Received: August 7, 2020. Revised: September 2, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Dysarthric Speech Recognition using Convolutional Recurrent Neural Networks

Hussain Albaqshi[1] **Alaa Sagheer**[1,2]*

[1]*Department of Computer Science, College of Computer Sciences and IT, King Faisal University, Saudi Arabia*
[2]*Department of Computer Science, Aswan University, Egypt*
* Corresponding author's Email: asagheer@kfu.edu.sa

Abstract: Automatic speech recognition (ASR) transcribes the human voice into a text automatically. Recently, ASR systems has reached, almost, the human performance in specific scenarios. In contrast, dysarthric speech recognition (DSR) is still a challenging task due to many reasons including unintelligible speech, irregular phonemes articulation, along with scarcity and heterogeneous of data. Most of the existing DSR works are employed the ASR systems that trained on an unimpaired speech to recognize such impaired speech, which of course is impractical and inefficient. In this paper, we developed a deep architecture of the convolutional recurrent neural network (CRNN) model and compared its performance with the vanilla convolutional neural network (CNN) model. We train both models using the samples of the Torgo dataset, which contains a mixed of impaired and unimpaired speech data. The experimental results show that the CRNN model attains 40.6% against 31.4% for the vanilla CNN. This indicates the effectiveness of the proposed hybrid structure of the CRNN to improve the recognition of dysarthric speech.

Keywords: Dysarthric speech recognition, Speech disorder, Torgo database, Convolutional neural networks, Recurrent neural.

1. Introduction

Dysarthria is a neuro-motor articulation disorder disease result in weakness of the speech muscles at the human, such as tongue and lips [1]. It caused as a result of many reasons including paralysis, poor coordination, and weakness of the muscles that produce speech. It may also result as a side effect of stroke, a Parkinson's disease, and a cerebral palsy or any traumatic brain injury [2]. The person with dysarthria is unable to talk regularly where speech will be nearly unintelligible and phonemes articulation will be irregular. For this reason, it is reported that dysarthric speech is slower than a regular speech by 15 times since tongue, lips, and jaw are difficult to move as in a normal speech [3, 4].

In most cases, dysarthria is accompanied by a physical disability with limited body movements and uncontrolled coordination, leading eventually to a difficulty in using communication applications that are based on a joystick or a keyboard. All these aspects make those people with dysarthria facing difficulties and isolation in their life due to difficulty of communication with those who are around along with limited interaction with electronic devices including computers and phones [5].

In the recent years, with the great progress in developing automatic speech recognition (ASR) systems, broad range of commercial applications, where ASR as user interface, have become ever more useful and pervasive. Several articles in the literature deployed the ASR systems on dysarthric speech datasets, which yield a very poor performance [5]. Certainly, any ASR system trained on un-impaired speech will not be suitable to be validated using dysarthric speech data in the scope of the large mismatch of acoustic and articulatory characteristics between dysarthric and normal speech [6, 7]. In other words, ASR systems are ineffective and impractical to process dysarthric speech recognition (DSR) systems [5]. Consequently, appropriate DSR systems specifically tailored to persons with dysarthria would be more efficient and practical than ASR systems.

The best way to develop such a DSR system will be accomplished by training the system using datasets include both normal speech and dysarthric speech [8, 9]. Most of the existing works that oriented to solve the DSR problem have focused on capturing the acoustic cues or features of the dysarthric speech. To this target, various extraction techniques for acoustic features have been used including hidden Markov models (HMM) [10], and Gaussian mixture model (GMM) [11].

However, with the emerging of deep learning approaches, DSR can be improved if it is implemented using deep neural networks (DNN) models rather than HMM and GMM. One of the advantages of using DNNs is that the inputs can be raw data, such as pixels of an image, rather than extracting specific input features as the case in GMM and HMM. Another major disadvantages of GMM and HMM is that they fail to model long-term dependencies exist in speech signal [8, 12].

Convolutional neural networks (CNN) and recurrent neural networks (RNN) are common deep neural networks models used widely to process different kinds of signals such as speech and image [13, 14]. In this paper, we merge the two models to have the convolutional recurrent neural network (CRNN). The CRNN combines the structures of both original models and, therefore, acquires the benefits of both models to ensure efficient processing to DSR. The developed CRNN consists of two CNN blocks; each has three layers, namely, a convolution layer, a Relu layer, and a dropout layer. The second CNN block is similar to the first block, except that we replaced the covolution layer with a unidirectional RNN block. The RNN block includes two RNN layers include 60 neurons in total. In the proposed structure, the convolutional layer of the first CNN block will extract the local features of the input speech, whereas the RNN block will extract the overall or global feature structure by aggregating the local features extracted by the convolution layer. Briefly, the CNN plays the function of features extractor whilst the RNN plays the function of features integrator.

