Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 6825555, 12 pages
https://doi.org/10.1155/2022/6825555

Research Article

Improvement in Automatic Speech Recognition of South Asian Accent Using Transfer Learning of DeepSpeech2

Muhammad Ahmed Hassan,[1] **Asim Rehmat**,[2] **Muhammad Usman Ghani Khan**,[3] **and Muhammad Haroon Yousaf**[4]

[1] *Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, Lahore, Pakistan*
[2] *Department of Computer Engineering, University of Engineering and Technology, Lahore, Pakistan*
[3] *Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan*
[4] *Department of Computer Engineering, University of Engineering and Technology, Taxila, Pakistan*

Correspondence should be addressed to Muhammad Haroon Yousaf; haroon.yousaf@uettaxila.edu.pk

Received 26 May 2022; Revised 25 July 2022; Accepted 5 August 2022; Published 4 October 2022

Academic Editor: Muhammad Fazal Ijaz

Automatic speech recognition (ASR) has ensured a convenient and fast mode of communication between humans and computers. It has become more accurate over the passage of time. However, in majority of ASR systems, the models have been trained using native English accents. While they serve best for native English speakers, their accuracy drops drastically for non-native English accents. Our proposed model covers this limitation for non-native English accents. We fine-tuned the DeepSpeech2 model, pretrained on the native English accent dataset by LibriSpeech. We retrain the model on a subset of the common voice dataset having only South Asian accents using the proposed novel loss function. We experimented with three different layer configurations of model to learn the best features for South Asian accents. Three evaluation parameters, word error rate (WER), match error rate (MER), and word information loss (WIL) were used. The results show that DeepSpeech2 can perform significantly well for South Asian accents if the weights of initial convolutional layers are retained while updating weights of deeper layers in the model (i.e., RNN and fully connected layers). Our model gave WER of 18.08%, which is the minimum error achieved for non-native English accents in comparison with the original model.

1. Introduction

Automatic speech recognition (ASR) is a key component in making human-computer interaction (HCI) hassle-free because it is the most interactive and convenient mode of communication between automated systems and humans [1]. The interaction between human and voice-based systems is mostly accomplished in English language. Some of the common applications of voice-controlled systems in AI world are chatbots, humanoid robots, healthcare systems [2], self-driving cars, surveillance systems, industrial robots, and many more. From these applications, chatbots are the most widely used till date. They are everywhere around us. Regardless of region state or country, everyone is using it. Google Assistant [3], SIRI [4], Cortana [5], and Alexa [6], all of them are helping humans in each step of their daily life [7]. Form booking an appointment for barbers' shop to remainder for daily grocery items. Over half of the world's population now has a mobile phone of which 17% of them are smartphones [8], which means around a billion smartphones and 700 million windows users [9] out there in this world. And all of them are equipped with chatbots. Mobile phones come with SIRI or Google Assistant and Windows users have Cortana.

Similar to chatbots, self-driving cars have similar issues. According to Google Waymo, autonomous vehicles have reached 8 million test miles in July 2018 [10]. This clearly states, how close humans are to bringing 5th level of autonomous vehicles to reality. These autonomous systems are designed to operate through voice commands with English according to the British Council, 1.75 billion people in this world speak English [11]. America has the biggest share with around 268 million native English-speaking people [12], UK is second with 59.6 million speakers [13], Europe is third with collectively 70 million people, who speak English as a native language, [14] and Canada is fourth with 19 million native speakers [15]. Although the abovementioned countries have more population of native English-speaking people, but there are several non-native English speakers, present in these countries as well. Apart from this, there is number of counties, which uses English as a second language and their accent is different than that of native English speakers.

For instance, Pakistan and India are holding the biggest share of non-native English speakers, i.e., around 88.6 million people in Pakistan and 125 million in India speak English as a secondary language [16]. Due to regional differences, South Asians and Gulf countries vocal accent for English is different from native English-speaking countries [17]. This highlights the issue of usability, for the voice-controlled systems with dissimilar accents. Which eventually generates hurdles in the use of discussed human voice-controlled systems. Although ASR is being used widely, it is not flexible enough for non-native English accents—thus, 290 million people are unable to use its applications properly.

The motivation behind this research area is given as follows:

(i) Make more accurate the English ASR for non-native English speakers

(ii) Use existing ASR system's accuracy and make it more robust by adding an additional pipeline for non-native English Speakers.

(iii) Make model learnable for a small amount of data

For the last two decades, Hidden Markov's model (HMMs) and Gaussian mixture models (GMMs) were very effective in improving the recognition accuracies of ASR. However, in recent few years, deep neural network (DNNs) [18] has replaced GMMs although the remaining part of GMM-based recognition architecture is still kept for several experiments. These systems are called hybrid ASR systems [19] because they use classic HMM/GMM-based architecture and after the training, they replace the GMM with DNN. Likewise, recurrent neural networks (RNNs) are used as well in a similar manner for language modeling.

There is a plethora of work regarding ASR. For most of the ASR systems, models are trained using a native English accent. Some of the described models have achieved above 90.0% accuracy. Like Google's model for ASR [20] can be as accurate as the human itself in some cases. They claimed of achieving 95.0% of accuracy. But this model is not performing well for non-native English accents until now. Research and development in ASR continuously getting better [21] with large-scale pieces of training, deeper network architectures, and reduced word error rate (W.E.R) have been providing efficient results for ASR. Microsoft AI and Research Lab published research shows 5.1% W.E.R on 2000 switch-board evaluation set by adding additional acoustic model architecture to their system [22]. They called it CNN-bidirectional long short-term memory networks (CNN-BLSTM). Their work clearly reflects closeness to human efficiency. But the problem still occurs when it comes to South Asian accents.

In deep learning, the learning of the model is directly proportional to the amount of data. The more data model has for training, it learns more and makes general decision boundaries. Sometimes, if the model needs some modification in class labels then it requires training from scratch. The solution to this problem was provided by transfer learning [23]. Through transfer learning, the learning of the model can be transferred to new similar problems with some modifications to the last layers of the model. We can use the trained weights of the model and find the best weights for the last layer, that is, called the classification layer.

