Available online @ https://jjem.jnnce.ac.in
https: www.doi.org/10.37314/JJEM.SP0230
Indexed in International Scientific Indexing (ISI)
Impact factor: 1.395 for 2021-22
Published on: 08 December 2024

An Efficient Technique for Identification of Speaker using Neural Network

Niranjan V S[1], Prashant Ankalkoti [2]

[1*]Student, [2]Assistant Professor, Department of Master of Computer

Applications J N N College of Engineering, Shivamogga

niruvs18@gmail.com, psankalkoti@gmail.com

Abstract

Identifying people through the assistance of their voice differentiation is known as speaker recognition. The progress in speaker recognition is being carried on using neural networks and deep learning. One may recognize several of the traditional methods that used feature extraction but did not mimic human speech along with Gaussian Mixture Model (GMM), Universal Background Model and example Mel-frequency Cepstral Coefficients (MFCCs). Deep neural networks like Convolutional and Recurrent, for instance, have advanced to more end-to-end solutions by directly learning from raw audio data. Hence, models of end-to-end have been instituted to improve the work productivity and speed. Improvements like the CNNs with BN and the CTC has enhanced the models and training. This paper also covers the application of BN in voice recognition through the design of an end-to-end CNN effective system.

Keyword: *Mel-Frequency Cepstral Coefficients (MFCCs), Connectionist Temporal Classification (CTC), Batch Normalization (BN).*

1, Introduction

Automatic speaker recognition is moving forward using the employment of the neural networks enhanced by deep learning approaches. Hardware objects, which included features that are built by hand and manually, include Gaussian Mixture Models (GMMs) with Universal Background Models (UBM- GMM) and Mel-Frequency Cepstral Coefficients (MFCCs) for audio recognition systems. Although these algorithms proved very efficient in their respective purposes, theywere not capable of emulating all aspects of human speech. DNNs, especially convolutional and recurrent RNNs, have ushered in a shift toward more data-driven approaches that enhance the model performance and feature extraction. The development of advanced neural networks in recent years has greatly driven a revolution in speaker recognition, as it no longer necessitates feature extraction because they can be learned directly from raw audio data. Due to this change, there are end-to-end models which have been developed to reduce the process of converting inputs from audio to text outputs while at the same time enhancing the speed and efficiency of the models. The area has gone further with techniques including incorporating CNN and LSTM networks for effective feature extraction;the use of Siamese networks for speaker validation. This model uses complex loss functions including cross entropy and triplet loss which boosts recognition efficiency and enhances the capacity of the models to distinguish between different speakers. Together with such features as domain portability, loudness perception, it is applied for speaker identification, diarization, and recognition in conditions that are noisy. augmentation techniques to handle problems of noise in the environment and variability in the speech to be expected. In conclusion, it can be stated that employing neural network methods has greatly enhanced the effectiveness of

speaker recognition systems, which have become more reliable and versatile to be applied in various applications, such as forensic speaker recognition and voice assistants. [15] The authors who presented Speaker Recognition Based on Deep Learning were Zhong Xin Bai et.al. The deep learning advances in two aspects, including speaker verification, identification, diarylation, and robust recognition. They focus on the enhancement of structures that are most suited to the task, loss functions, efficient pooling methods, and feature extraction. Some of the challenges include; dependencies on manually designed features, large model parameters, large amounts of data, and computational resources needed while the acoustic environments in the real world are noisy and are often subjected to speech-overlap. [11] Hossein Saleh Ghaffari in his paper offered a new architecture of CNN for speaker verification which surpasses all such methods including GMM-UBM and i-vectors as it successfully isolates specific attributes of speakers and omits unwanted information. The method utilizes end to end training in a Siamese framework and selection of good pairs. Its use in real-world scenarios and environments is somewhat restricted due to its reliance on the built features such as MFCCs instead of raw audio and the necessity for a significant amount of labeled data and computational resources. To address the South African issues of crime and illicit migration, the feasibility of using neural networks for speaker identification is examined in the paper [8] by Ganesh K Vinayagamoorthy et.al. Here, an attempt is also made to train a neural network using features of voice signal in frequency domain and it reveals acceptable level of accuracy. Yet the approach is limited by text-dependent recognition and less flexible due to this way; to address changes in phonemes and improve the approach's resistance to non-trained voices, further research is needed. The paper by Hardik Dudhrejia et.al [10] reviews neural networks in ASR, with all its applications in human–computer interaction and enhancing the livelihoods of disabled people. It goes on to describe how deep neural networks, essentially RNNs and LSTMs, are more efficient than traditional HMMs. This classifies the ASR Systems into isolated, continuous, connected and spontaneous speech. However, the scope of the paper has been restricted to English only; on the other hand, does not consider problems like variation in accent, background noise, and real-time implementation. A paper by Douglas A. Reynolds et al., Speaker Verification Using Adapted Gaussian Mixture Models [7], describes the deployment of a GMM-based system that uses Bayesian adaptation and likelihood ratio tests for speaker verification. While it performs reasonably well in most test conditions, it still suffers to an extent due to its reliance on low-level acoustic information in nonideal conditions, which is lacking in higher levels such as speaking style. Such higher-level characteristics should be factored into future study with a view to facilitating more accurate and robust outcomes. Mel-spectrograms are utilized in CNNs for speaker identification by Ali Muayad Jalil et.al [2] in the paper, Speaker Identification Employing CNN for Clean and Noisy Speaker Samples. Their proposed CNNs have shown superiority over conventional methods like the i-vector and UBM-GMM methods under noisy conditions. Under clean speech conditions, the effectiveness of the CNNs was poorer, which indicates that further enhancements are in order to realize most superior performance overall concerning different levels of noise. Isolated-word recognition under noisy conditions using CNNs was performed by Ayad Alsobhani et.al [3], who reported 97.06% accuracy on a self-collected dataset of six control words spoken by