The rest of this paper is organized as follows. Section 2 discusses the related work to this paper. The problem statement of this paper and our motivation to solve it are given in section 3. Section 4 shows the details of the experimental settings and the methods that used throughout the paper. The experimental results are shown and discussed in section 5. Section 6 concludes this paper.

2. Related work

The related research of DSR using the traditional artificial neural networks (ANNs) started since more than two decades ago. Specifically, in 1990s, statistical causal models used for dysarthric speech recognition. The importance of this figure of benefit is that the intrinsic causal relationship between the intelligibility of dysarthric speech and the responses of speech recognition systems are made clear through their linguistic counterpart or through different phonemes [15]. In another trial, different words have been recorded form a number of individuals with dysarthria. In this experiment, 20 words out of 50 have been selected and repeated 22 time used with two multilayer neural networks [16]. One of the networks had the fast Fourier transform coefficients as inputs; the other network had the format frequencies as inputs. The data presented in this experiment have implications for individuals with dysarthria other than those with cerebral palsy.

A deep belief network (DBN) pre-trained on the UAspeech database [17] that includes about 9 and 3 hours per speaker in the training and testing sets, respectively, and almost of explored different methods for generating speaker dependent pronunciations [18]. A combination model of maximum a posteriori (MAP) and the maximum likelihood linear regression (MLLR) applied on the QoLT 2012 and KPOW database. The performance of three severity models were better than the baseline model, relatively reducing the word error rate (WER) by 17.9%, 17.2%, and 10.4% for universal, mild, and mild-to-severe models on average, respectively [4]. Other serious trials have continued, until the trail that ended with the development of the Torgo dataset [19], which adopted in this paper. This datatset contains the speech data of eight individuals with dysarthria and seven individuals without dysarthria, more details about this dataset will be provided in section 4.1. In 2011, a three-level cascaded adaptation procedure was applied on Torgo [20]. This three-level procedure consists of MLLR adaptation and MAP estimation that adapted a speaker independent model to the characteristics of the speaker's vocal. Using these two techniques, the WER was reduced. The pronunciation lexicon adaptation (PLA) was used and reduced the error rate further, where it showed a clear efficiency in the long utterances relatively.

In 2013, based on the digital short-time Fourier analysis, a phase vocoder applied on the Torgo dataset and modified the acoustics of dysarthric speech system by Gaussian mixture mapping [21]. In this system, FestVox implementation has used a method to resynthesize and pitch the feature extraction. FestVox system has trained the parameters for this model using 24th-order cepstral coefficients with a standard expectationmaximization approach and 4 Gaussian components. Gaussian mixture model transformed the dysarthric speech, purely synthetic speech and traditional HMMs trained with large amounts of data from the general population have used with evaluated utterances in each proposed transformation result [21].

In 2014, the impaired TIMIT dataset and the unimpaired Torgo dataset used to build a speaker independent model using the sample of the Nemours database [22]. Using this database, the author of [23] built a MLLR and a constrained MLLR (C-MLLR) adaptation models that produce a single Gaussian mixture model. When a speech model of unimpaired used, CMLLR technique performs better than MLLR. TIMIT speaker independent model for the mildly impaired speech is more accurate, while on the Torgo database the system performs well in recognizing the impaired speech for severely impaired and moderately cases. Breiefly, both TIMIT and Torgo showed that MLLR technique high WER than the CMLLR technique [23].

In 2015, a DSR system is developed using the support vector machine (SVM), the linear discriminant analysis (LDA), and the and k-nearest neighbor (kNN) classifiers [24]. The best feature were selected based on unweighted average recall from two pathological speech sub-challenge, the Torgo and the NKI CCRT speech corpus [24]. The classification performance using SVM showed the best performance by smoothed posterior score fusion of subsystems. In most cases, the posterior smoothing results is improved, except prosody subsystem case and the LDA classifier with feature-level fusion and the SVM classifier with feature-level fusion only in terms of classification accuracy that is unweighted [24].

