International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

A DNN-BASED REAL-TIME SPEECH-TO-TEXT CONVERSION

Sai Siva Ram Pasam[1], Sainadh Varma Reddicherla[2], Venkata Abhiram Nelluri[3] , Neeraj Pindi[4]

Department of ECE at R.V.R. & J.C. College Of Engineering

Abstract - The instant voice or recorded speech is converted to text format in real-time concurrent speech-to-text systems.As there are multiple ways of doing this operation, we challenge ourselves to verify the WER of the system. We created a real-time speech-to-text conversion system using the DNN techniques in the GOOGLE COLLABORATORY platform which is trained with the LJ Speech databases that contain the transcription of every speech to verify the word error rate. DNN has been proven to be the better technique for predicting the system's desired output. This speech recognition model, which we refer to as DeepSpeech 2, was created by combining 2D CNN, RNN, and CTC Loss.Due to this, the model is getting a lower Word Error Rate and loss.

Keywords: CNN, RNN, Speech Recognition, WER, CTC Loss.

1. INTRODUCTION

Natural language processing has historically been a promising subject for research. Natural Language Processing has a large number of uses. Speech recognition is one of natural language processing's most important applications. We have traditionally valued speech as the most crucial component of daily communication. We use a particular language to communicate our views. By using speech recognition, computers can comprehend our language (natural language). Voice recognition, often known as word-by-word recognition, is the process of taking speech characteristics and categorizing those using already recorded datasets.

A word needs to be sent to more advanced software for syntactic and semantic analysis in order to be recognized. This pattern-matching method evaluates audio signals by framing them into phonetics (the number of words, phrases, and sentences). To complete this activity, one must first record a voice sample and then convert it to

.wav format. When a word is recognized, parameters based on the spectrum are acquired.

Speech intensification is the process of improving a speech sample's overall comprehensibility or quality. Speech improvement involves de-reverberating and separating the unconstrained signals in addition to reducing noise in a speech sample. It is desirable to enhance the speech since when speech is processed through any of the instruments in the slab it gets impacted by the noise (background noise or otherwise) and the individuality of the voice varies with time which influences the complete recognition process. Finding devices that actually work in a variety of real-world settings has thus become a very difficult problem for scientists. Nonetheless, this criterion is crucial in defending the algorithm's effectiveness in terms of quality and understandability. Taking all these points into consideration, 2D CNN and RNN have been chosen to develop the model.

CTC (Connectionist Temporal Classification) loss is a widely used loss function in automatic speech recognition (ASR) systems, particularly in those using recurrent neural networks (RNNs). CTC loss was introduced in a paper by Alex Graves et al. in 2006, and it has since become a popular method for training ASR models. CTC loss has several advantages for ASR systems. It allows the model to output label sequences of variable length, which is important for transcribing speech signals that have different lengths. This is achieved by introducing a blank label and allowing the model to output the blank label multiple times in a row. It also simplifies the training process by removing the need for an alignment step between the input speech signal and the ground truth label sequence. Additionally, CTC loss can be used with any RNN-based ASR model, including LSTM and GRU networks.

Overall, CTC loss has proven to be an effective and flexible loss function for training ASR models, and it has become a standard component of many state-of-the-art speech recognition systems.

The remaining portions of the paper are summarized below. Related work and studies are presented in Section 2. The proposed methodology is described in Section 3. In Section 4, we describe the Experimental view. The results are presented in Section 5. Finally, the paper is concluded in Section 6.

2. RELATED WORK

[1] Takao Suzuki, Yasuo Shoji, IEEE, Digital Communications Laboratories, Oki Electric Industry Co. Ltd., Japan, pp. 1515-1519, 1989: From the viewpoint of speech communication services for the asynchronous transfer mode (ATM) network and in order to introduce the necessary conditions for speech processing over an ATM network, the authors have developed a novel speech-processing scheme applied at the end of the ATM network. For this speech processing, speech signals are processed basically by two techniques: silence deletion for speech compression and low-bit coding for 32-kb/s adaptive differential PCM (ADPCM). In order to reduce speech quality degradation caused by lost ATM cells in network congestion conditions, the authors propose a cell-reconstruction algorithm using waveform substitution for ADPCM-coded speech based on the pitch estimation method. In addition, to maintain good speech quality, some new algorithms for speech processing are introduced. It was confirmed through subjective evaluation tests that the proposed speech-processing scheme for the ATM network could provide good speech quality up to a cell loss rate of about 3%. Two kinds of custom LSIs for implementing these speech-processing algorithms are described.
Summary: Studied a new speech processing scheme for ATM switching systems.