30 people in the study. Different isolated and continuous speech experiments proved that the applied CNN model is noise-resistant and robust to variability. However, generalizability is limited to only this reliance upon a small dataset and variety across word length and pronunciation, which tells that more augmentation of data is required and fine-tuning for improvement to get accuracy consistent across a range of settings. By learning from raw speech signals, the CNNs of an end-to-end ASR system avoid the necessity to hand-design features like MFCC, as explored by Vishal Passricha et.al [13]. Experimental results on the TIMIT dataset show that fine-tuned CNN models are significantly more effective in feature extraction and classification compared with traditional systems. The studies based on CNN models for real-world ASR applications should normalize the execution computing complexities and enlarge the datasets for model validation. Chao Li et.al [5] propose the ResCNN and GRU system for speaker embedding in their paper Deep Speaker: anEnd-to-End Neural Speaker Embedding System. Triplet loss is used to project utterancesinto the speaker similarity hypersphere. It makes enormous improvements over previous DNN-based i-vector systems: it achieves 60% improvement in identification accuracy on many datasets and verifies mistakes that drop by half. These challenges include high computing requirements, long training times, and more extensive dataset validation to guarantee real-world robustness. For any realistic deployment, optimization on CPU efficiency and size of models remains ineluctable. In their paper, Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks, Ying Zhang et.al [14]. invent a system that shall implement the CONSTRAINTED RNN, by which CNNs are combined with CTC. Their proposal for an end-to-end framework reduces processing burdens by utilizing CTC, which avoids the necessity of training independent modules, and is more authentic in the sense of deducing possible errors in the feature extraction step related to their modeling of temporal correlations and spectrum changes. It performs very competitively on the TIMIT dataset, thus proving that CNNs are capable of capturing temporal relationships. Long-term dependency modeling remains a challenging task, with proper testing on larger datasets. Applying approaches like Batch Normalization might further enhance stability and performance.

2. Proposed Method

The proposed objective of this project is to design an efficient speaker recognition system, especially based on Convolutional Neural Networks (CNN). Thus, when applying the proposed approach, various benefits associated with the use of CNNs are realized: For instance, the implementation of CNNs enhances the efficiency and utilization of speaker recognition applications. This paper describes our CNN-based system indetail, aided by the latest and highly efficient CTC algorithm for the purpose of speaker recognition. It is distinguished by the fact that itobviates many limitations of previous models and architectures that were investigated using traditional techniques based on older methods. They found that our solution providesa more efficient as well as a more effective wayof processing human speech. The accompanying figure 2.1 shows the schematic view of the proposed approach in this work withsteps and components of the end-to-end CNN-based system. Such a diagram helps in a visual understanding of how this present work deviates and adds to the commonly used methods.