In 2016, using the features of Mel-Frequency Cepstral Coefficients (MFCC), a small portion of the Torgo dataset is trained using both classical and modern neural network architectures. In this experiment, a DNN-HMM hybrid neural network architectures and GMM-HMM classical architectures have been used to compare with other DNNs, where the hybrid DNN-HMM showed the best performance [25]. The SVM, GMM, and the hybrid GMM/SVM systems used to test and compare in the assessment of a dysarthric speaker identification context. Relevant features used in both techniques based on MFCCs and distinctive auditory-based cues where different front-end processing used with SVM. Correct and high classification rate achieved by GMM compared to SVM. GMM/SVM has achieved best performance. Both Nemours and Torgo databases have used with different and changing durations that cannot process effectively by SVM [26].

The Torgo database and MOCHA-TIMIT with scattering coefficients, MFCCs comparison, wavelets and vocal 'tract variables' to phonological features used deep-belief networks and sum-product networks (SPNs). Through the use of an SVM classifier, the relationships between acoustic features three types and articulatory configurations are sought. For more accurate classification, over a broad array of phonological provided by MFCCs. Acoustic-articulatory inversion applied DBNs, but aspects interested several uniquely by SPNs, including a function of partition that is guaranteed given certain limitations of the network structure to be tractable. Although DBNs are less accurate than SPNs when using scattering transforms, very similar results have got by the more recent of SPN methods and the DBN [27].

In 2017, DBNs are applied again on Torgo to predict the posterior probabilities of the states in the RBM greedily as layered pre-trained and HMM to build speech as decoder with utilizing Weighted Finite State Transducers framework. Using DBNs returned better result where trained model tends to perform better when intelligibility scores have been higher by the test speakers [28]. Korean Phonetically Optimized Words (KPOW) databas, Korean Phonetically Balanced Words (KPBW) database, Korean Phonetically Rich Words (KPRW) database and SI dysarthria adaptation were used for dysarthtic speech recognition with KL-HMM and compared with GMM-HMM and DNN-HMM. The framework of KL-HMM showed that is effective for dysarthric speakers to improve the performance [29].

In 2018, a speaker with specific acoustic models trained on Torgo by tuning different parameters of acoustic model, using cepstral features normalized by speaker and building sequence discrimination strategies and dropout with complex DNN-HMM models and using generalized distillation framework to improve speakers of dysarthric with severe and severe-moderate speakers on control and dysarthric speech. For moderate and mild dysarthric speakers, the DNN of distilled student did not give any performance gains [5]. With extracting features using the reflection coefficients of perceptual linear prediction (PLP), MFCC, and filter bank, the HMMs showed better results provided by PLP and MFCC that applied on samples of six dysarthric speakers of Torgo speech where PLP is the most suitable [30].

3. Problem statement and motivation

As there are big numbers of individuals with dysarthric over the world, they are facing problems in their life and communications with their community because of their speech, and the experiments of dysarthric speech recognition have not enough improved the accuracies in this domain. Also, as the first author of this paper, is one of those individuals who are having a special need with dysarthria and facing an issue in his personal life when he communicates with the community, where his speech is not clear and non-understandable for everyone. All these reasons, and more, are motivated us to propose this research and address the improvement of DSR to help those people to be intervene with their surround and facilitate their life

The purpose of this paper is to find and apply suitable machine learning algorithms to improve the DSR to convert their speech to be clearer and more understandable for other people. In general, most of the algorithms used to improve DSR did not achieve satisfactory results, as mentioned in the literature review section. Most of the current state-of-the-art techniques are using ASR approaches to treat DSR problems. These techniques either show results with low accuracies or show results with high accuracies but on a part of the DSR corpus not a complete set of corpus. A special solution for DSR problem should be developed mainly for individuals with dysarthria, who usually are talking slowly. Furthermore, the existing state-of-the-art classical techniques, such as HMM and GMM, are biologically implausible and have excessive power consumption.

Inspired by the reported advantages of CNN and RNN models, the proposed CNN that combined with RNN improves the performance of DSR problems and adapt with DSR features better than the original CNN. The main thrust of this paper is to study the speech disorder of people with special needs who have neural weakness or feel some difficulty in controlling their nerves. To this target, we develop the CRNN in a Python environment using the Torgo database that contains samples of audio files of a number of single English words. The target is to improve the speech of individuals with dysarthria to be more understandable when they talk any English words. Applying this system to the Arabic language is one of our future targets.

