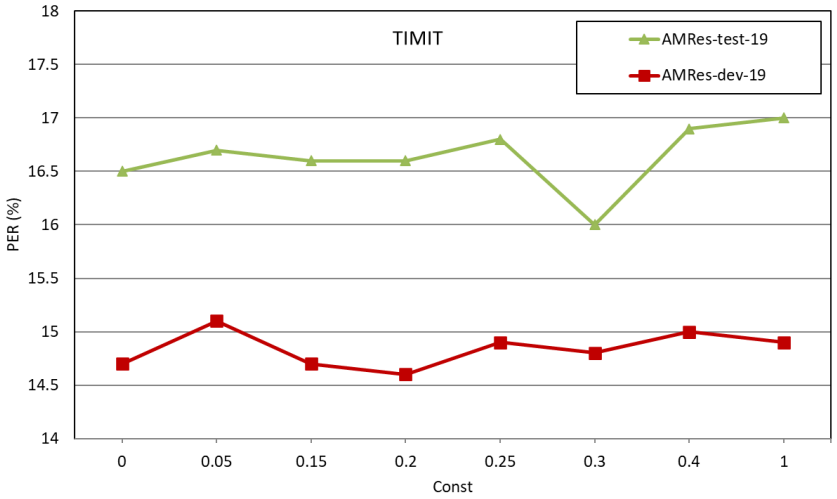


Figure 19. The error rate of sound detection in relation to various $const$ values for the FNNRA technique within the TIMIT dataset.

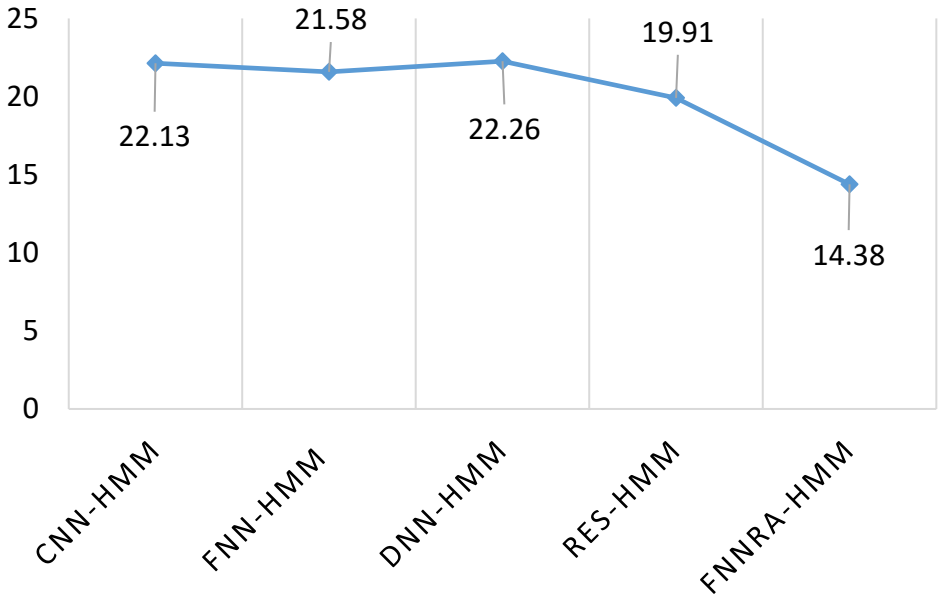


Figure 20. Phoneme error rate (PER) comparison of different proposed method structures on the FarsDAT dataset.