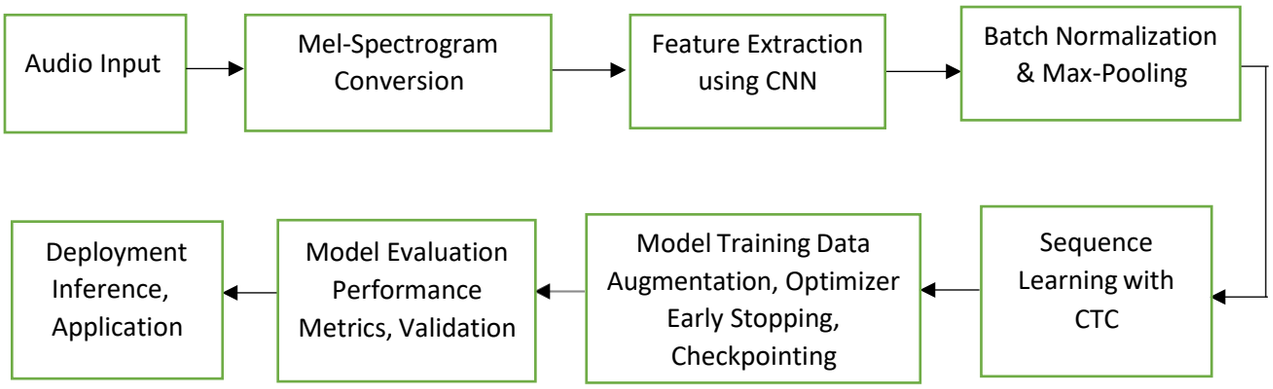

Figure 2.1: Block Diagram of Batch Normalization

The mel-spectrograms of the incoming speech sounds that comprise the TIMIT dataset must be derived before the network can be built. We have several convolutional layers in the network design that is being suggested, and for quick and reliable training, we use Batch Normalization after each layer. To obtain hierarchical feature maps and decrease the spatial dimensions of the input, it is equally necessary to apply max-pooling layers. Nonetheless, to obtain temporal correlations and spectral variations in the voice data, experimental procedures together with the above-mentioned methods have to be followed strictly. To deal with learning stability, the batch normalization known as BN at the beginning was introduced. BN scales the mini-batch samples for each activation layer input and make it standardized with a mean of zero and variance of one. This distils away fine-grained parameter initiation and allows more aggressive learning rates making the training of Deep Neural Networks (DNNs) much quicker. In the proposed model, BN is applied right before the activation layer (like ReLU layer) and in between the convolutional layer. The normalized input $\hat{z}^{(i)}$ is defined as

$$\hat{Z}^{(i)} = \frac{z^{(i)} - E[z^{(i)}]}{\sqrt{Var[z^{(i)}]}}$$

Where z is the d-dimensional input, E[.] is the expectation and Var[.] is the variance.

If the non-linear function of the incoming data is already normalized, layer input normalization will affect the layer representation in that it can strip off the activation non-linearity and retain linearity. In order to address this issue, scaling and shifting factors are recommended. The BN is then calculated as

$$h^{(i)} = \gamma^{(i)}\hat{z}^{(i)} + \beta^{(i)} \quad \ldots\ldots (2)$$

where $\gamma^{(i)} and\ \beta^{(i)}$ are scaling and shifting factors, respectively, that are learned from each feature map and are given by

$$\gamma^{(i)} = \sqrt{Var[z^{(i)}]} \quad \ldots\ldots (3)$$

$$\beta^{(i)} = E[z^{(i)}] \quad \ldots\ldots (4)$$

BN applies scaling and shifting factors, having calculated the mean and variance of each feature map throughout the mini-batch calculation process.

In place of segmenting words before translation, the convolutional neural network approach leverages a connectionist temporal classification (CTC) layer which allows us to directly map between input sequence and

output. Instead of word segmentation prior to transmission it made use of a continuous convolution neurons type layer known as CTC Connectionists Temporal Classification. The techniques provided by stepwise regression have been implemented during the process of learning so as to accelerate convergence (elimination) towards the optimal model. The experiment additionally demonstrated that using the "Adam" optimization method can enable effective adjustments for controlling gradient descent. Metrics like as word error rate (WER) and character error rate (CER) are used to assess the model's performance on the TIMIT dataset, indicating its ability to learn temporal relations essential for precise speaker recognition. The inclusion of Batch Normalization (BN) has contributed significantly to greater network stability during learning, thereby enabling a significant improvement both for transfer between English and other languages and for speaker recognition performance under noisy environments.

3. Results and Discussion

As demonstrated in Figure 3.1, with batch normalization, the loss curves for a CNN-based speaker recognition model smooth out and become stable. This means that with the incorporation of BN, the learning process is more efficient, with quick convergence, few oscillations in loss, and improved generalization that hence improves overall performance. According to the graphical representation shown in figure 3.1, BN can also stabilize training dynamics and improve accuracy. By remedying learning instability and slow convergence common in older approaches, BN improves the results and shows its utility in optimizing voice recognition tasks.