[2] J. S. Lim, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-26, no. 5, pp.471 - 472 1978: An intelligibility test was performed to evaluate a correlation subtraction method for the enhancement of degraded speech due to additive white noise. Results indicate that such a scheme does not significantly increase speech intelligibility at the S/N ratios where the intelligibility scores of unprocessed speech range between 20 and 70 percent.
Summary: Studied the evaluation of a correlated subtraction method for enhancing speech degraded by additive white noise.

[3] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Chen, Z. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In International conference on machine learning (pp. 173-182). PMLR.
Summary: This is the original paper that introduced the DeepSpeech2 model. The paper describes the architecture of the model, which is based on a stack of convolutional and recurrent neural networks, and uses Connectionist Temporal Classification (CTC) loss to transcribe speech to text. The model was trained on large datasets of English and Mandarin speech, and achieved state-of-the-art performance on several benchmark datasets.

[4] Hannun, A. Y., Maas, A. L., & Ng, A. Y. (2014). Deep speech: scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.
Summary: This paper describes an earlier version of the DeepSpeech model, which was the precursor to DeepSpeech2. The model is based on a deep neural network architecture that uses time-delay neural networks to capture temporal dependencies in speech signals. The paper includes experiments on the Wall Street Journal dataset and the CHiME-2 dataset, which show that the model can achieve competitive performance with traditional speech recognition systems.

[5] Kim, T., & Kim, K. (2017). Convolutional neural networks for speech recognition: An overview. In 2017 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 19-22). IEEE.
Summary: This paper provides an overview of convolutional neural networks (CNNs) and their use in speech recognition. CNNs have been used to extract features from speech signals and have been shown to be effective in recognizing phonemes and words. The paper also discusses some of the challenges in using CNNs for speech recognition.

[6] Li, X., Li, Y., Zhang, Y., Li, Z., Li, J., Li, Y., ... & Yan, S. (2020). ESPnet-ST: All-in-One Speech Translation Toolkit. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7241-7252).
Summary: This paper presents the ESPnet-ST toolkit, which is an open-source toolkit for building end-to-end speech translation systems. The toolkit is based on deep neural networks and includes a range of pre-trained models and tools for data preparation, training, and evaluation.

3. METHODOLOGY

The methodology for speech-to-text using Deep Neural Networks (DNNs) in this paper typically involves the following steps:

1. Data collection: The first step in any speech-to-text project is to collect a large dataset of audio recordings and their corresponding transcriptions. The audio recordings should be diverse in terms of speaker gender, accents, background noise, and speech styles to ensure that the model is trained on a variety of speech patterns.

2. Data preprocessing: Before training the model, the audio recordings need to be preprocessed. This includes

converting the audio to a spectrogram or other frequency-domain representation, segmenting the audio into shorter frames, and using various signal processing methods to improve the audio quality.

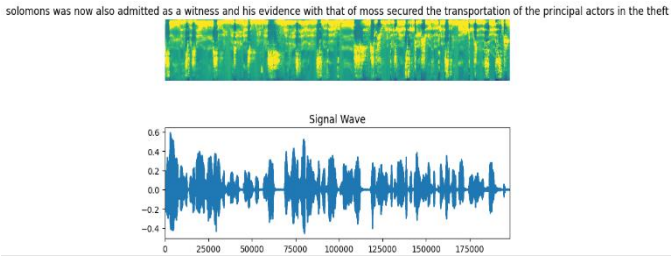

Fig. 1.Audio converted to Spectrogram.

4. Model training: Once the data has been preprocessed, the DNN model can be trained using supervised learning. The model takes in the audio features and predicts the corresponding text transcription. In order to reduce the discrepancy between the predicted and target transcriptions in the training data, the model's parameters are adjusted during the training process.

5. Model validation: The model is validated after training on a different set of data not used during training. The model's accuracy is evaluated by comparing its predicted transcriptions to the target transcriptions for the validation data.

6. Deployment: Once the model has been trained and validated, it can be deployed to recognize speech in real-time. The speech input is processed in the same way as during training, and the model outputs the corresponding text transcription.

4. EXPERIMENTAL VIEW

A. Process Flow

Understanding the process flow of a speech-to-text conversion system is essential for designing, developing, and deploying an accurate and efficient ASR system. By breaking down the process into distinct stages, it becomes easier to identify potential bottlenecks, fine-tune the system, and optimize its performance.

