International Journal of Advanced Science and Technology
Vol. 29, No. 7s, (2020), pp. 899-905

Recognition of Different Voice Pathologies with Data Augmentation Strategies

Dr.B.Muni Lavanya[1], Dr.Midde Ranjit Reddy[2], Appawala Jayanthi[3]
[1]*Assistant Professor (Adhoc),* [2]*Associate Professor,* [3]*Assistant Professor*
[1,2,3]*Department of CSE*
[1]*JNTUA College of Engineering, Pulivendula*
[2]*Srinivasa Ramanujan Institute of Technology (Autonomous), Anantapuramu*
[3]*Institute of Aeronautical Engineering, Hyderabad*
[1]*munilavanya45@gmail.com,* [2]*midderanjit@gmail.com*
[3]*jayanthi.a33@gmail.com*

Abstract

In the past few years, there is a lot of improvement in performance of speech analysis methods. Especially, deep learning algorithms plays a crucial role in speech synthesis. Different voices of different genders are going to be classified with the help of deep CNN methodology are mainly focused in this research work. For identifying voice pathologies, different deep learning algorithms are most suitable and for performing this task, there is a lot of requirement of training datasets. But there are limited resources of real-world audio data and here the different audio augmentation methods are to be used for increasing the training dataset. Speech augmentation is mainly used for the training of ANN and it will make the predictions effective. This Strategy is also going to avoid the overfitting and it will improve the model's robustness. Speech classification has shown a improved accuracy when compared to the existing models.

Keywords— *convolutional neural networks, Data augmentation, over fitting, voice disorder.*

1. INTRODUCTION

Speech is the most important and natural method for communicating the human beings. Communications disorders affect the personal ability to communicate which causes problems in speech, language and hearing all or three. The symptoms vary looking on the actual form of communication disorder, but they typically focus on problems in communicating. Voice Therapy and counseling are both playing an effective role especially dealing with various voice pathologies. Languages pathologies, Speech Disorder, Childhood Fluency Disorder, Social Communication Disorder, and Unspecified Communication Disorder are the most important types of disorders. Language disorder is the disorder which is the problem while expressing by themselves and other observed voices, these are not related to hearing issues. Language pathology is also referred as receptive expressive language which is common in young kid. Language disorder is usually noticed in childhood first. People with social communication disorder (SCD) will more precisely perceive people who have significant issues utilizing verbal and nonverbal correspondence for social reason. Social conversations behaviors which include eye contact, facial expressions, and language are influenced by sociocultural and person factors. Signs and symptoms of social communicative disorder consist of deficits in social interaction, social cognition, and pragmatics. Specific behaviors concerned with social communication disorder rely upon the individual's age, his or her predicted level of development and the conversation context. Some examples of behaviors affected by SCD include, using appropriate greetings, converting language and communication fashion based on setting or partner and repairing communication breakdowns. The disorder in which speech sound of the children may have a continual problem while speaking words or sounds. The correct articulation of the phonemes (person sounds) that make up spoken phrases is called speech sound manufacturing. The development of speech sound needed for both the phonological knowledge of speech sounds and the ability to manage the respiratory and vocalized jaw, tongue and lips. Children with speech sound disorder may additionally have issue with the phonological expertise of speech sounds or the

ISSN: 2005-4238 IJAST

capability to coordinate the movements essential for speech. But a few speech errors will be provoked by the injury to brain, Intellectual or progressive disability and difficulties with the hearing or listening to the hear loss, consisting of history of different ear infections and various abnormalities that will affect the speech, which includes cleft palate and cleft lip. Childhood-onset fluency disorders is also known as stuttering is a speech disease that involves continuous and considerable problems with speech glide and ordinary fluency. Stuttering is not unusual amongst young kids. Most kids outgrow this developmental stuttering. Children and adults who stutter may gain from remedies consisting of audio therapy, by the use of different devices to improve speech fluency or cognitive interactive therapy. Unspecified communication disorder can be defined in phrases of severity from slight to profound, and can be first visible in childhood, and can be first visible in childhood, with genetic etiology or acquisition through environmental influences at any point in development. MFCC is a feature extraction method using voice signal. Extraction of Features is the process of making sense of a worth or vector that can be utilized as a thing or a character personality certain thing. MFCC is a broadly utilized strategy in different zones of voice handling field, since it is viewed as significant in speaking to discourse signal. Data augmentation is the process of deforming audio using the available data and used to generate additional training data while using the concept of data augmentation, the semantic meaning of labels does not change.