Recently, DeepSpeech2 [24] provides a deep learning-based architecture that gave promising results in English and Mandarin languages. The deep speech architecture consists on 1D convolutional layers, RNN layers, and fully connected layers. The DeepSpeech2 architecture gives awesome results for two very different languages. It means it has the capability to learn the features of different languages. So, we decided to evaluate DeepSpeech2 on non-native English speakers and improve the quality of DeepSpeech2 by transfer learning.

The most basic limitation in training these models is the limited available dataset, i.e., non-native English speakers' dataset is limited and not widely available. Consequently, most of the models are unable to recognize non-native English accents. To cover this gap, we proposed a system for the recognition of English language, specifically for non-native English accent speakers. Our system will recognize and generate a transcription of human voice using deep learning model named DeepSpeech2 [25]. Our proposed solution will address the following points to reduce the word error rate (W.E.R) on South Asian accents for English language automatic speech recognition (ASR).

(1) We propose a hybrid model based on DeepSpeech2 with two pipelines that learn both English and non-native English accent.

(2) We propose a novel loss function that optimizes the model's weights better.

(3) We achieve 18.08% for non-native English and 7.0% WER for English accent. 2 [26] model by fine-tuning its model on using non-native English accent dataset.

2. Literature Review

Automatic speech recognition is not new to this era of 4th industrial revolution wave. It all started in the middle of 19th century. In 1950, researchers from bell labs build a system named "Audrey" [27] to recognize a digit for single person [28]. Audrey was a six-foot-high relay rack, capturing considerable power in addition to streams of cable. It was capable of recognizing digits from speech using phonemes.

Although in the 1950s computer systems were not so good, they had limited computational speed and memory. But Audrey was perfect in recognizing digits from 1 to 9 with more than 90% of accuracy [29]. It also produces above 70% accuracy in the case of some selected unknown speakers. But Audrey was not comfortable with unknown voices, which means lesser accuracy. From 1960 to 1970, most of the exploration and phonetic segmentation work was completed. Some major techniques used for ASR in 60 s were brute force approach and template. It was good in results but it is hard to scale. A big breakthrough in that era was speech understanding research (SUR), the project of DARPA [30].

Later on, in the 1980s some of remarkable discoveries were found. First appearance of hidden Markov models (HMM) [31] changed the way of speech recognition. With HMM, neural networks also played their role. Layered feed-forward networks with sigmoid function are used to train the model for speech recognition [32]. A three-layer net was constructed and a back-propagation learning procedure is used to train the network. Results of these neural nets are better than HMMs. Time delay neural network (TDNN) achieved 1.5% of word error rate over HMM's 6.3% of word error rate [33]. With the development of HMM and neural networks, many breakthroughs were achieved. DARPA started new speech projects, HMM become popular. Rabiner at bell labs performs well using HMM, AT&T performed the first large-scale deployment of speech recognition named (voice recognition, call processing) VRCP [34, 35]. Automated systems were deployed, in the late 1990s United Airlines launches an automatic flight information system.

In the last decade of research and development in the area of ASR, sensor networks [36] computer vision [37], and natural language processing grows rapidly. Deep learning innovation played a vital role in it. In a recently published Microsoft research [38], very impressive results had been recorded. LSTM was preferred on RNN-LMs to achieve proficiency in reducing W.E.R. Models were built using CNTK. Human versus machine errors is analyzed, which indicates substantial equivalence. In this research, NIST 2000 dataset was used, which produces 4.9% of word error rate [39]. But NIST 2000 dataset was originally recorded from calls with native English accent. That is why it is not very much accurate with South Asian accent.

Kadyrov et al. [40] proposed an ASR based on spectrogram images of speech signals. They achieved 98.34% accuracy. But they used a self-generated dataset. They did not evaluate the model on different accents of English speakers, we consider the different accents of the English speakers and evaluate them on standard benchmark.

Another method to gain efficiency in automatic speech recognition was through active learning. In it, a gradient-based active learning method was used [41]. Active learning aims to label only the most informative data. It helps in reducing labeling costs. In a result, it outperformed the confidence score method used in ASR. Deep learning approaches have achieved significant accuracies in ASR. CNN is key player in achieving these accuracies. Mostly less than 10 layers CNN architecture is used to design models for learning features. But Yisen Wang proposed deep and wide CNN architecture. This architecture is known as RCNN-CTC, it consists of residual connections and connectionist temporal classification loss function [25]. Resulting in 14.92% W.E.R on WSJ dev93 and 6.52% W.E.R on Tencent chat datasets.

One of the core difficulties in automatic speech recognition is noise. Because real-time speech data are filled with different noises like background noises, sampling rates, and codec distortion. Google recently published research to overcome this issue [42]. They trained their model on 162,000 hours of speech. Their goal is to make a generally robust system. Previously most of the models are domain robust, like noise. Google applied various techniques to ensure the robustness of the system by using multiple codecs for encoding inputs in the presence of background noise [43]. More interestingly their model performs very well in new unseen conditions. Their multidomain model trained on 10 hours of data outperformed a model trained for 700 hours of speech data on a new domain only [44].

The survey of previous ASR methodologies is also described in Table 1.

All of the abovementioned outstanding results provide an overview of ASR history and development. But there is one common problem with all these systems, which is non-native English accents. All datasets, used for those training, are recorded or collected from native English-speaking sources, which are different from South Asian English accents. Our proposed system is for the recognition of English language specifically for the non-native English accent as Asian accent is different from native English language, by reducing the (W.E.R) on South Asian accent. Our framework will perceive and create interpretation of human voice utilizing DeepSpeech2, where a few adjustments are suggested in the system layers.

3. Methodology

This paper makes a contribution toward automatic speech recognition for English language in native English and non-native English accents. This research work is inspired by DeepSpeech2 model for English and Mandarin languages. The whole system architecture is shown in Figure 1. The first step involves the preprocessing of the dataset, the second part is feature extraction from audio signals, and the third part is proposed two pipelined CNN-RNN models. The last stage of the system is the decoder, which is used for post-processing of the predicted transcriptions. Each module of the proposed system is explained below:

3.1. Data Preprocessing.

Common voice (CV) dataset was not recorded in a controlled environment, which means that volunteers used their own devices for the recording of CV dataset. The recording was completed with the help of microphones and Internet browsers. Due to the fact that the recording took place in an uncontrolled environment, too much noise was introduced in the background of recorded audio, for instance, the distortion at the beginning of the audio, similar to the noise generated by the microphone,

Table 1: Literature survey of previous methodologies.