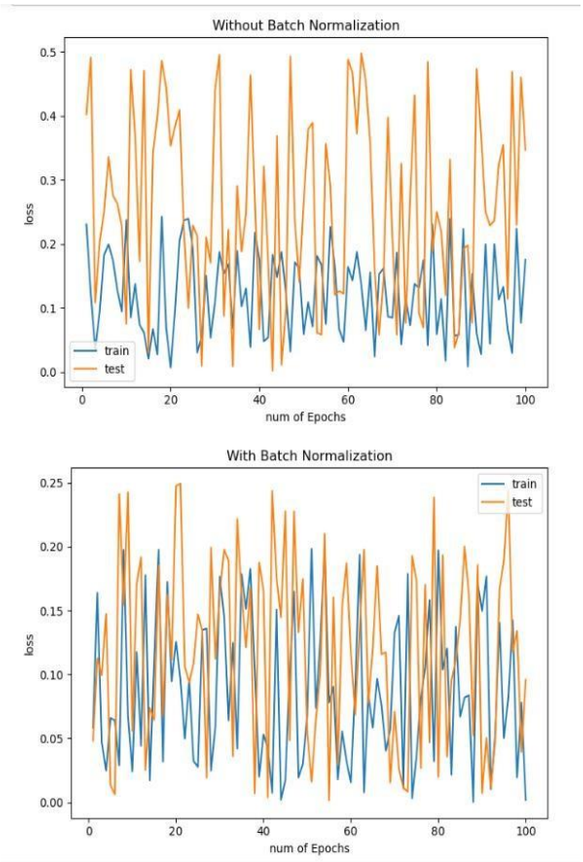

Figure 3.1: Training and Validation Loss Comparison with and without Batch Normalization.

4. Conclusion

With the help of Batch Normalization (BN) incorporated into CNN-based ASR models, its training becomes more stable and efficient. BN results in better generalization and higher model accuracy as the loss values for the training and validation datasets are smaller and decrease in a more harmonious manner. It has advantages that improve the stability and higher convergence rates so CNN based systems can be more robust in various noise and multiple language recognition systems. In particular, BN is shown to be instrumental in the enhancement of the current speaker identification methods that are already in use today.

References

1. Ali Muayad Jalil.et.al. Speaker identification using convolutional neural network for clean and noisy speech samples, Department of Computer Engineering University of Mustansiriyah Baghdad, Iraq 2019.

2. Ayad Alsobhani.et.al. Speech Recognition using Convolution Deep Neural Networks, Faculty of Engineering, Department of Electricity,University of Babylon, Iraq 2021.

3. Brecht Desplanques.et.al. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, ID Lab, Department of Electronicsand Information Systems, imec-Ghent University, Belgium 2020.

4. Chao Li.et.al. Deep Speaker: an End-to-End Neural Speaker Embedding System 2017.

5. Daniel Garcia-Romero.et.al. X-vector DNN Refinement with Full-length Recordings for Speaker Recognition, The Johns Hopkins University, Baltimore, MD 21218, USA 2019.

6. Douglas A. Reynolds.et.al. Speaker Verification Using Adapted Gaussian Mixture Models, M.I.T. Lincoln Laboratory, 244 Wood St., Lexington, Massachusetts 2000.

7. Ganesh K. Venayagamoorthy.et.al. Voice Recognition using Neural Network, Electrical and Computer Engineering, Missouri University of Science and Technology 1998.

8. Gurpreet Kaur.et.al. Speaker and Speech Recognition using Deep Neural Network, I.K Gujral Punjab Technical University, Kapurthala, India 2017.

9. Hardik Dudhrejia.et.al. Speech Recognition using Neural Networks, Department of Computer Engineering, G H Patel College of Engineering & Technology Vadodara, India 2018.

10. Hossein Salehghaffari, Speaker Verification using Convolutional Neural Networks Department of Electrical and Computer Engineering, NYU Tandon School of Engineering (Polytechnic Institute), NY 11201, USA 2018.

11. Li Wan.et.al. Generalized End-to-End Loss for Speaker Verification 2020.

12. Vishal Passricha.et.al. Convolutional Neural Networks for Raw Speech Recognition, National Institute of Technology, Kurukshetra, India 2018.

13. Ying Zhang et al. Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks, Department of Computer. Science and Operations Research, University of Montreal 2017.

14. Zhongxin Bai.et.al. Speaker Recognition Based on Deep Learning: An Overview, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China 2021.