iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6395

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Speech Emotion Recognition using Deep Neural Networks

Balaji Dharamsoth[1], Zia Uddin Mohammed[2], Dr. Suresh Pabboju[3]
[1, 2, 3]*Department of Information Technology, Chaitanya Bharathi Institute of Technology*

Abstract: ***The aim of this project work is to propose a speech emotion recognition method based on speech features and speech transcriptions (text). Modelling emotional behaviours is a challenging task due to the variability in perceiving and describing emotions. We try to perform emotion analysis on the speech by collecting speech and textual features and applying a deep neural network model which can classify the sentiments of the speech. Ideally, we would like to experiment with several deep neural network models which take in different combinations of speech features and text as inputs. Speech features such as Mel-Frequency Cepstral Coefficients (MFCC) help retain emotion related low-level characteristics in speech whereas text helps capture the semantic meaning, both of which help in different aspects of emotion detection.***
Keywords: mfcc features, confusion matrix, emotional speech recognition.

I. INTRODUCTION

Speech Emotion recognition is the interpretation of human voice by a computational device. This means that a computers' computational software containing mathematical algorithms can calculate our speech emotion through a series of inputs. Research of speech recognition allows machine to understand the sentiments of the speaker and use that information during human-machine interaction. A speech carries information like pace, pitch of the voice, content etc. Emotion plays an important role in interpersonal human interaction. Human-machine interfaces will benefit from incorporating emotional capabilities to recognize the affective States of the users. Studying and understanding the emotional modulation conveyed on expressive speech is an important step toward designing robust machine learning frameworks that exploit the underlying production of emotional speech. The aim of this project is to experiment several deep learning algorithms like RNN, Bi-RNN on speech features along with its textual features in order to classify the emotions of the speech. We will compare the performance of these models over different combinations of speech and text features in an attempt to achieve accuracies greater than existing state-of-art methods.

II. PROPOSED SYSTEM

We consider speech transcriptions along with its corresponding speech features like MFCC, which together provide semantic relationships and the necessary features required to distinguish among different emotions accurately. Experiments have been performed on audio features, audio-visual features, audio-transcript features independently as well as together to achieve accuracies greater than existing state-of-the-art methods. Different combinations of inputs have been used in different DNN architectures.

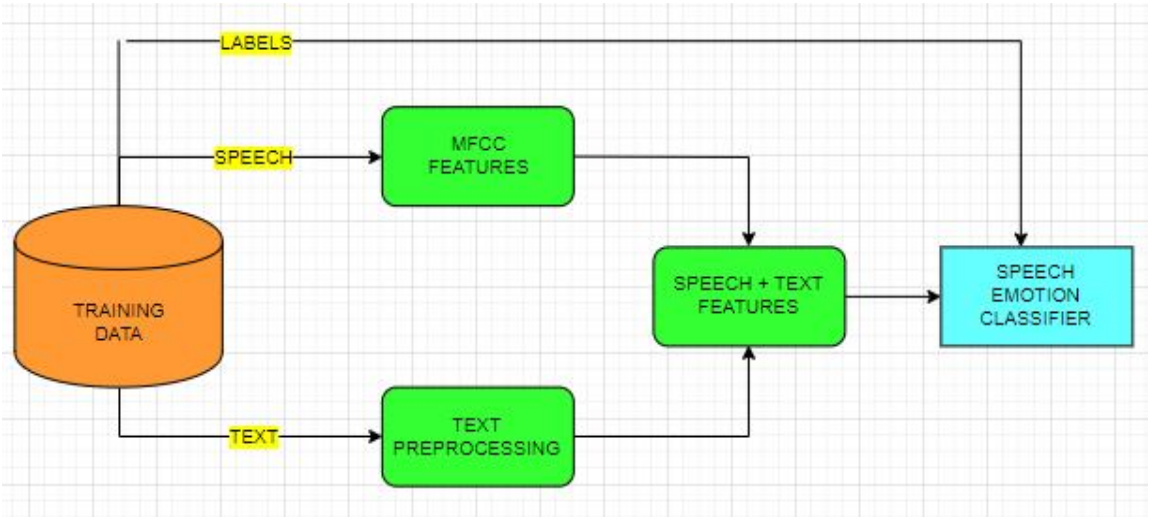


Fig: 1. Training Phase.