Year	Paper	Methodology	Language	Accent	WER
2021	[45]	Unsupervised CNN	English	—	12.83
2021	[46]	RNN encoder-decoder + CTC	English	—	8.3
2021	[47]	Quantifying bias	Dutch	—	16.85
2022	[48]	Transformers	English	—	39.9
2022	[49]	Semisupervised learning	English	—	9.4
2022	Proposed	Multi pipeline feature extractor learning	English	Native + South Asian	7

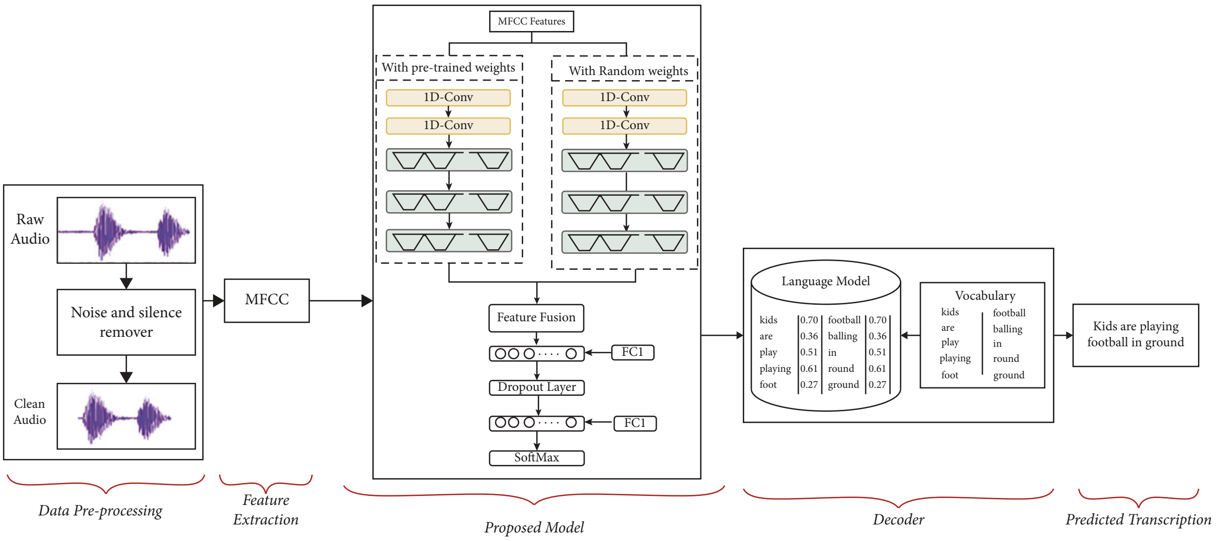


Figure 1: System architecture of the proposed system.

when plugged into the port. Moreover, it has empty gaps (unnecessary silences) between the words and sentences, for example, the speaker starts speaking after 0.5 to 1.5 second delay, and sometimes takes a long stay in between two sentences while reading paragraphs. Thus, CV dataset was useless in raw form and required tons of cleaning and preprocessing. We have cleaned all of South Asian separated audio files by removing the noise and deleting silences between sentences. We employed a self-generated Algorithm 1 to scan each audio file and perform the following activities on it to make sure data are useable for training and prediction. The preprocessing steps include as follows:

(1) Deletion of empty audio files

(2) Elimination of unnecessary silences between the sentences and words

(3) Removal of loud noisy sounds from the beginning of recordings

(4) Extraction of audio file in the FLAC format as per the requirement of network

In order to remove silence and loud noisy sounds, we used zero crossing rate (ZCR) methodology [50]. It is observed that speech section of audio file computes a low zero crossing value and in silent parts, it gives a higher zero crossing value [51]. It is because of the fact that zero crossing count indicates frequency, which is concentrated by energy in the spectrum of voice signals. Vocal sounds are produced by repeated flow of air through the glottis by excitation of the vocal tract, which usually generates a low zero crossing count. Whereas speech other than voice is formed by a narrow vocal tract to cause turbulent airflow that will eventually result in noisy sound and outputs a high zero crossing count.

$$\text{Z.C.R} = \frac{1}{T-1}\sum_{t=1}^{-1} I_R < 0\left(S_t, S_{t-1}\right), \qquad (1)$$

where $S = \text{Signal}$, $T = \text{Length of Signal}$, $T = \text{time}$,

After filtering audio files, we saved it to FLAC format because FLAC files are better in audio quality than mp3. The visual representation of audio signal before and after preprocessing is shown in Figure 2.

3.2. Feature Extraction. The MFCC features are used as input for the model. In the last few decades, these features show very excellent results in automatic speech recognition, semantic analysis through speech, gender classification, and emotion recognition through speech. MFCC features are calculated by the given equation as follows:

$$C_n = \sum_{j=1}^{J} \log(E(j))\cos\left[n\left(j-\frac{1}{2}\frac{\pi}{m}\right)\right], \quad \text{for } n = 1, 2, 3, \ldots, k, \qquad (2)$$

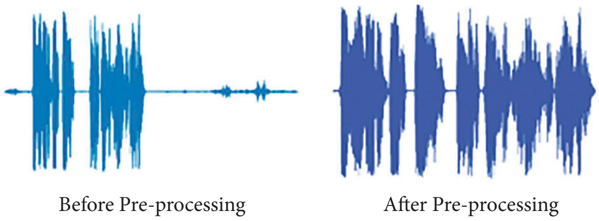


Figure 2: Audio signal processing to remove unnecessary silence and noise.

here E (J) is DFT bin energies that are obtained by the following equation:

$$E(J) = \sum W_k(f)|x(f)|^2. \tag{3}$$

The proposed model also implements the attention mechanism to weight the extracted features. The functionality of the attention layer can be expressed through the following equation:

$$C(i) = \sum_{i=1}^{n} a_i f_i, \tag{4}$$

here a_i represents the attention score of ith features and f_i is the actual value of ith feature.