4. Materials and methods

4.1 The Torgo database

The Torgo dataset contains the speech data samples of eight (three female and five male) individuals with dysarthria and seven (three female and four male) without dysarthria [19]. Each individual, either with or without dysarthria, recorded his/her data collection of words as array microphone and head-worn microphone, a sample is shown in Fig. 1. The data samples of Torgo are grouped by gender, per person, and a number of sessions for each person. Each session of each parson has audio files either array microphone or head-worn microphone where many of them are words and phrases.

Most of these data samples are labeled, which are grouped together in around 530 classes of labels of dysarthric speech. Most of these labels are words and other labels are sentences and phrases. The data objects (words, phrases, and sentences) are varied in the length of audio clips, where there are a number of objects has a few numbers of clips per class. Only the following label of classes out of the 530 classes have between 30 to 50 clips per class, and other classes have less than 30 clips per class as shown in Table 1.

Table 1. Object classes and their most frequencies in the Torgo dataset

No.	Class label	No. of clips per class
1	relax your mouth in its normal position	59
2	Sip	55
3	Xxx (just a sound)	52
4	Sigh	52
5	Air	45
6	Knew	44
7	Slip	43
8	Beat	35
9	Chair	35
10	Leak	35
11	Warm	34
12	Storm	34
13	Spark	33
14	say "Ah" for 5 seconds	32
15	Feed	32
16	Feet	32
17	Swarm	32
18	Know	30
19	Witch	30

In our experiments of this paper, we focused on the words that has large number of frequencies, as shown in Table 1 except the words number 1, 3 and 14, in total with 16 words. We divided these 16 data samples as 80% for training subset, 20% for testing subset. In the results section we will show the results of both phases; training and testing using the standard CNN model and the proposed CRNN model.

4.2 Hardware and software platforms

All experiments of this paper are implemented on a laptop equipped with VMware of Ubuntu 18.04.3 operating system. The processor of the VMware is Intel Core™ i7-3632QM CPU @ 2.20GHz, 2.9 GB RAM, x64 based processor under python 3.5.2 software environment. For the DNNs models, the Tensorflow library was used with scikit-learn and librosa libraries as back-end.

In our experiments we applied Mel Frequency Cepstral Coefficient (MFCC) to extract the acoustic features from the audio clips with setting samples rate as 16000, clip duration as 1000 milliseconds, Duration of frequency analysis window as 30 milliseconds and window stride (which shows how far we move between frequency windows) as 10 milliseconds where we set range to randomly shift the training audio as 100 milliseconds. We reshaped the features to be in fixed size of two dimensions.

4.3 Deep neural network models

The proposed model CRNN is the combination of two DNN models, namely CNN and RNN. In the following, we describe both of these models and, then, the proposed model.

4.3.1. Convolutional neural networks (CNN)

The CNN is a common deep neural network model used widely to process different kinds of signals such as speech and image. It is a feedforward neural network architecture inspired by the natural visual perception mechanism of the human beings [31]. It consists of multi-layers of two repeated layers, namely, the convolution layer and the pooling layer, of course beside the input layer. The original CNN has been successfully employed in many ASR system due to its ability to extract local speech features through the function of the repeated convolution and pooling layers, as shown in Fig. 2.

In the experiments of this paper, we applied the CNN to compare with the model CRNN. The employed CNN block consists of three repeated layers: the first layer includes a convolution layer using weight (filter) with standard deviation 0.01 and size as $20\ \times 8 \times 64 \times 1$ where 20 is the filter height, 8 is the filter width, 64 is the filter number (i.e. count), and setting the stride with 1. Then we added the results with zero of biases b as in the Eq. (1)

Figure. 1 A sample from the torgo dataset (this picture is adapted from the original article [19])

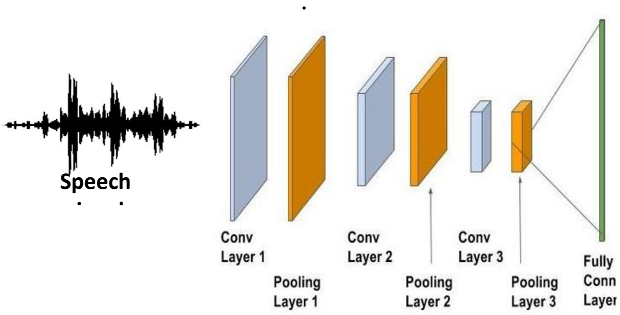


Figure. 2 The standard CNN block

$$y = w \times x + b \qquad (1)$$

Where y is the output of the first layer of the CNN, *w* is the filter of CNN, *x* refers to the input signal.