2. LITERATURE REVIEW

IlyesRebai et al [1] discussed performance can be increased using deep learning based system in speech identification task and various models have been established in the few years ago using various learning methods. Data augmentation techniques are very much useful in increasing the training data which is an effective way for the neural network training process and it will make various predictions. Ensemble based (EM) approaches also have significant consideration in machine learning.

Jan Schluter and Thomas Grill[2] mentioned about audio transformation such as stretching and time shifting to increase training data. Training data can be increased to acquire invariance that are difficult to integrate with model. This methodology is not methodically investigated for audio signals. Pitch shifting is used as an audio augmentation technique. It appears to fill in certain holes in vocal range revealed by our lower training sets.

Justin Salamon & Juan Pablo Bello[3] the proposed a model for environmental sound classification using data augmentation and produces state of the art results. The results are improved based on the mix of a profound, high-limit model and an augmented training set.

VijayadityaPeddinti et al[4] introduced data augmentation technique to avoid from overfitting and improve power of the models. The change in the speed of the sound level is an augmentation strategy and introduced an audio augmentation technique with low usage cost.

BhavikVachhani et al[5] proposed a techniques to classify Dysarthria which alludes to a discourse issue brought about by injury to the cerebrum territories worried about engine parts of discourse offering ascend to exertion full, slow, slurred or prosodic partner irregular discourse. An engineered discourse information is produced and used to order the seriousness of the ailment. The outcome shows that the 3450 sound control expressions are modified into dysarthria articulations utilizing different increase parameters.

Eric Bouteillon [6] presented a semi-supervised technique to create an effective audio tagging as well as a novel facts augmentation method for multi-labels audio tagging named by the author spec Mix. Purpose of these techniques is to use machine learning techniques with less volume of reliable, manually-labeled training data & a larger volume of noisy audio information in a multi-label audio tagging challenge with a huge vocabulary setting.

Stefan kah et al[7-12] offered recognizing feathered creature species in sound chronicles is a troublesome assignment of research. They utilized various convolution neural systems to create highlights removed from obvious portrayals of subject accounts. The feathered creature species which contain 2000 examples containing 37.955 sound records in flying creature CLEF 2018 training dataset. The outcomes show that there is still a significant number opportunity to get better

ISSN: 2005-4238 IJAST

specifically for the soundscape area, which plausible is the most basic genuine worldwide application methods.

3.1 A OVERVIEW OF PROPOSED MODEL

The first step in our proposed approach is to load the audio file. Then perform the feature extraction technique called MFCC from the audio sample. Then, use the train the deep neural networks with the training data and make predictions with the test data. The Deep CNN architecture is used for classifying the outcome into normal or phonological disorders.

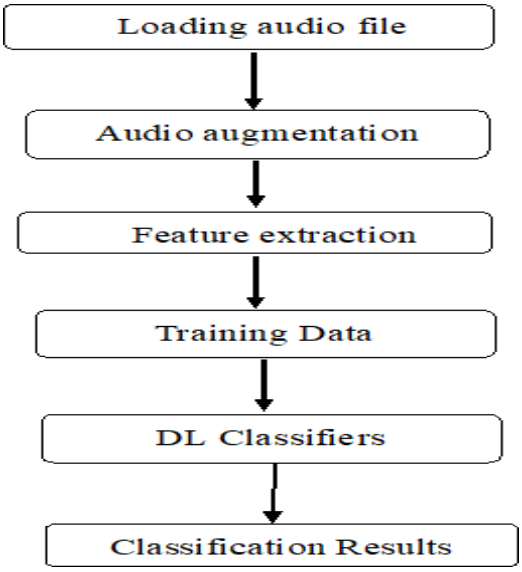

Figure 1: Architecture Diagram

3.2. DATA COLLECTION

Original Training data is extracted from primary school children to recognize various disorders. 20 normal children voices were recorded and voices from 18 children with phonological disorders were also recorded. 33 voices with speech sound disorders were also collected [13].

DISORDERS	Data
Regular Disorders	30 files
Phonological Disorders	22 files
Speech & Language Disorders	42 files

Table 1: Types of Disorders

3.3. METHODS

DATA AUGMENTATION

It is an essential method of state-of-the-art systems for both image and speech recognition to increase the training data and improving the performance of the model [14-18].