3.3. Proposed Network. The proposed network architecture of DeepSpeech2 has been used with an extra pipeline for non-native English accents and a novel loss function described in the loss function section. The DeepSpeech2 architecture contains two 1D convolution layers, three bidirectional RNN layers, and one fully connected layer. The 1D convolution layers extract the features from the signal by convolving the 1D-kernel over the Mel frequency cepstral coefficients (MFCC) features, where RNN layer is a state full layer, that extracts the temporal information from the features extracted by convolution layers. The fully connected layer then predicts the text using the features extracted by both convolutional and RNN layers. The proposed model consists of two pipelines of convolution layers. These two pipelines are introduced to extract features of English and non-native English accents. As the DeepSpeech2 model was trained on the English accent dataset so, we used the DeepSpeech2 model with its trained weights to extract English accent features. The structure of the second pipeline is same as DeepSpeech2 model having initial DeepSpeech2 weights that are further fine-tuned using non-native English accent dataset.

MFCC features have been used as input for both pipelines of model. First convolution layer contains 32 filters of size 11 × 41 × 1 with a stride size of 3 × 2. Second convolution layer contains 32 filters of size 11 × 21 with a stride size of 1 × 2. Both convolution layers perform padding to avoid the down sampling of data. Three bidirectional RNN layers are stacked followed by convolution layers. The last bidirectional RNN layer of both pipelines out 2,048 features. The features from both pipelines are concatenated and further passed to FC1 layer having 4,096 neurons. Now the proposed network extracts double features (2048 + 2048 = 4096) than deep speech features (2048). The excess of features leads the network to overfit. To avoid overfitting dropout layer is used after FC1 layer. An attention layer is also introduced after the dropout layer for weighting the features from English and non-native English accent pipelines. The weighted features are further passed to FC2 layer having a number of neurons equal to vocabulary size. The softmax layer is used to predict the probability of each character. The probabilities of each character can be calculated by the following equation (5):

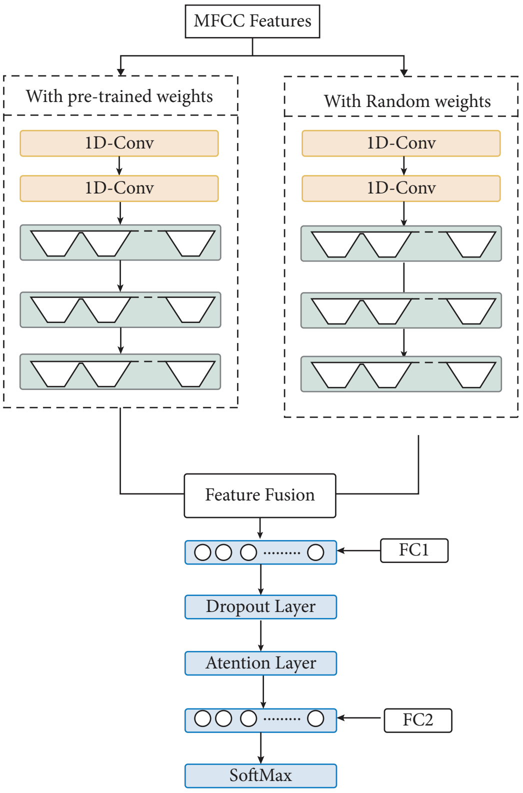


Figure 3: Proposed network architecture.

$$p(c) = \frac{e^c}{\sum_{k=1}^{k} e^c}, \tag{5}$$

here p (c) shows the probability of z class and e^c shows the score of z class that is produced by FC2 layer. K represent the size of the vocabulary.

The architecture of the proposed network is shown in Figure 3.

3.4. Decoder. The transcriptions produced by the model are mostly correct without English language constraints like spacing and sentence boundary, etc. To handle this problem a language model is used by Amodei [24]. We extend the vocabulary of LibriSpeech dataset [52] with the common voice dataset's transcriptions. A decoder is developed using a language model and vocabulary that accepts the predicted transcriptions from RNN model and produced the

```
Read CSV of common voice dataset
Choose "Pakistani," "Indian," "Dutch," and "Sri-Lankan" accent rows from country column
Select "filename" columns from CSV file
Create a dictionary for JSON file writing
for $i$ in filename do
    flacConvert($i$)
    SilenceRemover($i$)
    Duration($i$)
    Finding transcript with respect to audio file
    for lineIndex in range($x$) do
        transcript = split[lineIndex]
        Write JSON file
        data = {"audio_filepath": "1.FLAC," "duration": duration, "text": transcript}
        with open('json-other-train.txt," "a") as outfile:
        json.dump(data, outfile)
    end for
end for
flacConvert($i$) (Convert mp3 -> FLAC)
Calculate audio file length
Calculate audio file sample rates
Calculate duration = audio file length/sample rate
Calculate loudness
Calculate peak amplitude
Splitting audio file into chunks with silence > 0.5 second (considering it silent if quieter than −16 dBFS)
for chunk in enumerate (chunks) do
    List.Append(chunk)
end for
```

Algorithm 1: Dataset preprocessing algorithm.

Table 2: Sample transcriptions generated by proposed model and decoded by the language model.

RNN output	Decoder output
Kids are playing foot balling round	Kids are playing football on the ground
She is trying together toys	She is trying to get her toys
Birds flying roups	Birds fly in groups

transcription that satisfies the English language constraints as shown in Table 2.

4. Experiments and Results

In this section, the experimental setup of model training, evaluation measure, and results are discussed.

4.1. Experimental Setup. The training is done for three different layer configurations of the network. In configuration A, the convolutional layer freezes, and learning of RNN layer and FC layers takes place by modifying their weights. Configuration B is made by freezing the convolution layer and FC layer and modifying weights of RNN layer. In configuration C, the RNN and FC layers are frozen and let the convolution layers learn. These three configurations are shown in Table 3.

Table 3: Different configuration settings for non-native English accent pipeline.

Configuration	Freeze layer	Training layer
A	None	Conv, RNN, and FC
B	Conv and FC	RNN
C	Conv	RNN and FC

While training, one pipeline for non-native English accent was trained and the other pipeline was used as is with pretrained weights of the original DeepSpeech2 model. The English accent pipeline was freezed by setting the learning rate 0 for all layers. The proceeding implementation details for non-native English accent pipeline.