After that, we applied the ReLU activation function and the dropout function. The output of the dropout layer is used as input to the pooling layer with filter and stride as $1 \times 2 \times 2 \times 1$ and the output of max pooling used as input to the following, i.e. second, convolution layer with filter of size $10 \times 4 \times 64 \times 64$ as four dimensions with standard deviation 0.01 and zeros of biases. Finally, we flatten the shape of the second convolution layer to be in one dimension and applied Eq. (1) on the second convolution where x is the results of the second convolution layer, w is the weight of the second layer or filter with random values from a truncated normal distribution of (results of second CNN × number of labels) and b is the bias with zero value. The last (unrepeated) layer is the fully connected layer, which produce the output of CNN block.

4.3.2. Recurrent neural networks (RNN)

The RNN is another deep neural network model able to process data in a chronological order that is often has a varied length such as speech [32]. RNN is able to capture the key characteristics of dysarthric speech by modelling long term temporal structures. As each word, or a phoneme, in a speech should has connectivity perception, which is a shortcoming of the classical models or shallow neural networks, RNN can handle the issue in terms of its recursive dependency architecture [33]. For each input sequence, RNN performs the same task, where each single output is dependent on the previous computation, see Fig. 3.

4.3.3. The proposed CRNN model

The In the proposed CRNN architecture, we applied two blocks of the CNN model with one output layer. The first CNN block includes three layers; namely, convolution layer, Relu layer, and dropout layer. In the convolution layer we applied the Eq. (1) in subsection 4.3.1, where x is the input as a MFCC feature, w is the weight used Xavier initialization with size $10 \times 4 \times 48 \times 1$, valid padding, 2×2 stride, and b is zero of bias. Then, we applied the Relu activation and a dropout on the Relu layer output. The second CNN block is similar to the first block, except that we replaced the coevolution layer with a unidirectional RNN block. The RNN block includes two RNN layers include 60 neurons in total. Each neuron in the RNN has normalized to Gated Recurrent Unit (GRU) cell. After that, we applied the Eq. (1), x is the results of RNN and w is weight with standard deviation 0.01. Then we apply the Relu activation function on the result of RNN and a dropout on the result of Relu function. An overall visual representation is depicted in Fig 4.

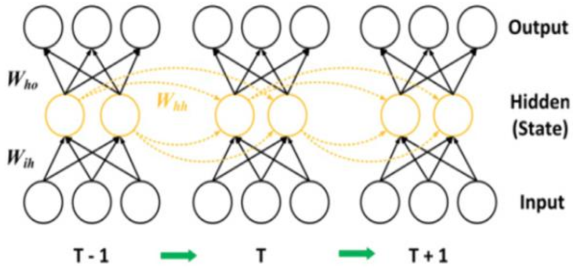


Figure. 3 The processing of time sequence in the standard RNN model

5. Experimental results

The main objective of the proposed experiment is to investigate the impact of adding the convolution layer in the standard CNN model into the standard RNN model. For all models that employed in the experiments, we applied a cross entropy mean, softmax of a cross entropy with logits and the ADAM optimizer [35] to calculate the confusion matrix of expected results as labels and predicted results as predictions. Our experiment is divided in two phases, training phase and assessment phase. In the training phase, we trained both models on the training subset of the Torgo dataset. The assessment phase compress of two sessions, in the first session we reused the training subset of data to check the validity of each model.

After multiple experimental trials, we found that for both models, the training accuracy have increased at the first 300 steps and reached 100% and saturated at this level, as shown in Fig. 5. Figs. 6 (a) and (b) shows the overall performance of both models until saturation using the training data subset applied on the selected set of 16 dysarthric words.

In the second session, where we used the testing data subset, we merged between the training phase and assessment phase in order to track the performance of each model and investigate the impact of training phase on the assessment phase. To perform this investigation, we make an assessment step at every2,000 training steps through different values of learning rates. We found that the performance of each model improves as we increase the number of training steps. CNN reached the saturation level, i.e. no improvements, after 16000 training steps, whereas CRNN reached to this level after 22000 training steps. Tables 2 shows the assessment accuracies for each model using the testing subset data at each step.