3.3.1 TECHNIQUE OF DATA AUGMENTATION

(i) Adding white noise

(ii) Shifting the sound

(iii) Stretching the sound

ISSN: 2005-4238 IJAST

ADDING WHITE NOISE

White noise is a random audio signal having equal amplitude at different frequencies. The samples of a white noise signal can be continuous in time or organize alongside one or greater spatial dimensions. The audible audio is heard by the human ear the range between 20 and 20,000HZ.

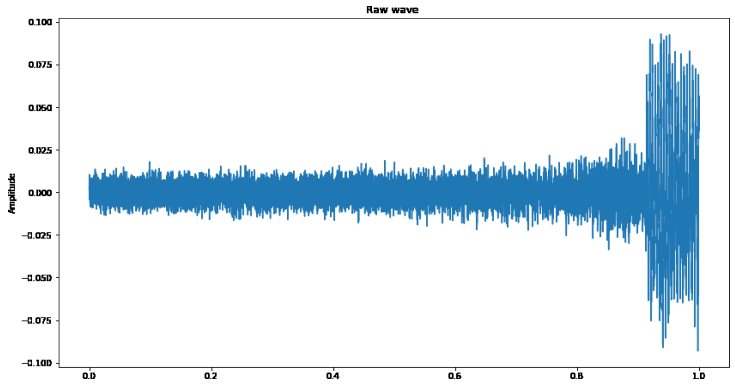

Figure 2: Adding white noise

SHIFTING THE SOUND

Sound shifting is a technique where the original pitch of a sound is raised or decreased through a pre-distinct speech interval. Another way to increase the pitch and minimize time or contrariwise and it was recorded by a repetition of an audio waveform at a different rate.

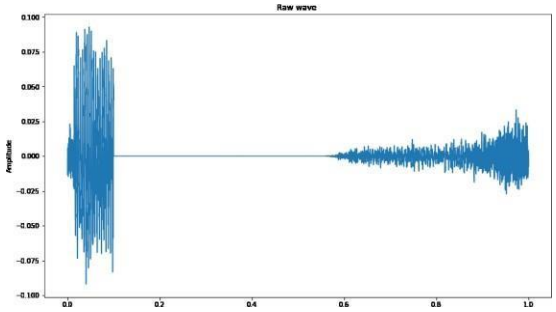

Figure 3: shifting the sound

STRETCHING THE SOUND

The process of changing the speed or duration of an audio signal is called stretching. It is used to conform an allotted time slot such as 30 second commercial or 1-hour broadcast. These methods are often used to suit the pitches and tempos for mixing of two pre-recorded clips cannot be executed or resample.

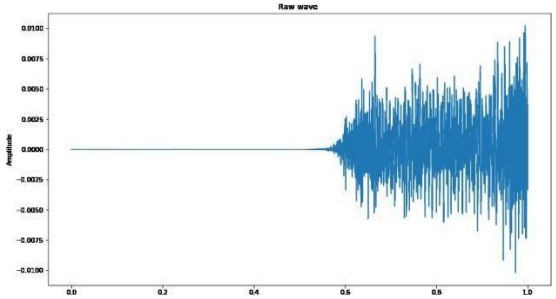

Figure 4: Stretching the sound

ISSN: 2005-4238 IJAST

3.4 . SPEECH FEATURES EXTRACTION

In Feature extraction, the unwanted data has been removed from the speech signals. Mel Frequency Cepstral coefficient (MFCC) is the most widely used feature extraction technique. Prior to MFCC, they are many feature extraction techniques such as LPCS. MFCC technique is used for diverse frequencies that can be received by human ear so it is used to characterize the sound signals as humans. Each audio file is sampled at the rate of 22050 Hz. The frame size is 512 bits. Then 40 MFCC features are extracted for each frame. MFCC summarizes the frequency distribution to analyse both frequency and time characteristics of sounds. The block diagram of an mfcc is shown in fig.