The training of model is done using ReLU activation in all of the layers. The proposed loss function as discussed in Section 4.1.1 is used as a criterion. Stochastic gradient descent (SGD) optimizer is used with dynamic learning rate, starting from $l = 1 \times e^{-3}$ with decay rate of $1 \times e^{-1}$. The model is trained for 200 epochs. The experiments are performed on the system having Nvidia's 1080 Ti GPU having 3584 Cuda cores and 11 GB cud memory. The system contains 16 GB DDR3 RAM and a 2.8 quad-core processor. The total dataset is split in a 70–30 ratio. A total of 70% of the data are utilized for learning the model and the remaining 30% of data are used to evaluate the learning of the model. The model's architecture is designed and trained on paddle paddle framework. Some of the main python libraries for data manipulation and audio signal processing are used such as Pandas, Numpy, SciKit, and PyAudio.

The model hyperparameters are shown in Table 4.

4.1.1. Loss Function. Most existing ASR models were optimized using some cost functions that take the predicted output y' and ground truth y. The difference y and y' is used to optimize the model parameters. The proposed optimization technique intends to reduce the feature gap between the South Asian and European accents. So, the proposed loss calculation function uses the features of South Asian and European accent audio from FC2 layer, and the mean square difference of these features is used as a loss to optimize the model parameters as shown in the following equation:

$$d(a,e) = \frac{1}{N}\sqrt{(a_1 - e_1)^2 + (a_2 - e_2)^2 + \cdots + (a_i - e_i)^2 + \cdots + (a_n - e_n)^2}, \tag{6}$$

here a is feature vector of South Asian accent and e is feature vector of European accent on the same transcription and N is number of samples. The objective of this loss function is to decrease the difference between the feature vector of South Asian and European accents to reduce the accuracy gap in ASR for South Asian and European accents.

4.2. Dataset. We have used common voice [53] dataset for fine-tuning of DeepSpeech2 [54] models. common voice dataset was recorded for more than 18 languages by Mozilla for the purpose of research. It consists of total 1087 hours of audio files, from which 780 hours were validated with transcription. This dataset was recorded with both male and female voices with the ratio of 47% and 11% at a sampling frequency of 16 kHz. A detailed description of dataset accent according to the region is listed in Table 5.

As stated earlier, we are focusing on non-native English accents, so we used a subset of this dataset by filtering out Pakistani, Indian, Dutch, and Sri-Lankan speaker's recordings, with the help of Algorithm 1. Because accent variation is affected by the geographical area in which the speaker grows up and lives as well as by factors such as social class, culture, education, and working environment. All of these factors have an impact on the accuracy of the automatic speech recognition system. After splitting desired recorded files, we got a total of 10,219 audio files with an average playtime of 5 seconds. Table 6 shows the further splitting of training and testing classes accordingly with respect to gender.

4.3. Evaluation Measures. The model evaluation parameters that we have selected are word error rate (WER), match error rate (MER), and word information rate (WIR).

4.3.1. Word Error Rate. WER is one of the most common evaluation parameters for ASR models and it provides a good comparison between the results of our proposed model and other related work done so far. The WER tells the rate of error in the transcript generated by ASR by comparing it to the original transcript. It can be calculated by the following equation:

Table 4: Parameters and their values while implementation.

Parameters	Value
Operating system	Ubuntu 18.04
Frame work	Paddle paddle
Language	Python 3.7
CPU	Core-i7 (7th gen.)
RAM	16 GB (DDR3)
GPU	1080 Ti (11 GB memory, 3584 cores)
Batch size	256
Epochs	200
Drop out	0.4
Learning rate	$l = 1 \times e^{-3}$
Loss function	Proposed Section 4.1.1
Optimizer	SGD

Table 5: Percentage of different English accents in common voice dataset.

Share (%)	Accent
23	United States English
9	England English
4	India and south Asia (India, Pakistan, and Sri Lanka)
3	Canadian English
3	Australian English
1	New Zealand English
1	Southern African (South Africa, Zimbabwe, and Namibia)
1	Scottish English

$$\text{W.E.R} = \frac{S + I + D}{N}, \tag{7}$$

where N is the total number of words spoken in the original transcript.

$$N = S + D + H, \tag{8}$$

here, S is the number of substitutions. I is the number of insertions, D is the number of deletions, and H is the total number of hits, i.e., correctly transcribed words.

4.3.2. Match Error Rate. Match error rate tells the probability of given input-output word matches being incorrect. It can be calculated by equation.

$$\text{M.E.R} = \frac{S + I + D}{H + S + I + D} = 1 - \frac{H}{N}. \tag{9}$$

Unlike in WER, here N is sum of all four terms.

$$N = H + S + I + D. \tag{10}$$

4.3.3. Word Information Loss. Word information loss gives the probability of any input word is matched with an equal output word and vice versa. It can be calculated by equation.

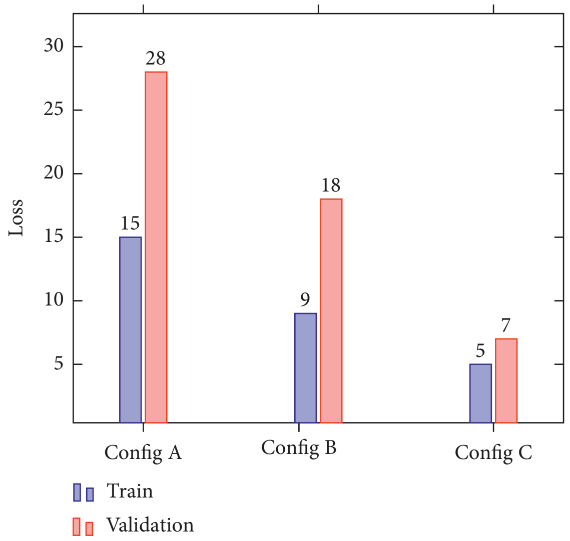


Figure 4: The image shows a calculated loss after training and validation at config A, config B, and config C.

$$\text{W.I.L} = \frac{H^2}{(H+S+I)(H+S+D)}. \quad (11)$$

All three evaluation metrics represent the errors and loss in the output, hence, the lesser the value is, the better the model predicts. However, WER is not an actual percentage as it has no upper bound limit because of the insertion I parameter. So, WER can only be used to compare different models while MER and WIL can be interpreted as how well the model performs.