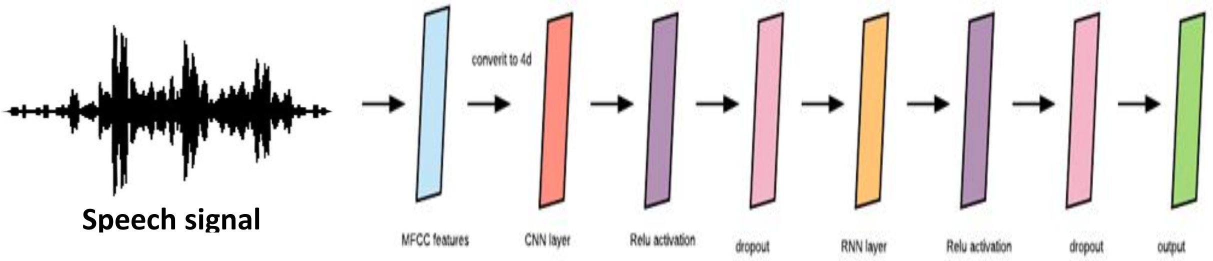
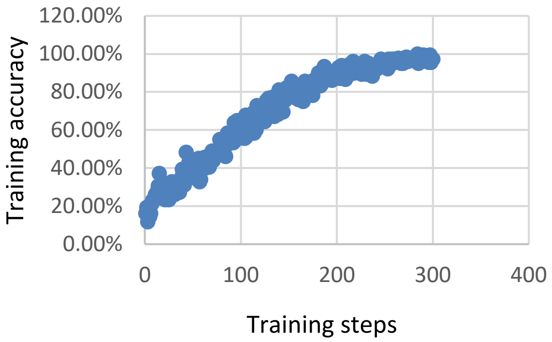
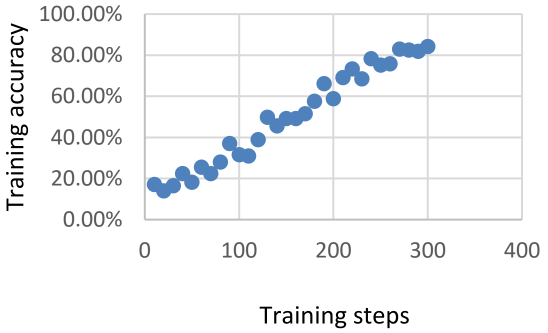


Figure. 4 The proposed CRNN architecture



(a)



(b)

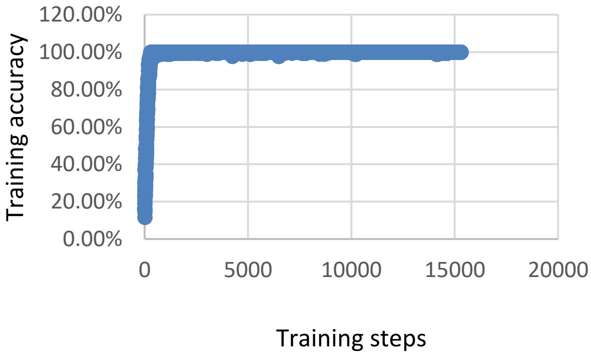
Figure. 5 Performance via the first 300 step of: (a) CNN (b) CRNN

As we see in Table 2, the assessment performance of the CNN model starts with accuracy value as 28.6% after 2000 training step. The performance improves after 4000 training steps to reach the accuracy level as 31.4%. After that, as we increase the number of training steps, we did not find any impact on the CNN performance, which means that CNN is saturated. On the CRNN side, it starts with accuracy value as 31.2% after 2000 training step. The performance improves after 6000 training steps to reach the accuracy level as 34.4%. The improvement is continued until 22000 training steps with accuracy 40.6%. After that, as we increase the number of training steps, we did not find any impact on the CRNN performance, which means that the CRNN is saturated and no further improvement.

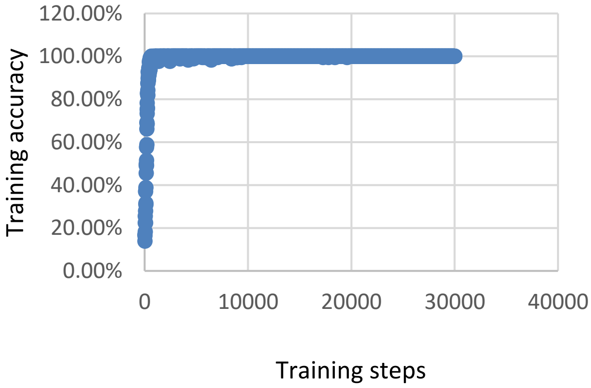
We can interpret the well performance of the proposed CRNN compared to the standard CNN in the scope of CRNN architecture. As the CRNN model includes multiple RNN units, where each can read the acoustic input data independently as a sequence of words, then this enable for better performance than CNN. In the mean time, in order to CRNN performs well, it requires plentiful training steps more than CNN. Also, this is a natural in the scope that the original RNN includes essentially intensive connections among its cells, which certainly requires intensive training session as well.