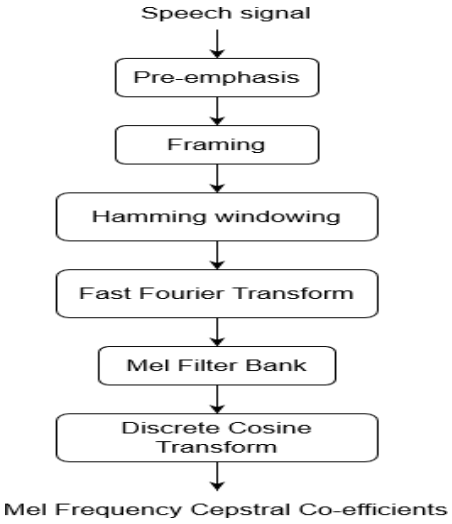

Figure 5: MFCC workflow

3.5 BUILDING THE MODEL

CONVOLUTIONAL NEURAL NETWORK

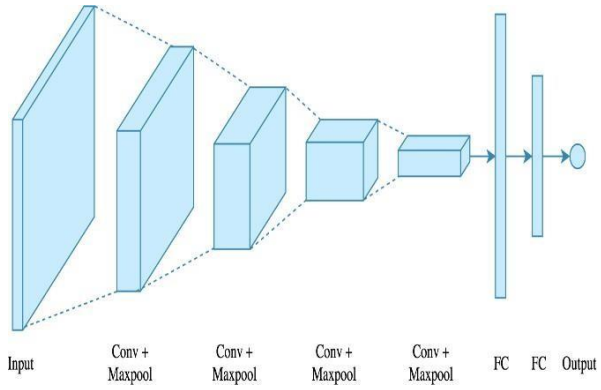

Figure 6: CNN architecture

MFCC features will be added into the Convolutional Neural Network model which is widely used in the deep learning methods. CNN model comprises of two convolution layers, a pooling layer, and different completely associated layers. Right now, convolution layers and pooling layer are intended to extricate the highlights, and the completely associated layers are treated as a classifier. CNN was trained over 100 epoch using binary cross entropy loss function and Adam optimizer is used with a learning rate of 0.001. Activation function is used in the layers except in output layer. In the output layer, softmax function is used. Input will pass through multiple convolution layers with increasing the filter size of 32, 64, 128, 512 and kernel size of 2*2 in all layers. A softmax function in the yield layer of CNN is utilized to n foresee the probabilities of each class for a sound, and the sound will be alloted to the class with the most elevated likelihood.

ISSN: 2005-4238 IJAST

3.6 EXPERIMENTAL RESULTS

Classification of voice disorder is performed using the CNN model with various augmentation techniques. For training the CNN model Adam optimizer is used with different batch sizes and epochs to get better accuracy. CNN is performed on the collected voice data to learn different voice disorders and is tested against the test data. The results show that 87% of children are classified correctly. Again, CNN is applied on augmented audio file and the performance is evaluated. It shows 11% improvement over the original data.

DATASET & MODEL	ACCURACY
Original dataset &CNN	87%
Augmentation using original dataset & Deep CNN	98%

Table 2:- Evaluation of results using Deep CNN model

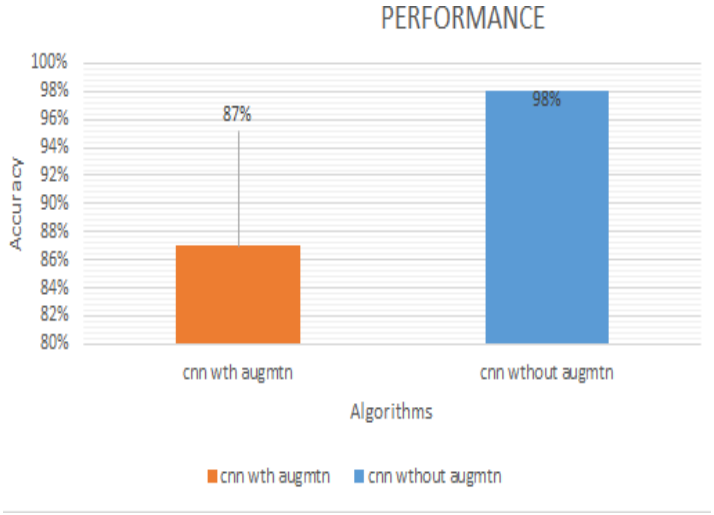

Figure 7: Accuracy comparison of CNN model with and without augmentation

4. CONCLUSION

In this research work, the deep convolutional neural network architecture combined with audio augmentation technique provides best results in classifying the normal and pathological voices. To overcome the difficulty of collecting the pathological voices, the audio augmentation techniques are implemented with low cost. The CNN model with augmentation techniques outperforms the CNN model with original data. In future, the work is improved further with another hyper parameter. Also, other feature extraction techniques can be used for further improvement.