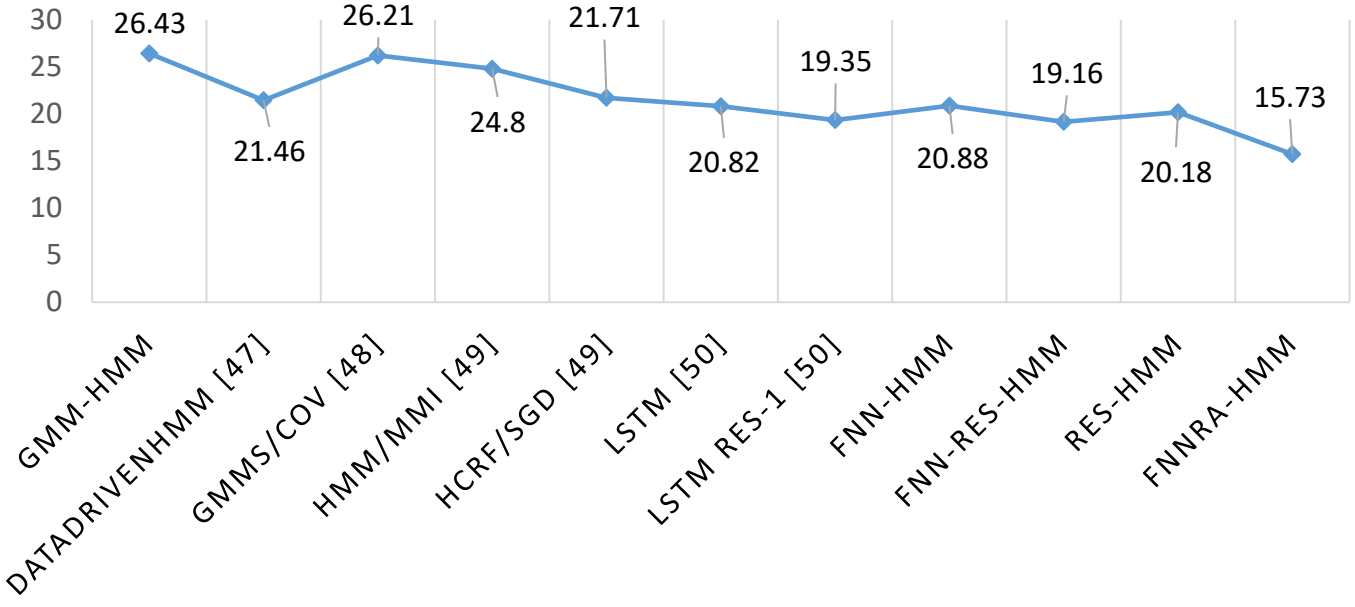
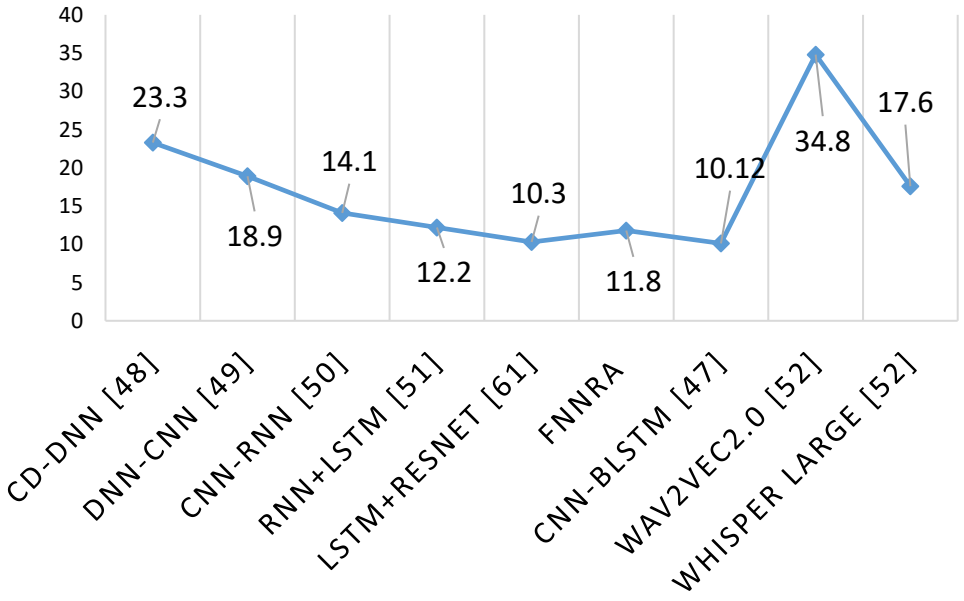


Figure 21. Contrasting different configurations of the proposed approach with sophisticated methods on the TIMIT dataset.



1.

Figure 22. Contrast among varied configurations of the suggested technique and cutting-edge methodologies in the CallHome dataset.

TABLES

Table 4. Comparison of Parameter Counts between the Proposed Model and Other Models.

	Network model	Number of layers	Number of parameters	Number of parameters	Number of parameters
1	**Deep neural network**	10FC	$n \times h + 10h^2$	$420h + 10h^2$	10.92 M
2	**Convolutional neural network**	7Conv+3Fc	$7h \times f \times k + 3h^2$	$343h + 3h^2$	3.49 million
3	**Flexible Neural Network**	7Conv+3Fc	$9h \times f \times k + 3h^2$	$441h + 3h^2$	3.5 million
4	**Res**	7Conv+3Fc	$7h \times f \times k + 3 + 3h^2$	$346h + 3h^2$	3.50 million
5	**Flexible Neural Network-Res**	7Conv+3Fc	$9h \times f \times k + 7 + 3h^2$	$448h + 3h^2$	3.60 million
6	**Flexible Neural Network with Recursive Architecture**	7Conv+3Fc	$9h \times f \times k + 3 + 3h^2$	$444h + 3h^2$	3.60 million

Table 5. FARSDAT Dataset information [31]

Data set	Speakers Numbers	Utterance Numbers	Hours numbers
Train	224	3994	2.91
Development set	50	475	0.39
Test set	30	287	0.23
Total	304	4756	3.54

Table 6. Information about the TIMIT dataset

	#Speakers	#Utterance	#Hours
Train	462	3696	3.14
Complete test	168	1344	0.81
Core test	24	192	0.16
Total	654	5232	4.11