Table 2. Assessment performance of CNN and CRNN via different training steps

CNN		CRNN	
Training steps	Accuracy	Training steps	Accuracy
2000	28.6%	2000	31.2%
4000	31.4%	6000	34.4%
10000	31.4%	12000	37.5%
14000	31.4%	20000	37.5%
15300	**31.4%**	22000	**40.6%**



(a)



(b)

Figure. 6 Overall performance until saturation for (a) CNN (b) CRNN

6. Conclusion and future work

Recognizing dysarthric speech is a challenging problem more than recognizing the normal speech problem. The sounds rendered by those people who have dysarthria are ambiguous and unintelligible due to the lack of coordination of mouth articulators. In this paper, we investigated the performance of two deep neural network models; namely, the convolutional neural network (CNN) and the convolutional recurrent neural network (CRNN), to process the dysarthric speech recognition (DSR) problem. This investigation conducted in the context of a speaker-independent mode, using the samples of

Torgo dataset. The experimental results demonstrated that enhancing the standard RNN with a convolution layer will improve the performance and, in the same time, outperforms the standard CNN, as well. It is clear that CRNN has the potential to improve the DSR performance by attaining an assessment accuracy as 40.6% against 31.4% for CNN. Overall, it is clear both models are suffering due to the high variability in speech intelligibility of DSR. In our future work and from data perspective, we would like to improve accuracy by increasing the number of audio clips per subjects either by looking for other available datasets that has higher number of clips per subject with dysarthria. Another expected work is to record new dataset from different persons having dysarthria support the Arabic language. From the algorithmic perspective, we will improve our algorithm by using a spiking neuron instead of the normal neuron in our algorithms. Based on the spike-timing-dependent plasticity function, the CNN will show better performance in the scope of the spiking neuron properties such as its membrane potential activation function and ability to fire when it reaches a specific threshold.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Author Contributions

Conceptualization, H.A.; Methodology, A.S.; Software and Experiments, H.A.; Validation, H.A. and A.S.; Formal Analysis, H.A. and A.S.; Data Curation, H.A.; Supervision A.S.; Writing—Original Draft Preparation, H.A and A.S.; Review & Editing A.S.

References

[1] P. Enderby, “Disorders of communication”, *Dysarthria. Handbook of Clinical Neurology*, pp. 273-281, 2013.

[2] J. Duffy, “Motor Speech Disorders: Clues to Neurologic Diagnosis”, *Parkinson’s Disease and Movement Disorders*, pp. 35-53, 2000.

[3] F. Rudzicz, G. Hirst, and P. Lieshout, “Vocal Tract Representation in the Recognition of Cerebral Palsied Speech”, *Journal of Speech, Language, and Hearing Research*, pp. 1190-1207, 2012.

[4] M. Kim, J. Yoo, and H. Kim. “Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models”, In: *Proc. of the Annual Conf. of the International Speech Communication Association*. Interspeech, 2013.

[5] N. Joy and S. Umesh, “Improving Acoustic Models in TORGO Dysarthric Speech Database”, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 637-645, 2018.

[6] E. Sanders, M. Ruiter, L. Beijer, and H. Strik, “Automatic Recognition of Dutch Dysarthric Speech: A Pilot Study”, In: *Proc. of The 7th International Conf. on Spoken Language Processing*, Denver, Colorado, USA, pp. 16-20, 2002.

[7] M. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, and J. Carmichael, “A speech-controlled environmental control system for people with severe dysarthria”, *Medical Engineering Physic*s, pp. 586-593, 2007.

[8] S. Caballero Morales and S. Cox, “Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers”, *EURASIP Journal on Advances in Signal Processing*, 2009.

[9] M. Rughani and D. Shivakrishna, “Hybridized Feature Extraction and Acoustic Modelling Approach for Dysarthric Speech Recognition”, *ArXiv*, 2015.

[10] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, “Hmm-Based and Svm-Based Recognition of the Speech of Talkers with Spastic Dysarthria”, In: *Proc. of 2006 IEEE International Conf. on Acoustics Speech and Signal Processing Proceedings*, pp. 1060-1063, 2006.