REFERENCES

[1] Rebai, I., BenAyed, Y., Mahdi, W., &Lorré, J. P. (2017). Improving speech recognition using data augmentation and acoustic model fusion. *Procedia computer science*, *112*, 316-322.

[2] Schlüter, J., & Grill, T. (2015, October). Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks. In *ISMIR* (pp. 121-126).

[3] Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, *24*(3), 279-283.

[4] Ko, T., Peddinti, V., Povey, D., &Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[5] Vachhani, B., Bhat, C., &Kopparapu, S. K. (2018, September). Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In *Interspeech* (pp. 471-475).

[6] Bouteillon, E. SPECMIX: A Simple Data Augmentation And Warm-Up Pipeline To Leverage Clean And Noisy Set For Efficient Audio Tagging.

[7] Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., &Eibl, M. (2017, September). Large-Scale Bird Sound Classification using Convolutional Neural Networks. In *CLEF (Working Notes)*.

[8] Huang, C. L., & Hori, C. (2013, October). Classification of children with voice impairments using deep neural networks. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (pp. 1-5). IEEE.

[9] Teixeira, J. P., Fernandes, P. O., &Alves, N. (2017). Vocal Acoustic Analysis–Classification of Dysphonic Voices with Artificial Neural Networks. *Procedia computer science*, *121*, 19-26.

[10] Verde, L., De Pietro, G., &Sannino, G. (2018). Voice disorder identification by using machine learning techniques. *IEEE Access*, *6*, 16246-16255.

[11] Ummadi Janardhan Reddy, Pandluri Dhanalakshmi, Pallela Dileep Kumar Reddy Image Segmentation Technique Using SVM Classifier for Detection of Medical Disorders Ingénierie des Systèmes d'Information, Vol. 24, No. 2, pp. 173-176, April 2019

[12] Guan, H., &Lerch, A. (2019, January). Learning Strategies for Voice Disorder Detection. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)* (pp. 295-301). IEEE.

[13] Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., &Eibl, M. (2017, September). Large-Scale Bird Sound Classification using Convolutional Neural Networks. In *CLEF (Working Notes)*.

[14] G. Ramu, P. Dileep Kumar Reddy, Appawala Jayanthi "A Survey of Precision Medicine Strategy Using Cognitive Computing" International Journal of Machine Learning and Computing, Vol. 8, No. 6, December 2018 DOI: 10.18178/IJMLC2018.8.6.741 (Scopus) (UGC Approved) Journal No: 48748, pp 530 to 535

[15] Jaya Prakash, R., & Devi, T. (2019). Resolving presentation attack using CNN (convolutional neural network). Test Engineering and Management, 81(11-12), 5454-5458. Retrieved from www.scopus.com

[16] Ramu, G. A secure cloud framework to share EHRs using modified CP-ABE and the attribute bloom filter. Educ Inf Technol 23, 2213–2233 (2018). DOI https://doi.org/10.1007/s10639-018-9713-7

[17] Ramu, G., Reddy, B.E., Jayanthi, A. et al. Fine-grained access control of EHRs in cloud using CP-ABE with user revocation. Health Technol. 9, 487–496 (2019). http://dx.doi.org/10.1007/s12553-019-00304-9

[18] P. Dileep Kumar Reddy, R. Praveen Sam, C. Shoba Bindu "Optimal Blowfish Algorithm based Technique for Data Security in Cloud"Int. J. Business Intelligence and Data Mining, ISSN online 1743-8195, ISSN print 1743-8187, Vol. 11, No. 2, 2016.Pp.171–189.DOI: 10.1504/IJBIDM.2016.10001484. (Inder Science)(UGC Approved). Journal No: 16481

[19] Ahn, H., Lee, J. -., & Cho, H. -. (2019). Implementation of multi-object recognition algorithm using enhanced R-CNN. Test Engineering and Management, 81(11-12), 1-8. Retrieved from www.scopus.com

[20] J. Somasekar a, G. Ramesh , Gandikota Ramu, P. Dileep Kumar Reddy, B. Eswara Reddy e, Ching-Hao Lai, "A dataset for automatic contrast enhancement of microscopic malaria infected blood RGB images", Data in brief, Elsevier, https://doi.org/10.1016/j.dib.2019.104643,2352-3409/2019.