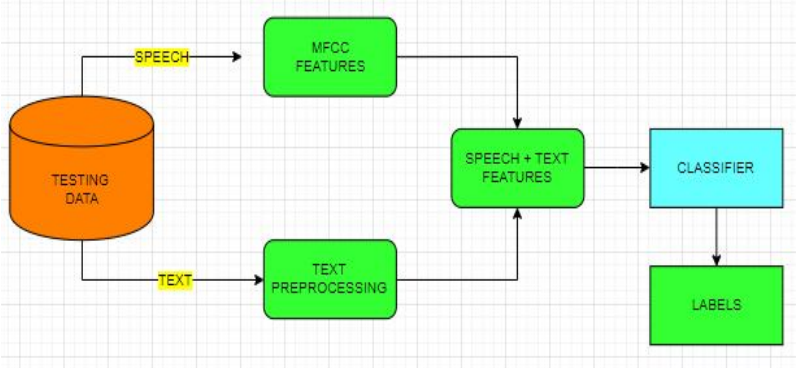


Fig: 2. Prediction Phase.

III. EXPERIMENTAL SETUP

A. IEMOCAP Data Set

Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. An acted, multimodal and multi speaker database. Collected at Signal Analysis and Interpretation Laboratory (SAIL) lab at University of Southern California (USC). We used this database in this work. The IEMOCAP corpus comprises of five sessions where each session includes the conversation between two people, in both scripted and improvised topics and their corresponding labelled speech text (both phoneme and word level). Each session is acted upon and voiced by both male and female voices to remove any gender bias. The data thus collected is then divided into small utterances of length varying between 3-15 seconds, which are then labelled by evaluators. Each utterance is evaluated by 3-4 assessors. The assessors had the option of labelling every utterance among 10 different emotion classes (neutral, happiness, sadness, anger, surprise, fear, disgust frustration, excited, other). In our experiments, we have considered only 6 of them (anger, excitement (happiness), neutral, frustration, disgust and sadness) to remain consistent with earlier research. We chose utterances where at least 2 experts were in with their decision and only used improvised data, again being consistent with prior research, as the scripted text shows a strong correlation with labelled emotions and lead to lingual content learning, which can be an undesired side effect.

B. Deep Neural Networks

We are building different models based on RNN and Bi-RNN with LSTM's by training them on different combinations of audio and audio-transcript features. We are using sequential models as it helps us remember the context and to make the decision based on the experience. We are also using K-Fold cross validation sampling technique and training our model on various splits in order to achieve better accuracy and to overcome over-fitting of the model.

C. RNN Architecture

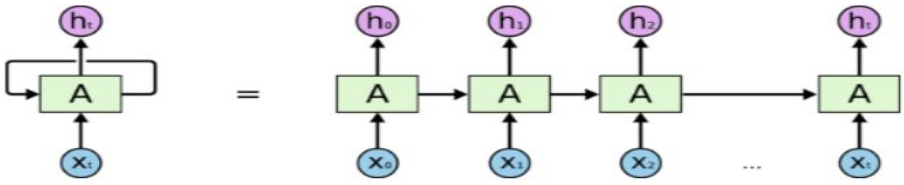


Fig:3. RNN Architecture.

D. BI-RNN Architecture

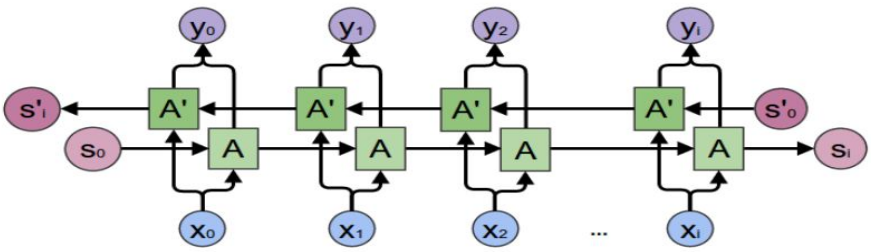


Fig:4. Bi-Directional RNN Architecture.

IV. MODELS AND RESULTS

Here we will discuss the different models which we have generated on speech and textual features and we will record their performance and compare them in order to get the best model out of them which can give us state-of-the-performance.

A. *Speech based Models*

1) Our first model is an RNN model with 1 hidden layer with 300 hidden neurons. The input layer consists of 39 neurons since we are considering 39 audio features and the output layer contains 6 neurons as we are classifying only six emotions with Adam as optimization algorithm and categorical cross entropy loss function.