4.4. Results. Before any modifications happened in DeepSpeech2, we loaded pretrained model on LibriSpeech dataset. As the majority of speakers of this dataset were from US, the model achieved WER of around 6% on the test set on US English accents. As the model was trained on American people and the test was also containing American people, that is why we obtained 6% WER. However, the same model, when evaluated using common voice dataset, gave a drastically high WER of 43% for South Asian accents. The reason for 43% WER is pretrained model of DeepSpeech2, which is not trained for non-native American or non-native English guys, that is why we use parallel pipeline for processing for the non-native English peoples. This is called a learning network. When it finds the input of non-native speakers for English it learns and updates its weights accordingly, meanwhile, we have frozen the learning rate for the freeze network, hence, we can save the performance of our model for the native English users and weights would not be changed for this pipeline. Whereas if the user is not-native English speaker then the learning network will entertain that user and update the weights. That is why our model is working better than other models due to its learning controls.

The training of DeepSpeech2 is done in three different layer configurations as mentioned in Table 3. The loss comparison graph of configurations A, B, and C is shown in Figure 4. The purpose of experimenting through these configurations is that we want to make DeepSpeech2 perform better for South Asian accents by transfer learning. We experimented with different dropout ratios and the best results were achieved with a dropout ratio 0.7. All the results shown here of different configurations have the same dropout ratio of 0.7.

(1) For configuration A, where all layers were learned using a common voice dataset, the model achieved W.E.R of 35.35% on the validation set. The training and validation cost of this modification is shown in Figure 5(a).

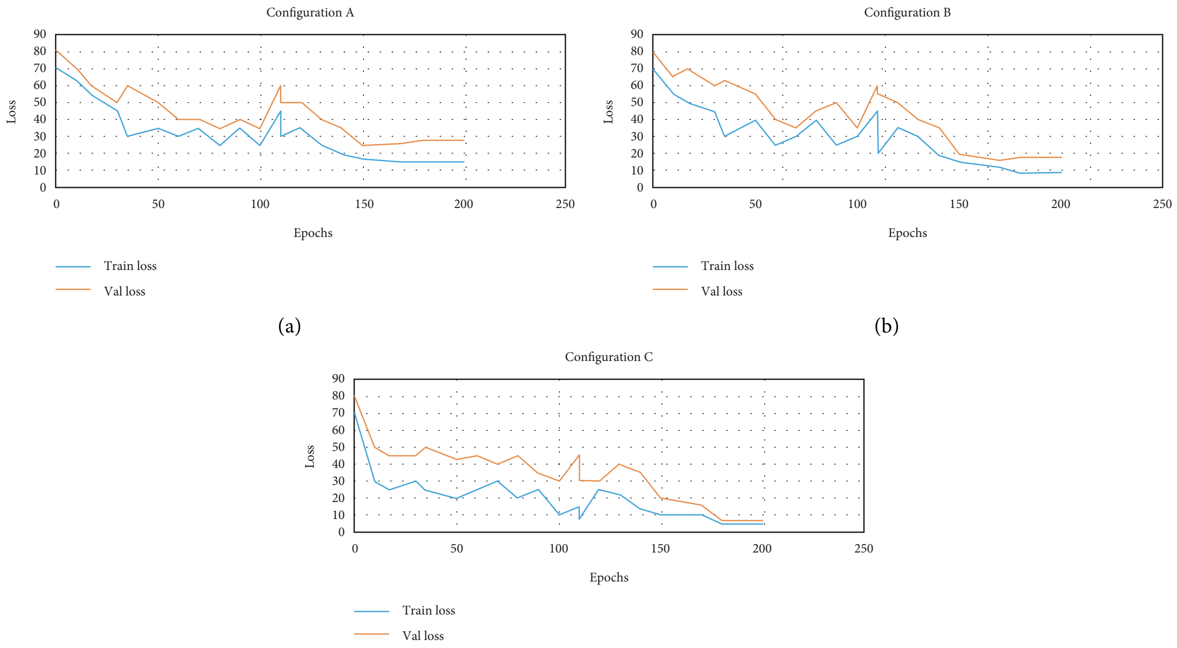
(2) For configuration B, the weights of RNN layer have been learned by freezing the both CNN and FC layers. In this configuration, the model retained its low-level features learned from LibriSpeech dataset and learned only the new high-level features from the common voice dataset. The WER achieved on the validation set is 20.419%. Training and validation cost for these configurations is shown in Figure 5(b).

(3) Finally, for configuration C, we learned the weights of RNN and FC layers by freezing the convolutional layer. W.E.R achieved on the validation set is 18.0859%. The cost of training and validation is shown in Figure 5(c).

(4) We performed fine tuning on both the European accent and South Asian accent datasets. The results deduced after fine-tuning showed that it behaved differently for both datasets, as for European accent the W.E.R increased from 6.8% to 7.0% whereas for South Asian accent W.E.R is reduced from 51% to 18.08%.

The WER, MER, and WIL rate of the model on test set are shown in Table 7.

We contrasted our DeepSpeech2 algorithm to Apple Dictation, Bing Speech, Google Speech API, and wit.ai, which are all for profit speech technologies. Our test is



(a)

(b)

(c)

Figure 5: Training is done for 200 epochs. The cost or loss of training in different configurations is shown. (a) The trend of loss function for training and validation of the model in configuration A, (b) the trend of loss function for training and validation of model in configuration B, and (c) the minimum loss for training and validation of model in configuration C.

Table 6: Total number of audio files in training and test set of non-native English accents with respect to gender.

Utterances	Training set	Test set
Male	4,444	1,112
Female	3,730	933
Total	8,174	2,045

Table 7: Results of all three-layer configurations of the proposed model in terms of WER, MER, and WIL.

Config	South Asian accent			Western accent		
	WER (%)	MER (%)	WIL (%)	WER (%)	MER (%)	WIL (%)
A	35.35	29.68	30.98	10.36	12.65	11.22
B	20.419	18.54	19.35	9.68	8.69	9.85
C	18.0859	15.36	15.25	7.0	6.8	7.2

Table 8: The comparison between nonoptimized and optimized models with respect to WER and CER.