[11] S. Shahamiri and S. Binti Salim, “A Multi-Views Multi-Learners Approach Towards Dysarthric Speech Recognition Using Multi-Nets Artificial Neural Networks”, *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, pp. 1053 – 1063, 2014.

[12] H. Sharma and M. Hasegawa-Johnson, “Acoustic model adaptation using in-domain background models for dysarthric speech recognition”, *Computer Speech and Language*, pp. 1147-1162, 2013.

[13] M. Alom, T. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. Nasrin and V. Asari, “A State-of-the-Art Survey on Deep Learning Theory and Architectures”, *Electronics*, 2019.

[14] J. Schmidhuber, “Deep Learning in Neural Networks: An Overview”, *Neural Networks*, pp. 85-117, 2015.

[15] B. Sy and D. Horowitz, “A statistical causal model for the assessment of dysarthric speech and the utility of computer-based speech recognition”, *The IEEE Transaction on Biomedical Engineering*, pp. 1282 – 1298, 1993.

[16] G. Jayaram and M. Abdelhamied, "Experiments in dysarthric speech recognition using artificial neural networks", *Journal of Rehabilitation Research and Development*, pp. 162-169, 1995.

[17] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research", In: *Proc. of the Annual Conf. of the International Speech Communication Association, Interspeech*, pp. 1741-1744, 2008.

[18] H. Christensen, P. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech", Interspeech, pp. 1159-1163, 2013.

[19] F. Rudzicz, A. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria", *Language Resources and Evaluation*, pp. 523–541, 2012.

[20] K. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech", In: *Proc. of The IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Prague, Czech Republic, 2011.

[21] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible", *Computer Speech and Language,* pp. 1163-1177, 2013.

[22] X. Menendez-Pidal, J. Polikoff, S. Peters, J. Leonzio and H. Bunnell, "The Nemours database of dysarthric speech", In: *Proc. of Fourth International Conf. on Spoken Language Processing*, pp. 1962–1965, 1996.

[23] M. Mustafa, S. Salim, N. Mohamed, B. Al-Qatab and C. Siong, "Severity-Based Adaptation with Limited Data for ASR to Aid Dysarthric Speakers", *PLOS ONE*, 2014.

[24] J. Kim, N. Kumar, A. Tsiartas, M. Li and S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech", *Computer Speech and Language*, pp. 132-144, 2015.

[25] C. España-Bonet and J. Fonollosa, "Automatic Speech Recognition with Deep Neural Networks for Impaired Speech", In: *Proc. of International Conf. on Advances in Speech and Language Technologies for Iberian Languages,* pp. 97-107, 2016.

[26] K. Kadi, S. Selouani, B. Boudraa, and M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge", *Biocybernetics and Biomedical Engineering,* pp. 233-247, 2016.

[27] F. Rudzicz, A. Frydenlund, S. Robertson, and P. Thaine, "Acoustic-articulatory relationships and inversion in sum-product and deep-belief networks", *Speech Communication,* pp. 61-73, 2016.

[28] J. Ren and M. Liu, "An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks", *(IJACSA) International Journal of Advanced Computer Science and Applications*, pp. 48-52, 2017.

[29] M. Kim, Y. Kim, J. Yoo, J. Wang, and H. Kim, "Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition", *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* pp. 1581-1591, 2017.

[30] J. Mathew, J. Jacob, K. Sajeev, J. Joy, and R. Rajan, R, "Significance of Feature Selection for Acoustic Modeling in Dysarthric Speech Recognition", In: *Proc. of The International Conf. on Wireless Communications, Signal Processing and Networking (WiSPNET).* Chennai, India, 2018.

[31] A. Li, M. Yuan, C. Zheng, and X Li, "Speech enhancement using progressive learning-based convolutional recurrent neural network", *Applied Acoustics*, 2020.

[32] A. Ogawa and T. Hori, "Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks", *Speech Communication*, pp. 70-83, 2017.

[33] A. Sagheer and M. Kotb, "Time series forecasting of petroleum production using deep LSTM recurrent networks", *Neurocomputing*, pp. 203-213, 2019.

[34] L. R and D. Sherly, "Automatic Speech Recognition using different Neural Network Architectures – A Survey", *(IJCSIT) International Journal of Computer Science and Information Technologies,* pp. 2422-2427, 2016.

[35] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", *arXiv*. 2017.