2) Our second model is an RNN model with 2 hidden layers consisting of 300 and 200 hidden neurons respectively with the input layer containing 39 features and the output layer containing 6 neurons for classifying 6 emotions with Adam as optimization algorithm and categorical cross entropy loss function.

3) Our third model is also an RNN model with 3 hidden layers. Each layer consists of 300, 200 and 100 neurons respectively. The input features are audio features and output layer have 6 neurons with Adam as optimization algorithm and categorical cross entropy loss function.

Table 1. Performance comparison of RNN with Speech Features

Hidden Layers	Neurons (layer1, layer2, …)	(Loss, Accuracy)
1	300	(4.1894, 0.4498)
2	300, 200	(5.5632, 0.4529)
3	300, 200, 100	(4.8221, 0.4701)

4) Our fourth model is based on Bi-RNN model with 1 hidden layer containing 256 hidden neurons in each forward and backward network collectively making a layer of 512 hidden neurons and a dropout layer with 0.5 dropout rate. Input has 39 speech features and the output has 6 neurons with Adam as optimization algorithm and categorical cross entropy loss function.

5) Our fifth model is based on Bi-RNN model with 1 hidden layer containing 512 hidden neurons in each forward and backward network collectively making a layer of 1024 hidden neurons and a dropout layer with 0.5 dropout rate. Input has 39 speech features and the output has 6 neurons with Adam as optimization algorithm and categorical cross entropy loss function.

6) Our sixth model is also a Bi-RNN model with 2 hidden layers containing 256 and 128 hidden neurons a dropout layer with 0.5 dropout rate. Input has 39 speech features and the output has 6 neurons with Adam as optimization algorithm and categorical cross entropy loss function.

Table 2. Performance comparison of Bi-RNN with Speech Features

Hidden Layers	Neurons (layer1, layer2, …)	(Loss, Accuracy)
1	256	(0.6440, 0.4937)
1	512	(0.6211, 0.5173)
2	256, 128	(0.6134, 0.5027)

B. *Speech with Text based Models*

1) Our first model is an RNN model with 1 hidden layer with 500 hidden neurons. The input layer consists of 139 neurons since we are considering 39 audio and 100 audio-transcript features, and the output layer contains 6 neurons as we are classifying only six emotions with Adam as optimization algorithm and categorical cross entropy loss function.

2) Our second model is an RNN model with 2 hidden layers consisting of 500 and 300 hidden neurons respectively with the input layer containing 139 features and the output layer containing 6 neurons for classifying 6 emotions with Adam as optimization algorithm and categorical cross entropy loss function.

3) Our third model is also an RNN model with 3 hidden layers. Each layer consists of 500, 300 and 100 neurons respectively. The input features are audio and text features and output layer have 6 neurons with Adam as optimization algorithm and categorical cross entropy loss function.

Table 3. Performance comparison of RNN with Speech and Text Features

Hidden Layers	Neurons (layer1, layer2, …)	(Loss, Accuracy)
1	500	(5.1050, 0.4720)
2	500, 300	(5.1914, 0.4824)
3	500, 300, 100	(6.0309, 0.4843)

4) Our fourth model is based on Bi-RNN model with 1 hidden layer containing 256 hidden neurons in each forward and backward network collectively making a layer of 512 hidden neurons and a dropout layer with 0.5 dropout rate. Input has 39 speech features and 100 textual features, and the output has 6 neurons with Adam as optimization algorithm and categorical cross entropy loss function.

5) Our fifth model is based on Bi-RNN model with 1 hidden layer containing 512 hidden neurons in each forward and backward network collectively making a layer of 1024 hidden neurons and a dropout layer with 0.5 dropout rate. Input has 39 speech features and 100 textual features, and the output has 6 neurons with Adam as optimization algorithm and categorical cross entropy loss function.

6) Our sixth model is also a Bi-RNN model with 2 hidden layers containing 256 and 128 hidden neurons a dropout layer with 0.5 dropout rate. Input has 39 speech features and 100 textual features, and the output has 6 neurons with Adam as optimization algorithm and categorical cross entropy loss function.

Table 4. Performance comparison of Bi-RNN with Speech and Text Features

Hidden Layers	Neurons (layer1, layer2, …)	(Loss, Accuracy)
1	256	(0.6841, 0.4837)
1	512	(0.7012, 0.4545)
2	256, 128	(0.6684, 0.4729)

C. Speech with Text based Model with K-Fold CV

Here we have generated a model using LSTM based Bi-RNN with 1 hidden layer containing 256 neurons and a dropout layer with a dropout rate of 0.5. The model is trained on speech and textual features using K-Fold cross validation sampling technique and tested it on different 'k' values and recorded the performance in the below table.