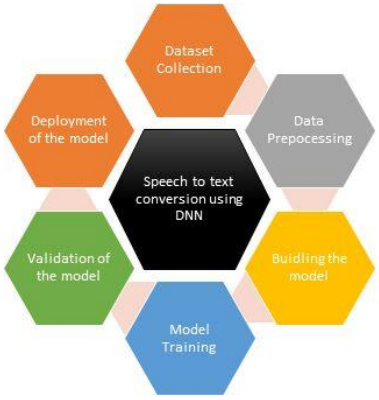

Fig.2.Flow chart of the proposed model

B. Dataset – The LJ Speech dataset

A public domain speech dataset with 13,100 brief audio clips of the speakers reading passages from seven non-fiction books makes up this collection. For every clip, a transcription is offered. There is roughly 24 hours' worth of clips, ranging in length from 1 to 10 seconds. Transcripts.csv contains the necessary metadata. One record appears on each line in this file, which is separated by the pipe character (0x7c). The fields are:

1. **ID**: this is the name of the corresponding .wav file
2. **Transcription**: words spoken by the reader (UTF-8)
3. **Normalized Transcription**: transcription with numbers, ordinals, and monetary units expanded into full words (UTF-8).

Each audio file has a sample rate of 22050 Hz and is a single-channel 16-bit PCM WAV. This dataset could be used on a variety of applications. Some of them are:

Voice-controlled applications: Synthetic speech command datasets can be used to train models that can recognize spoken commands and control applications like music players, navigation systems, and home automation systems.

Accessibility: Speech recognition technology can be used to make applications more accessible for people with disabilities. Synthetic speech command datasets can be used to train models that can recognize spoken commands and enable people with disabilities to use applications more easily.

Automotive: Speech recognition technology can be used in the automotive industry to enable drivers to control in-car entertainment, navigation systems, and other features using voice commands. Synthetic speech command datasets can be used to train models that can recognize these commands and improve the user experience.

After preprocessing the dataset, it is split into two sets, of which the validation data have 10%, and the training set the remaining 90%.

C. Model Architecture

Convolutional Neural Networks (CNNs) are frequently used for image recognition tasks but can also be modified for speech recognition tasks. In speech recognition, recurrent neural networks (RNNs) are a common type of neural network. The next likely scenario can be predicted using patterns and RNNs, which are built to identify sequential characteristics in data. Here is a general summary of the architecture of the model proposed for speech-to-text conversion:

Input: The input to the CNN is a sequence of audio samples, represented as a spectrogram as shown in Fig.1.

Convolutional layers: Convolutional layers make up the first few layers of the CNN and are trained to extract low-level features from the audio input. Each convolutional layer creates a set of feature maps by applying a set of teachable filters to the input.

Recurrent layers: The RNN layer is used to process these features over time and generate an output sequence of phonemes or words. The RNN layer has the ability to maintain a memory of previous inputs, allowing it to capture temporal dependencies in the speech signal.

Dense layer: Dense layer is commonly used in the output layer of speech-to-text conversion systems, where they are used to map the intermediate features learned by the RNN layers to the final output labels, such as phonemes or words. A vector of probabilities, in which each component represents the probability of the corresponding output label, is the result of the dense layer. The final output is determined by which label has the highest probability.

Classification layer: In speech-to-text conversion, the classification layer is the final layer of the neural network that produces the output transcription of the spoken input. The classification layer is typically a fully connected layer a softmax activation function. The classification layer plays a crucial role in the accuracy of the ASR system. To calculate the loss function, which gauges the discrepancy between the predicted output and the actual output (i.e., the ground truth transcription), the classification layer's output probabilities are used. The weights of the neural network are updated based on the value of the loss function during the training process, with the goal of minimizing the loss. The number of neurons in the classification layer depends on the size of the output symbol set.

5. EXPERIMENTAL RESULTS & ANALYSIS

The results of the model are analyzed mainly on the basis of two parameters i.e., WER and Loss.

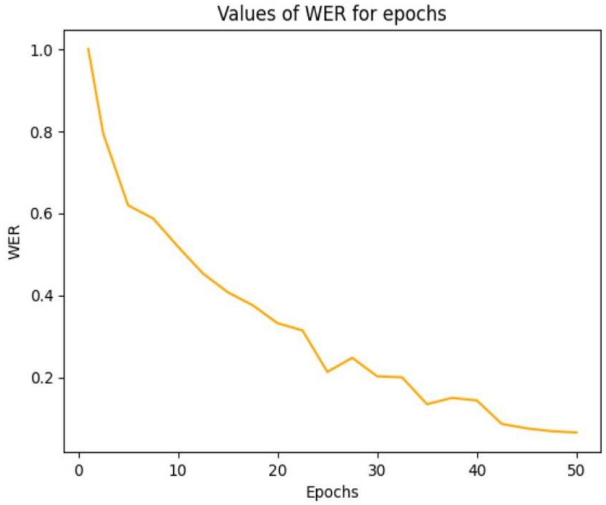

Fig.3. WER vs Epochs

Fig.3. represents the behavior of WER of the model with respect to the epochs. At the start, WER is drastically high and it continues to reduce with an increase in Epochs. Getting the lower WER is one of the main objectives of this project. The WER in the study provided achieved 0.0653 for 50 epochs. This demonstrates the model's successful performance.