Model	Nonoptimized		Optimized	
	WER	CER	WER	CER
DeepSpeech2	15.57	4.52	—	—
DeepSpeech2 KENLMc	10.46	3.68	10.45	2.96
DeepSpeech2 KENLMo	10.75	3.79	10.66	2.89
DeepSpeech2 and KENLM (c + O)	9.9	3.61	9.91	2.80

intended to monitor success in noisy situations. This circumstance complicates the evaluation of web audio APIs: whenever the SNR is just too small or, in certain situations, whenever the phrase is too lengthy, such algorithms will provide no results. As an outcome, we limit our analysis to phrases under that all algorithms gave a not-a-void outcome. Table 8 shows the outcome of assessing each system on our test files.

Table 9: Comparison of DeepSpeech2 and proposed model's WER on English and non-native English accent.

Model	WER (English accent) (%)	WER (non-native English accent) (%)
DeepSpeech 2	6	43
Proposed	7	18.08

The comparison of the proposed model and DeepSpeech2 model is shown in Table 9.

5. Conclusion

ASR is being used extensively to enable natural human-machine interaction, but not pliable enough for South Asian accent for English language for which almost 200 million people are unable to use its applications. Our contribution towards resolving this obstacle is the proposed system, that is, inspired by DeepSpeech2. The proposed method provides the two pipelined deep learning architectures that achieve minimum character error rate (CER) and word error rate (WER) on common voice (CV) benchmark. By setting up different experimental configurations and modifications, we are successful in achieving minimum WER, that is, reduced from 43% to 18.08% at a lower validation cost. As this work focused on South Asian's English accents so, there is a little bit of increase in WER and CER for English speakers. The system will be further scalable towards targeting other South Asian languages like Bengali, Urdu, Hindi, and others with more robust datasets and higher accuracy and the training of both pipelines parallelly.

Data Availability

There are two datasets that are used in experiments for the proposed research. The first one is LibreSpeech dataset is audio signals in English language. The total length of the dataset is 1000 hours. The annotation is provided along the dataset in form of a transcription of the audio signal. The readers can find the dataset at this (https://www.openslr.org/12) link. The second dataset, that is, used for transfer learning of DeepSpeech2 is a common voice dataset version 7.0. This dataset also includes the audio signal and their transcription in English language. The total length of dataset is 2637 hours and 75879 different voices. The size of the total dataset is 65 GB. The reader can find the dataset at this (https://commonvoice.mozilla.org/en/datasets) link.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. Rafaeli, "From new media to communication," *Sage Annual Review of Communication Research: Advancing Communication Science*, vol. 16, pp. 110–134, 1988.

[2] N. R. Pradhan, A. P. Singh, S. Verma et al., "A novel blockchain-based healthcare system design and performance benchmarking on a multi-hosted testbed," *Sensors*, vol. 22, no. 9, p. 3449, 2022.

[3] G. López, L. Quesada, and L. A. Guerrero, "Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces," in *Proceedings of the International Conference on Applied Human Factors and Ergonomics*, Springer, Los Angeles, CA, USA, 2017.

[4] J. F. Weaver, *Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws*, ABC-CLIO, Beijing, China, 2013.

[5] P. Domingues and M. Frade, "Digital forensic artifacts of the cortana device search cache on windows 10 desktop," in *Proceedings of the 2016 11th International Conference on Availability, Reliability and Security (ARES)*, pp. 338–344, IEEE, Salzburg, Austria, September 2016.

[6] V. Kepuska and G. Bohouta, "Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home)," in *Proceedings of the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 99–103, IEEE, Las Vegas, NV, USA, January 2018.

[7] M. B. Hoy, "Alexa, siri, cortana, and more: an introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[8] C. Shepard, A. Rahmati, C. Tossell, L. Zhong, and P. Kortum, "Livelab: measuring wireless networks and smartphone users in the field," *ACM SIGMETRICS—Performance Evaluation Review*, vol. 38, no. 3, pp. 15–20, 2011.

[9] Y. F. Chang, C. Chen, and H. Zhou, "Smart phone for mobile commerce," *Computer Standards & Interfaces*, vol. 31, no. 4, pp. 740–747, 2009.

[10] J. A. Pozzi, *Weaponization of artificial intelligence*, Ph.D. dissertation, Utica College, Utica, NY, USA, 2018.

[11] M. A. Hasman, "The role of English in the 21st century," *TESOL Chile*, vol. 1, no. 1, pp. 18–21, 2004.

[12] J. C. Bennett, *The Anglosphere Challenge: Why the English-speaking Nations Will Lead the Way in the Twenty-First Century*, Rowman & Littlefield, Lanham, MD, USA, 2004.

[13] M. Berns, K. De Bot, and U. Hasebrink, *The Presence of English: Media and European Youth*, Vol. 7, Springer Science & Business Media, Berlin, Germany, 2007.

[14] A. Ädel and B. Erman, "Recurrent word combinations in academic writing by native and non-native speakers of English: a lexical bundles approach," *English for Specific Purposes*, vol. 31, no. 2, pp. 81–92, 2012.

[15] G. Andreou and I. Galantomos, "The native speaker ideal in foreign language teaching," *Electronic Journal of Foreign Language Teaching*, vol. 6, no. 2, pp. 200–208, 2009.

[16] M. Azam, A. Chin, and N. Prakash, "The returns to English-language skills in India," *Economic Development and Cultural Change*, vol. 61, no. 2, pp. 335–367, 2013.

[17] R. M. McKenzie, "Social factors and non-native attitudes towards varieties of spoken English: a Japanese case study," *International Journal of Applied Linguistics*, vol. 18, no. 1, pp. 63–88, 2008.

[18] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal*

Processing, pp. 8599–8603, IEEE, Vancouver, Canada, May 2013.

[19] D. T. Toledano, M. P. Fernández-Gallego, and A. Lozano-Diez, "Multi-resolution speech analysis for automatic speech recognition using deep neural networks: experiments on timit," *PLoS One*, vol. 13, no. 10, Article ID e0205355, 2018.

[20] T. Ashwell and J. R. Elam, "How accurately can the google web speech api recognize and transcribe Japanese l2 English learners' oral production?" *The JALT CALL Journal*, vol. 13, no. 1, pp. 59–76, 2017.

[21] M. A. Al Amin, M. T. Islam, S. Kibria, and M. S. Rahman, "Continuous Bengali speech recognition based on deep neural network," in *Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–6, IEEE, Cox'sBazar, Bangladesh, February 2019.