Table 5. Performance comparison of Bi-RNN with K-Fold CV

No. of Splits (k)	Neurons (layer1, layer2, …)	(Loss, Accuracy)
3	256	(0.6644, 0.8072)
5	256	(0.3791, 0.8383)
7	256	(1.0856, 0.7247)

D. Interpretation

The following points can be interpreted based on the above recorded results,

1) RNN's with Speech features (Table 1), Speech & Text features (Table 3) performed poorly. On increasing hidden layers, the model does show a very slight improvement, but it doesn't justify the trade-off between complexity and performance.

2) Bi-RNN's with Speech features (Table 2) performed similar to the RNN's and when trained on Speech & Text features (Table 4) the model shows a minute improvement in the accuracy with no decrease loss and was overfitting.

3) Bi-RNN's with K-fold Cross Validation (Table 5) has outperformed other models and has shown significant improvement in the accuracy with a very less validation loss.

4) Bi-RNN using Speech & Text features with K-Fold CV where k=5 has shown the state-of-the-art performance (Accuracy: 0.8383, Loss: 0.3791).

E. Output - Confusion Matrix (State-of-the-Art Performance)

Fig:6. Confusion Matrix for the Bi-RNN Model on Speech & Text features using K-Fold Cross Validation with K=5.

V. CONCLUSIONS AND FUTURE SCOPE

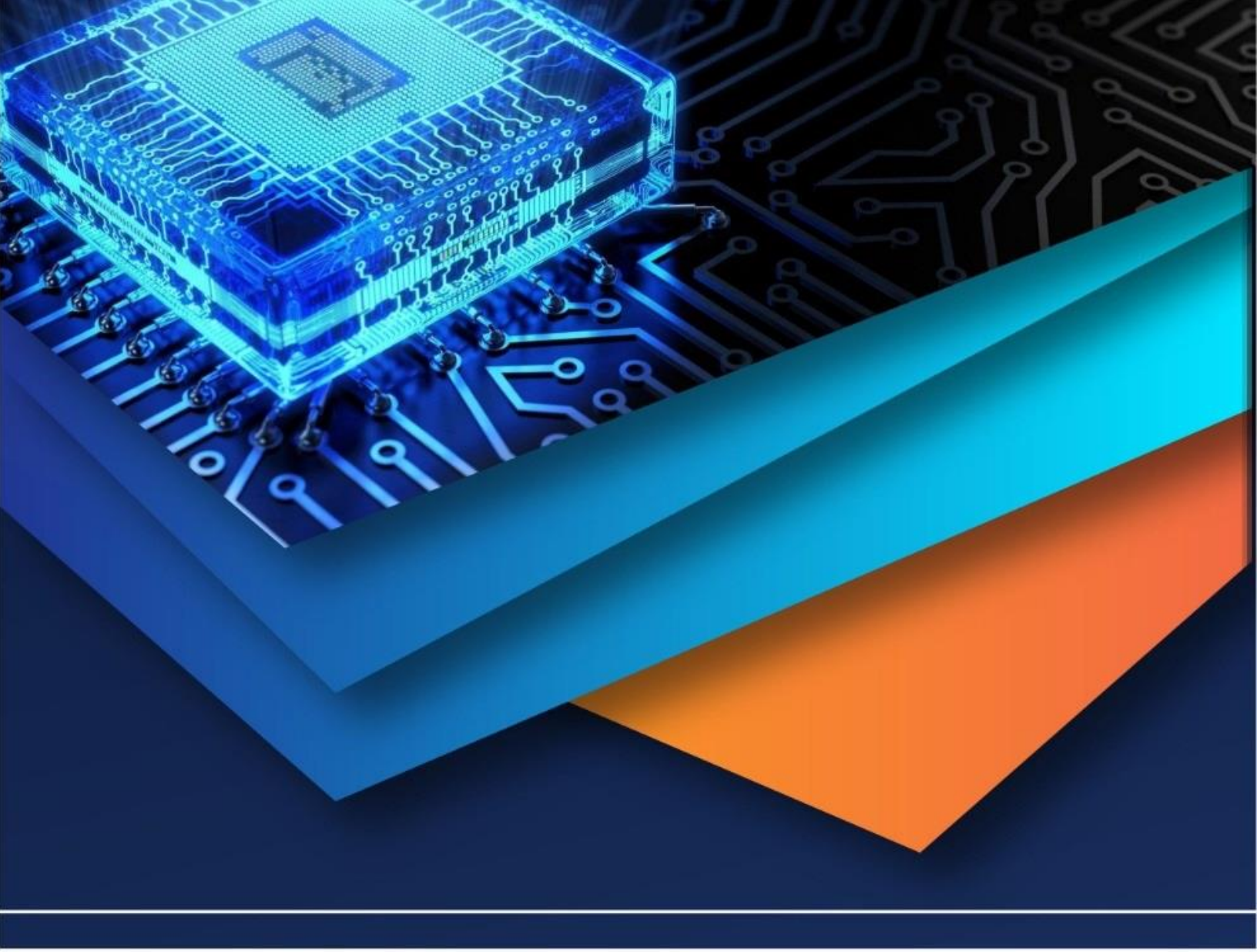
In this project, we have proposed multiple sequential based architectures such as RNN and Bi-RNN to work with speech features and transcriptions. Bi-RNN model trained on speech features provides better accuracy than its RNN based counterpart, which further improves when combined with text. The current Spectrogram-MFCC based model results in an overall emotion detection accuracy of 78.4%, an almost 7% improvement to the existing state-of-the-art methods. Better results are observed when speech features are used along with speech transcriptions. The combined MFCC-Text model gives an overall accuracy of 83.8% an almost 13% improvement over current benchmarks respectively. The proposed models can be used for emotion-related applications such as conversational chatbots, social robots, etc. where identifying emotion and sentiment hidden in speech may play a role in the better conversation. We will try to do more experiments on other public benchmark databases to analyse our work. The other direction of research is to use this architecture to deal with multimodal features for emotion recognition in real time applications. We will continue to further study speech emotion recognitions based on other deep neural nets with the aim to study how to improve the recognition rate of speech emotion recognition.