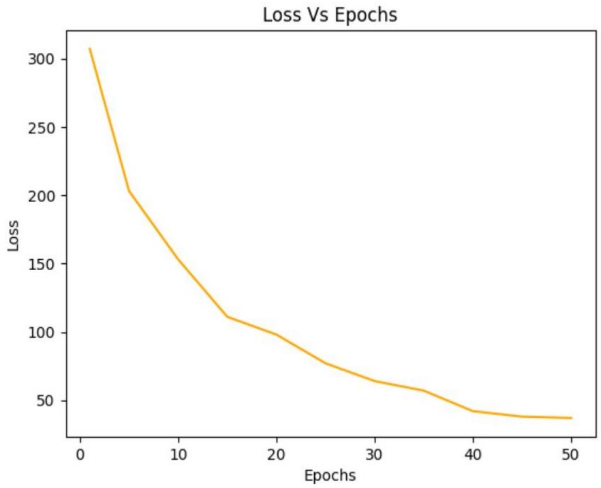

Fig.4. Loss vs Epochs

In Fig.4., the loss of the model for the dataset is represented against the epochs. It is clearly seen that loss is getting reduced with an increase in epochs. This is clearly describing that the proposed model is well-performing.

6. CONCLUSION

In this paper, a combination of 2D CNN and RNN-based models are used for speech-to-text conversion. The CNN model is used for feature extraction and RNN for speech recognition, while the CTC Loss is used for sequential prediction of the letters. The combination of these two models has eventually improved the overall performance of the speech-to-text conversion system, especially in noisy environments where the quality of the speech signal is poor. However, with the specific configuration of the models and the training process we have got the best results even in a noisy environment. Not only validating with the dataset we have also recorded our own voices in .wav format and got the best results. We have taken WER as the key parameter to focus on and modified the model to reduce the word error rate. Moreover, the WER of the model is nearly 6% due to the combination of CNN and RNN along with the CTC loss.

REFERENCES

1 Graves, A., Mohamed, A. R., Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6645-6649).

2 Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.

3 Sak, H., Senior, A., Beaufays, F. (2015). Long short-term memory-based recurrent neural network architectures for large vocabulary speech recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4520-4524).

4 Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... Wu, D. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In International Conference on Machine Learning (ICML) (pp. 173-182).

5 Lu, X., Li, H., Wang, J., Zhang, B., Yao, K. (2020). A survey on end-to-end speech recognition. IEEE Access, 8, 188564-188583.

6 J. D. Tardelli, C. M. Walter, “Speech waveform analysis and recognition process based on non-Euclidean error minimization and matrix array processing techniques”. IEEE ICASSP, pp. 1237-1240, 1986.

7 Takao Suzuki, Yasuo Shoji, “A new speech processing scheme for ATM switching systems”. IEEE, Digital Communications Laboratories, Oki Electric Industry Co. Ltd., Japan, pp. 1515-1519, 1989.

8 J. S. Lim ”Evaluation of a correlated subtraction method for enhancing speech degraded by additive white noise”, IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-26, no. 5, pp.471 -472 1978.

9 R. E. Kalman and R. S. Bucy, “New results in linear filtering and prediction theory,” Trans. ASME Series D, J. Basic Engineering, pp. 95108, 1961.

10 Gabrea, M.: ‘Adaptive Kalman filtering-based speech enhancement algorithm’. IEEE Canadian Conf. on Electrical and Computer Engineering, 2001, vol. 1, pp. 521–526.

11 Jeong, S., Hahn, M.: ‘Speech quality and recognition rate improvement in car noise environments’, Electron. Lett., 2001, 37, (12), pp. 800–802.

12 Ma, J., Deng, L.: ‘Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model’, IEEE Trans. Speech Audio Process., 2003, 11, (6), pp. 590–602.

13 Amodei.D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Chen, Z. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In International conference on machine learning (pp. 173-182). PMLR.

14 Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., ... & Watanabe, S. (2018). Espnet: End-to-end speech processing toolkit. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6114-6118). IEEE.

15 Hannun, A. Y., Maas, A. L., & Ng, A. Y. (2014). Deep speech: scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

16 Zen, H., Agiomyrgiannakis, Y., Chen, Y. C., & Huang, C. C. (2020). LibriSpeech: An ASR corpus based on public domain audiobooks. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 2443-2448).

17 Zhang, Y., Chan, W. K., & Wong, K. F. (2017). Hybrid deep neural network-Hidden Markov Model (DNN-HMM) based speech recognition. In Proceedings of the 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (pp. 146-150). IEEE.

18 Kim, T., & Kim, K. (2017). Convolutional neural networks for speech recognition: An overview. In 2017 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 19-22). IEEE.