[22] A. Nicolson and K. K. Paliwal, "Deep XI as a front-end for robust automatic speech recognition," 2019, https://arxiv.org/abs/1906.07319.

[23] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[24] D. Amodei, S. Ananthanarayanan, R. Anubhai et al., "Deep speech 2: end-to-end speech recognition in English and Mandarin," in *Proceedings of the 2016 International Conference on Machine Learning*, New York, NY, USA, 2016.

[25] A. Hannun, C. Case, J. Casper et al., "Deep speech: scaling up end-to-end speech recognition," 2014, https://arxiv.org/abs/1412.5567.

[26] Y. Yang, A. Lalitha, J. Lee, and C. Lott, "Automatic grammar augmentation for robust voice command recognition," in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6376–6380, IEEE, Brighton, UK, May 2019.

[27] S. Joshi, A. Kumari, P. Pai, S. Sangaonkar, and M. D'Souza, "Voice recognition system," *Journal of Research*, vol. 2, p. 11, 2017.

[28] K. Ifrat, K. Israt, I. H. Saimun, and F. Akter, *Articulatory feature based automatic speech recognition using neural network*, Ph.D. dissertation, United International University, Dhaka, Bangladesh, 2018.

[29] R. Pieraccini and I. Director, "From audrey to siri," *Is Speech Recognition a Solved Problem*, vol. 23, p. 123, 2012.

[30] D. S. Pallett, "Benchmark tests for darpa resource management database performance evaluations," in *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Glasgow, UK, 1989.

[31] M. Rahul, R. Agrawal, N. Kohli, and K. Hbtu, "Layered recognition scheme for robust human facial expression recognition us-ing modified hidden Markov model," *Journal of Multimedia Processing and Technologies*, vol. 10, no. 1, p. 18, 2019.

[32] M. A. Tahir, H. Huang, A. Zeyer, R. Schlüter, and H. Ney, "Training of reduced-rank linear transformations for multi-layer polynomial acoustic features for speech recognition," *Speech Communication*, vol. 110, pp. 56–63, 2019.

[33] S. Jain, A. Goel, and P. Arora, "Spectrum prediction using time delay neural network in cognitive radio network," in *Smart Innovations in Communication and Computational Sciences*, pp. 257–269, Springer, Berlin, Germany, 2019.

[34] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, 2019.

[35] S. Tang, X. Qin, X. Chen, and G. Wei, "Video quality assessment for encrypted http adaptive streaming: attention-based hybrid RNN-HMM model," in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, UK, 2019.

[36] P. N. Srinivasu, G. JayaLakshmi, R. H. Jhaveri, and S. P. Praveen, "Ambient assistive living for monitoring the physical activity of diabetic adults through body area networks," *Mobile Information Systems*, vol. 2022, Article ID 3169927, 18 pages, 2022.

[37] A. Vulli, P. N. Srinivasu, M. S. K. Sashank, J. Shafi, J. Choi, and M. F. Ijaz, "Fine-tuned densenet-169 for breast cancer metastasis prediction using fastai and 1-cycle policy," *Sensors*, vol. 22, no. 8, p. 2988, 2022.

[38] A. Singh and A. M. Joshi, "Speaker identification through natural and whisper speech signal," in *Optical and Wireless Technologies*, pp. 223–231, Springer, Berlin, Germany, 2020.

[39] R. Shirani Faradonbeh and A. Taheri, "Long-term prediction of rockburst hazard in deep underground openings using three robust data mining techniques," *Engineering with Computers*, vol. 35, no. 2, pp. 659–675, 2019.

[40] S. Kadyrov, C. Turan, A. Amirzhanov, and C. Ozdemir, "Speaker recognition from spectrogram images," in *Proceedings of the 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, pp. 1–4, IEEE, Nur-Sultan, Kazakhstan, April 2021.

[41] D. M. Ziegler, N. Stiennon, J. Wu et al., "Fine-tuning language models from human preferences," 2019, https://arxiv.org/abs/1909.08593.

[42] N. Amara, J. Olmos-Peñuela, and I. Fernández-de Lucio, "Overcoming the "lost before translation" problem: an exploratory study," *Research Policy*, vol. 48, no. 1, pp. 22–36, 2019.

[43] T. Vaillancourt and M. Jelinek, "Method and system for encoding left and right channels of a stereo sound signal selecting between two and four sub-frames models depending on the bit budget," 2019, https://patents.google.com/patent/EP3353784A1/en.

[44] R. Gale, L. Chen, J. Dolata, J. van Santen, and M. Asgari, "Improving ASR systems for children with autism and language impairment using domain-focused DNN transfer techniques," in *Proceedings of the Interspeech 2019*, Graz, Austria, 2019.

[45] J. Ni, L. Wang, H. Gao et al., "Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition," 2022, https://arxiv.org/abs/2203.15796.

[46] J. Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[47] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," 2021, https://arxiv.org/abs/2103.15122.

[48] B. Bharathi, B. R. Chakravarthi, S. Cn, N. Sripriya, A. Pandian, and S. Valli, "Findings of the shared task on speech recognition for vulnerable individuals in Tamil," *Diversity and Inclusion*, pp. 339–345, 2022.

[49] Y. Zhang, D. S. Park, W. Han et al., "Bigssl: exploring the Frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–14, 2022.

[50] M. Jaber, "Voice activity detection method and apparatus for voiced/unvoiced decision and pitch estimation in a noisy

speech feature extraction," 2007, https://patents.google.com/patent/US20070198251.

[51] L. Firmansah and E. B. Setiawan, "Data audio compression lossless flac format to lossy audio mp3 format with huffman shift coding algorithm," in *Proceedings of the 2016 4th International Conference on Information and Communication Technology (ICoICT)*, pp. 1–5, IEEE, Bandung, Indonesia, May 2016.

[52] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, IEEE, Brisbane, Australia, April 2015.

[53] S. M. Zeitels, R. R. Casiano, G. M. Gardner, N. D. Hogikyan, J. A. Koufman, and C. A. Rosen, "Management of common voice problems: committee report," *Otolaryngology-Head and Neck Surgery*, vol. 126, no. 4, pp. 333–348, 2002.

[54] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," 2017, https://arxiv.org/abs/1701.02720.