VI. ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to Dr. Suresh Pabboju Professor, Dept. of IT, our project guide for his valuable guidance and constant support, along with his capable instructions and persistent encouragement. We are grateful to our Head of the Department, Dr. Suresh pabboju, for his steady support and provision of every resource required for the completion of this project. We would like to take this opportunity to thank our principal, Dr. P. Ravinder Reddy, as well as the management of the institute, for having designed an excellent learning atmosphere.

REFERENCES

[1] Reza Lotfian, (Student Member, IEEE), and Carlos Busso, (Senior Member, IEEE), "Lexical Dependent Emotion Detection using Synthetic Speech Reference", date of current version March 1, 2019, Digital Object Identifier 10.1109/ACCESS.2019.2898353.
[2] Giovanni Dimauro, Vincenzo Di Nicola, Vitoantonio Bevilacqua, Danilo, and Francesco Girardi, "Assessment of Speech Intelligibility in Parkinson's disease using a Speech-To-Text System", date of current version November 7, 2017, Digital Object Identifier 10.1109/ACCESS.2017.2762475.
[3] Reza Lotfian (Student Member, IEEE), and Carlos Busso (Senior Member, IEEE), "Effective Spectral and Excitation Modelling Techniques for LSTM-RNN-Based Speech Synthesis Systems", date of current version March 1, 2019, Digital Object Identifier 10.1109/ACCESS.2019.2898353.
[4] Emilio Granell and Carlos-D. Mart´ınez-Hinarejos, "Multimodal Crowd sourcing For Transcribing Handwritten Documents", IEEE/ACM Transactions on Audio, Speech, And Language Processing, Vol. 25, No. 2, February 2017.
[5] Yuki Saito , Shinnosuke Takamichi , Member, IEEE, and Hiroshi Saruwatari , Member, IEEE, "Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks", IEEE/ACM Transactions on Audio, Speech, And Language Processing, Vol. 26, No. 1, January 2018.
[6] Geoffrey S. Meltzner, James T. Heaton, Yunbin Deng, Gianluca De Luca, Serge H. Roy, and Joshua C. Kline, "Silent Speech Recognition as an Alternative Communication Device for Persons with Laryngectomy", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 12, December 2017.
[7] S. Chandrakala and Natarajan Rajeswari, "Representation Learning Based Speech Assistive System for Persons with Dysarthria", IEEE Transactions on Neural Systems And Rehabilitation Engineering, Vol. 25, No. 9, September 2017.
[8] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2013, pp. 7962–7966.
[9] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2014, pp. 3829–3833.
[10] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2015, pp. 4455–4459.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY