Applied Acoustics 190 (2022) 108643

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Spectral warping and data augmentation for low resource language ASR system under mismatched conditions

Mohit Dua [a], Virender Kadyan [b,*], Neha Banthia [c], Akshit Bansal [a], Tanya Agarwal [a]

[a] Department of Computer Engineering, National Institute of Technology, Kurukshetra, India
[b] Speech and Language Research Centre (SLRC), School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India
[c] Department of Computer Engineering, Indian Institute of Information Technology, Sonepat, India

ARTICLE INFO

Article history:
Received 10 May 2021
Received in revised form 14 December 2021
Accepted 13 January 2022
Available online 29 January 2022

Keywords:
Children speech recognition
Formant modification
TDNN
TTS
MFCC
FDLP

ABSTRACT

The performance of an Automatic Speech Recognition System (ASR) system deteriorates while using it on children speech, due to large variations and mismatch of acoustic and linguistic variables between spoken utterances of adults and children. Another important reason for the low efficiency of ASR models is the data scarcity of children speech data for low resource-language like Punjabi. The proposed work in this paper tries to address the both challenges i.e. acoustic and linguistic variations challenge, and data scarcity problem, thereby improves performance of a children speech ASR system for Punjabi language. To handle the first issue of acoustic and linguistic variations, the proposed work uses formant modification as a spectral warping technique to reduce the variation between children speech and adult speech. Further, to address the second issue of data scarcity, this paper discusses training of ASR models on augmented children speech data. Also, the work combines well established Mel-Frequency Cepstral Coefficients (MFCC) features extraction technique with Frequency Domain Linear Prediction (FDLP) to propose MFCC-FDLP hybrid approach for front end feature extraction. For implementing the data augmentation, Tacotron 2, an end-to-end Text to Speech (TTS) generative model has been used. The proposed work uses MFCC, FDLP and hybrid MFCC + FDLP techniques for front end feature extraction, Time Delay Neural Network (TDNN) for backend acoustic modeling, and trigram language model to implement continuous Punjabi language ASR systems. To increase robustness of the proposed ASR system, we have included a batch of lexically diverse words in our pronunciation model which achieved a relative improvement of 29.44%.

1. Introduction

Speech is considered as the primary means of communication between humans. But nowadays communication is not just limited to humans, but even to machines. Automatic Speech Recognition (ASR) is a technique used to facilitate interaction between machines and humans. In the modern era, consumers reap the benefits and ease provided by the device that utilizes Automatic Speech Recognition. For instance, speech based virtual assistants like Amazon Alexa, Google Assistant and Apple Siri are very popular, offering a wide variety of services like controlling smart home devices, following voice commands to perform different tasks [1,2]. Research on automation of simple tasks that require human machine interaction has attracted a lot of attention in the last few decades [3,4]. A lot of research, studies and data collection have been done for high resource languages like English, Spanish etc. However, there is a big scope of improvement for the low resource and regional languages such as Punjabi [5]. There are many challenges in building ASR systems in regional languages such as data scarcity, high cost involved in building transcripts and acoustic variability.

For effective training of ASR models, a large amount of speech data is required. In low resource languages like Punjabi, the ASR performance drops dramatically when the amount of training data is reduced [6]. Collecting such a large amount of data has its own challenges. Data scarcity causes an overfitting problem, since training data is not sufficient for the ASR system to work properly. Data scarcity can be tackled using data augmentation, which is a very popular method to incorporate speaker/acoustic variability in training speech data to increase robustness of ASR systems. Further, inter speaker variability such as age, gender, accent, speaking rate and formant frequencies of the speakers also pose difficulties.

* Corresponding author.
E-mail addresses: er.mohitdua@nitkkr.ac.in (M. Dua), vkadyan@ddn.upes.ac.in (V. Kadyan).

https://doi.org/10.1016/j.apacoust.2022.108643
0003-682X/

Data augmentation also handles the speaker variability problems by generating multiple readings of the training utterances from different speakers. There are various existing Data augmentation techniques such as Generative Adversarial Networks (GAN) [7], Prosody Modification [8], spectrogram augmentation [9]. Data augmentation is widely categorized as in-domain and out domain techniques. In-domain data augmentation is done by using voice conversion (VC) to alter the acoustic attributes. It includes techniques like GAN [7], Vocal Tract Length Perturbation [10] and Stochastic Feature Mapping (SFM) [11]. Out-domain data augmentation uses unseen utterances to enhance existing dataset like speed prosody modification [8]. To make the ASR model to recognize acoustic and lexically diverse utterances, TTS (end to end speech synthesis model) is used to include unseen utterances in the training data, resulting in a more robust recognizer.

There is a lot of research done on ASR for Punjabi language using adult speech corpus, exploring feature extraction techniques like MFCC [12] and Perceptually based Processing (PLP) [13], FDLP [14], acoustic models methods such as Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) and/or Deep Neural Networks (DNN) [15], and language modeling approaches like tri-phone, mono-phone phoneme models and N-grams. Most of the ASR system available publicly works well with Punjabi adults' speech i.e., high-SNR (Speech to Noise Ratio) speech [16]. However, in case of low-SNR speech or in the case of children speech, the system performance collapses. These mismatched conditions occur due to training on adults' speech and then testing on children's speech [17]. The difference in the vocal tract dimension between the adults and children is the root cause of this mismatch [18]. Kumar et al. proved that formant frequencies decrease with the increase in vocal tract [17]. The range and amount of change in formant frequencies are smaller among older age groups than younger age groups [19]. Hence, issues such as linguistic variations and data scarcity are still posing a challenge for the low resource language such as Punjabi.

Motivated by these issues, the work in this paper contributes by using formant modification as a spectral warping technique for handling linguistic variation problem, and implementing training of ASR models on augmented children speech data for addressing data scarcity problem. Also, for front end feature extraction, the work proposes hybrid approach by combining well established Mel-Frequency Cepstral Coefficients (MFCC) features extraction technique with Frequency Domain Linear Prediction (FDLP) to propose MFCC-FDLP hybrid approach. For implementing the data augmentation, Tacotron 2, an end-to-end Text to Speech (TTS) generative model has been used. The proposed continuous Punjabi language ASR systems is developed using Time Delay Neural Network (TDNN) based acoustic modeling, and trigram based language model. Further, the work includes a batch of lexically diverse words in pronunciation model which increase robustness of the proposed ASR system.

The remainder of the paper is organized as: Section 2 discusses the literature Survey. Section 3 describes fundamentals of Tacotron 2 and formant modification. Section 4 discusses the proposed architecture. Section 5 discusses experimental setup and analyses the results, followed by Section 6 that concludes the proposal.

2. Literature survey

ASR systems require a large amount of training data for a system to work reasonably well. High resource languages like English, Spanish, Chinese have many state of the art speech corpuses for effective training of ASR Systems. However, such large amount of data of regional languages like Hindi, Punjabi etc., is not available and still, only a few researchers are working for the same [20]. In fact, it is estimated that for only about 1% of the world languages, required speech corpus, which is needed to train an ASR, is available [21].

Potamianos et al. [22] performed one of the oldest works in the field of child ASR models. They implemented age-dependent acoustic variability that reduced WER (Word Error Rate) by 10%. Shahnawazuddin et al. [23] explored prosody modification to map mismatch between adult speech and children's speech. The experiment resulted in improvement of the baseline system approximately by 50%. Kumar et al. [17] proposed a study that highlighted the large variation and mismatch in acoustic and linguistic attributes between children's and adults' speech. The paper implemented a linear predictive coding (LPC)-based formant modification [24] method to reduce the difference between adults' and children's speech, which in turn improved the performance of children speech ASR systems. The proposed technique improved the system performance over a hybrid DNN-HMM baseline model, vocal tract length normalization (VTLN) and speaking rate adaptation (SRA). Also, the proposed method was tested for noisy channels as well.

It is a challenging task to collect a traditional text-to-speech corpus for a low-resource language. It is difficult to synthesize speech on the "found" data [25], that is why Cooper proposed to use various sources of available "found" data. Oord et al. [26] introduced Wavenet, a deep neural network of time domain waveforms, which is significantly used in the complete Text-to-Speech synthesis model. Their paper implemented Wavenet in the TTS system for English and Mandarin languages. Li et al. [27] proved that the correct ratio of synthetic data and natural data can also improve the results. The authors accomplished their work via Global Style Token (GST). The improvements were done by increasing the depth of the recognition network and through hyper parameter tuning. Acoustic data perturbation, semi-supervised training, multilingual processing and speech synthesis are some of the techniques for data augmentation, which were proposed by Ragni et al. [28]. Recently, Rosenberg et al. [29] have used speech synthesis architecture of Tacotron to improve performance of speech recognition system. The work used speech synthesis mechanism to enhance acoustic diversity and lexical diversity by creating new training utterances and synthesizing training data with different speaker characteristics, which in turn generates new training utterances, respectively. The proposed study was implemented on two corpora from distinct domains for high resource language i.e., English.

In last few years, some researchers have explored the different possibilities to improve performance of Punjabi ASR by doing improvements in traditional feature extraction techniques and using advanced backend models. Guglani and Mishra [16] explored various feature extraction techniques like MFCC, PLP and compared their performance for Punjabi dataset. The paper also experimented with the mono-phone and tri-phone model using the N-gram language model and did a comparative study to find the best alternative. Gerosa et al. proposed several methods for speaker adaptive acoustic modeling to cope with inter speaker variability to make the ASR model more efficient [30]. The paper claimed that vocal-tract length normalization (VLTN) method with constrained MLLR based speaker normalization (CMLSN) method performs better than other methods. Kadyan et al. [31] discussed a comparative study of deep neural network-based Punjabi-ASR systems. The authors experimented for robust feature extraction techniques to bring high performance in Punjabi speech recognition systems. Kadyan et al. [32] also reduced acoustic mismatch by working on VLTN, explicit pitch and duration modification. All the three explored techniques proved to be effective.

Further, Shen et al. [33] implemented Tacotron 2, a neural network architecture for text to speech synthesis models. This paper combines a sequence-to-sequence Tacotron-model that generates

mel spectrogram and the modified Wavenet vocoder. The proposed model can be trained directly on the data and significantly improve the performance of ASR models by producing state-of-the-art sound quality that is very close to natural human speech. TTS produces significant results in producing synthetic speech that sounds almost like human speech. However, to train ASR models on single-speaker, abundant training data from a professional voice talent is required [34]. Deng et al. [35] proposed in his study that multi-speaker models can outperform single-speaker models when large amounts of training data of single-speakers are not available.

The front end feature extraction also plays an important role in implementation of an ASR system. For many years, MFCC remained the only choice for developing ASR systems. However, in last two decades many other feature extraction methods such as GFCC [20], FDLP [14] have shown their presence in ASR field. FDLP method was first introduced by J. Herre et al. [36] as a method for efficient coding of transients in transform coders. In [14], Athineos et al. has proposed a novel representation of the temporal envelope in different frequency bands by exploring the dual of conventional linear prediction method. With this technique of frequency-domain linear prediction (FDLP), the 'poles' of the model describe temporal peaks. In [37], Thomas et al. highlights the new feature extraction technique which utilized short-term spectral envelope and modulation frequency features. These features are derived from sub-band temporal envelopes of the estimated speech using FDLP method. Also, recently, researchers have shown improvements in the performance of ASR systems by using hybrid features for implementing front end feature extraction [20].

These recent speech synthesis advancements can also help in improving the performance of a low resource language such as Punjabi. In this paper, a small self-created children speech corpus of Punjabi language has been used. Rosenberg et al. [29] found that improvements in speech recognition performance is achievable by augmenting training data with synthesized material. Their results show the relative gain of 4% to WER using Tacotron-2 on the LIBRISPEECH corpus [38]. Motivated by the work, we have implemented Tacotron 2 to increase the robustness of the ASR system. With ample amounts of training data, overfitting and acoustic variability issues are tackled. Further, most of the ASR models are trained on adults' speech and don't prove to be robust on children's speech. There is a big difference in acoustic variability between children's speech and adults' speech. Kathania et al. [17] indicated an improvement of 27% in children ASR system performance using formant modification tested on PF-STAR [39]. This inspired us to apply formant modification on our ASR system on children speech. Hence, the proposed work in this paper is to build ASR models on children's speech of Punjabi language using augmented data as well as formant modification.

3. Preliminaries

This section discusses the fundamentals of Tacotron 2 and formant modification used in the implemented work to improve the performance of Punjabi ASR system.

3.1. Tacotron 2

Neural networks are not smart enough to begin with, as a poorly trained neural network is not able to produce the desired results. This happens due to the lack of adequate training data. Hence, to expand the available dataset, minor alterations are done to the existing training data. This process is called data augmentation. Data augmentation enables us to add relevant data, which is related to the way with which neural networks learn. A neural network keeps on becoming better as we feed more data to it. TTS (Text to Speech) is used as a data augmentation process for expanding speech dataset. Through this, we can first train our TTS model and then supply text to it to get corresponding audios. For data augmentation, several techniques are available like GAN, prosody modification, spectrogram augmentation etc. Tacotron 2 used in the proposed work, is also a TTS technique. The most important advantage of using Tacotron 2 is that it is able to generate natural sounding speech directly from text. We do not need to train it providing complex acoustic and linguistic features as input. It incorporates the ideas of Tacotron [29] and Wavenet [26]. It works by mapping a sequence of characters to a sequence of features that encode the audio sample. It uses a sequence-to-sequence model that optimizes the text-to-speech process. The features are an 80-dimensional audio spectrogram with frames computed every 12.5 ms. These features capture word pronunciation as well as characteristics of human speech like speed, volume and intonation. It then uses a Wavenet-like structure to convert these features to a 24 KHz waveform [26].

Fig. 1 shows system architecture for Tacotron 2. It has two parts: (1) a recurrent sequence-to-sequence network to predict features, which is able to predict a chain of *mel* spectrogram frames from an input character sequence; (2) an improved Wavenet that generates time-domain waveforms based on the *mel* spectrograms. In other words, a sequence-to-sequence Tacotron style model is able to generate *mel* spectrograms and a modified Wavenet vocoder is able to utilize these spectrograms to generate time-domain waveform pertaining to the audio sample.

The first part of Tacotron 2 is also called the encoder, whose first layer is the embedding layer with 512 dimensional vectors. The output of this first layer is directed to a block of three one-dimensional convolution layers having 512 filters with a length of 5. The next block in this encoder block consists of a bidirectional long short-term memory (LSTM). The second part of Tacotron 2 is called the decoder. At each decoding step, "attention" forms the context vector and updates the attention weight. The context vector given by cvi is the product of the encoder's output (eo) and attention weights (w). It is mathematically expressed as:

$$cvi = \sum_{j=1}^{T} wij\ eoj \tag{1}$$

The attention weights are calculated using the following formula:

$$w_{ij} = \frac{e^{exp(e_{ij})}}{\sum_{k=1}^{T} \exp(e_{ik})} \tag{2}$$

Here, e_{ij} represents energy. For calculating energy, we use:

$$e_{ij} = w^{T} tanh\left(Ap_{i-1} + Beo_{j} + Cl_{ij} + d\right) \tag{3}$$

where, p_{i-1} represents previous hidden state of LSTM network, eo_j is the j^{th} hidden encoder state,

A, B, C, D are trained parameters. The variable l_{ij} represents location signs, which are calculated as:

$$l_i = F * w_{i-1} \tag{4}$$

where, F being the convolution operation and w_{i-1} is the previous attention weight.

3.2. Formant modification

The performance of an ASR system degrades if the training and testing conditions are different. One of these mismatch conditions occurs, when we train our ASR system with adults' speech and test it on children's speech. The size of vocal tract of adult and child are different, and the children's speech is low-SNR speech. Due to this, the formants and other important features of speech are lost while

training and testing, which in turn drops the accuracy of the ASR system.

A formant is the broad spectral maximum that results from an acoustic resonance of the human vocal tract. It is usually defined as a broad peak, or local maximum, in the spectrum. Studies show the change in formant frequencies in people of different age groups [17]. Since the length of vocal tract is inversely proportional to formant frequencies, the increase in vocal tract decreases the formant frequencies.

Formant modification uses warping on LP (Linear Prediction) spectrum. The resulting LP spectrum is denoted by $R_\Gamma(f)$. It is obtained when we apply warping function $w(f)$ on the original LPC (Linear Predictive Coding) spectrum which is denoted by $R(f)$. Here, Γ is the warping factor.

$$R_\Gamma(f) = R(w_\Gamma(f)) \tag{5}$$

An estimate of speech signal denoted by $R(m)$ is obtained as a linear combination of the M samples values obtained before; this is classical working of the LPC method [24].

$$r(m) = \sum_{j=1}^{M} \Gamma r(m-j) \tag{6}$$

Then we take its Z-transform:

$$\widehat{R}(z) = \left(\sum_{j=1}^{M} \Gamma_j z^{-j}\right) R(z) \tag{7}$$

Here, z^{-j} are j unit delay filters r_j are the LPC filter coefficients. We use these to calculate LPC spectrum. The unit delay filter is replaced by an all-pass filter $F(z)$ to wrap the LPC spectrum. We use first order all-pass filter which is given by

$$F(z) = \frac{(z^{-1} - \Gamma)}{(1 - \Gamma z^{-1})} \tag{8}$$

Here, Γ is the warping factor which lies in the range of -1 and 1.

$$-1 < \Gamma < 1 \tag{9}$$

The warped frequency scale matches psycho-acoustic scale with proper value of Γ which is based on auditory perception. The formants can be shifted systematically on the application of warping function on the LPC coefficients. If Γ is positive, formant frequencies shift to lower frequencies. The residual $(r(m) - \widehat{r}(m))$ and the modified LPC coefficients are then used by standard LPC synthesizer to synthesize the speech signal which is called formant modified signal. This signal is used as input to the ASR system.

4. Proposed approach

This section describes the architecture of the proposed approach. Fig. 2 shows the architecture of the proposed ASR system. The Punjabi children speech corpus is split into two sets; train and test sets, with 80% in training and rest in testing. The training speech data along with its transcriptions is used to train Tacotron 2. Three types of augmented datasets are synthesized – original, random and sampled. Formant modification is applied on adult speech data to reduce acoustic variabilities. This formant modified adult data is combined with augmented data, training adult data and training child data to serve as training data for our ASR system. MFCC technique is used to extract features that along with the training speech data are input to the acoustic model. FDLP is representation of the temporal envelope in different frequency bands by exploring the dual of conventional linear prediction (LPC) when applied in the transform domain [34]. The proposed approach fuses FDLP and MFCC features. The hybrid features generate better results. The acoustic model generates mono-phones which are transformed to tri-phones. They are passed to the decoder. Pronunciation model, also known as lexicon, is created by linguists. It matches phones to words and outputs the probability of the possible words. The language model uses these words to select the one which generates proper meaning in the context of the sentence. Speaker adaptive training is done to make the model robust against unseen speakers. These models work together with the decoder to perform speech recognition. This generates our ASR system which is tested on the testing part of children speech corpus. The following subsection describe these steps in detail.

4.1. Data augmentation

The proposed system uses a TTS model based on Tacotron 2. Fig. 3 shows the architecture of Tacotron 2 model used in the proposed system implementation. As described earlier, Tacotron 2 is a combination of encoder-decoder network with an attention mechanism, and a Wavenet based Vocoder. It takes input as a sequence of text in Punjabi language, which is encoded by encoder. In the first part of the encoder, the character sequence is converted into a word embedding vector. The input text sequence embedding is encoded by 3 convolution layers each containing 512 filters of shape 5×1, followed by a bidirectional LSTM layer of 250 units for each direction. Tacotron 2 uses 'Local sensitive attention' which takes the encoder output as input and tries to summarize the full encoded sequence as a fixed length context vector for each decoder output step.

Decoder is an autoregressive recurrent neural network which predicts a *mel* spectrogram from the encoded input sequence one frame at a time. The output of the attention layer is passed through a small pre-net containing 2 fully connected layers of 256 hidden ReLU (Rectified Linear Unit) units. The pre-net output and attention context vector are concatenated and passed through a stack of 2 unidirectional LSTM layers with 1024 units. The output of the LSTM layer is projected through a linear transform to predict the target spectrogram frame.

The predicted *mel* spectrogram is passed through 5-layer convolution postnet layers, each composed of 512 filters with shape 5×1 filters with batch normalisation, followed by tanh activation on all but the final layer. The postnet layer predicts a residual to add to the prediction to improve the overall reconstruction. Finally, the *mel* spectrogram is transformed into time domain waveforms by modified Wavenet vocoder. The *mel* spectrograms are mapped to a fixed-dimensional embedding vector, known as deep speaker vectors (d-vectors). These d-vectors are frame-level speaker discriminative features that represent the speaker characteristics. Algorithm 1 gives the pseudo code for the data augmentation process used in the proposed work. The proposed system uses following three different approaches to generate d-vectors for inference to handle speaker diversity in the synthesized data.

- **Original:** In this case, d-vector is derived from the training utterance itself. If synthesized utterances are identical to the source, that implies this is perfect synthesis.
- **Sampled:** Here, we use a d-vector from some other utterance that was used during training for inference. In this case, speaker representations will be seen by synthesizer, but the source utterance and synthesized utterance will have different speaker characteristics.
- **Random:** D-vector is generated by a random 256-dimensional vector, and then projects it to the unit-hypersphere via L2-normalization. Random sampling is effective when d-vectors are evenly distributed.

Algorithm 1:*DataAugmentation*

Input: *Audiowavfiles + metadata*
Code:

```
//Initialisation Trainingsamples : Audiowavfiles+
metadatathatisusedtotrainTacotron − 2
//Initialisehyper − parametersinhparamsfile
encoder_n_convolutions = 3 //Encoder
for[iinrangeencoder_n_convolutions]:
 conv_layer = Initialiseiconvolutionlayers
//BidirectionalLSTMlayeriscreated. //Decoder
Ineachiteration : Producemelspectrogram
whichisfedbacktoitsinput
whileTrue : decoder_inputs = Prenet(decoder_input)
  decoder_outputs = Decode(decoder_inputs)
  gate_output = decoder_outputs[1]
  ifsigmoid(gate_output) > gate_threshold :
  break
  decoder_input = decoder_outputs[0]
//Theloopstopswhenagiventhreshold
forthestoptokenisreached.
//BoththeEncoderandDecoderuseLSTMlayers.
//PostnetLayerpostnet_n_convolution = 5
convolution_list = [] for[iinpostnet_n_convolution] :
convolution_list.append(convolutionlayer)
```

Algorithm 2:*FormantModification*

Input: *SpeechWaveform*
Code:

```
// Load a speech waveform
rd = Read(speech_waveform)
wt = create_output_file
for[line in rd] :
  [sampled_data, sampled_rate] = audioread(line)
  // Fit LPC to short time segments
  //'x' is a stretch of signal, fit order 'p' LPC mod els
  // Return the successive all−
pole coefficients as rows of 'a'
  // Return the per − frame gains in 'g'
 and the residual excitation in 'e'
  [a, g, e] = lpcfit(sampled_data , no_LPC mod el)
  // Choose a warping factor (alpha) between −1 to 1
  alpha = 0.1
  //Warpolesfunctionwarpsanall−
polepolynomialbysubstitution
  //Itisdefinedbyrowsofafirst− orderwarpfactoralpha
  //Negativealphashiftspolesupinfrequency
[B, A] = warpoles(a, alpha)
//LycsynthfunctionresynthesizefromLPCrepresentation
  //Eachrowof 'a'isaLPCfittoah − pointframeofdata

'e'isanexcitationsignal
  //Itreturns'd'asaresultingLPCresynthesis
  dw = filter(B(1, :), 1, lycsynth(A, g, e)
  //scalesaudiosignalytothespeakeratsamplerateFs
  soundsc(samplingdata, samplingrate)
  soundsc(dw, samplingrate)
  write(dw, sr)inoutputfile
```

4.2. Formant modification

Formant modification is used to lighten the difference between children's speech and adults' speech. LPC based formant modification produces an improvement of 27% compared to baseline with DNN. Therefore, we have used this process in our proposed approach. It is carried out to the LP spectrum wield wrapping. LP analysis is performed on speech signal that is fed to LPC coefficients and LP residual. Those LPC coefficients give the input to the wrapping function for the modified LPC process. The output of LP Residual and modified LP process generated the Formant modified speech signal going through LP synthesis.

4.3. Feature extraction

In feature extraction step, features are extracted from the formant modified speech signal using fusion of FDLP (Frequency

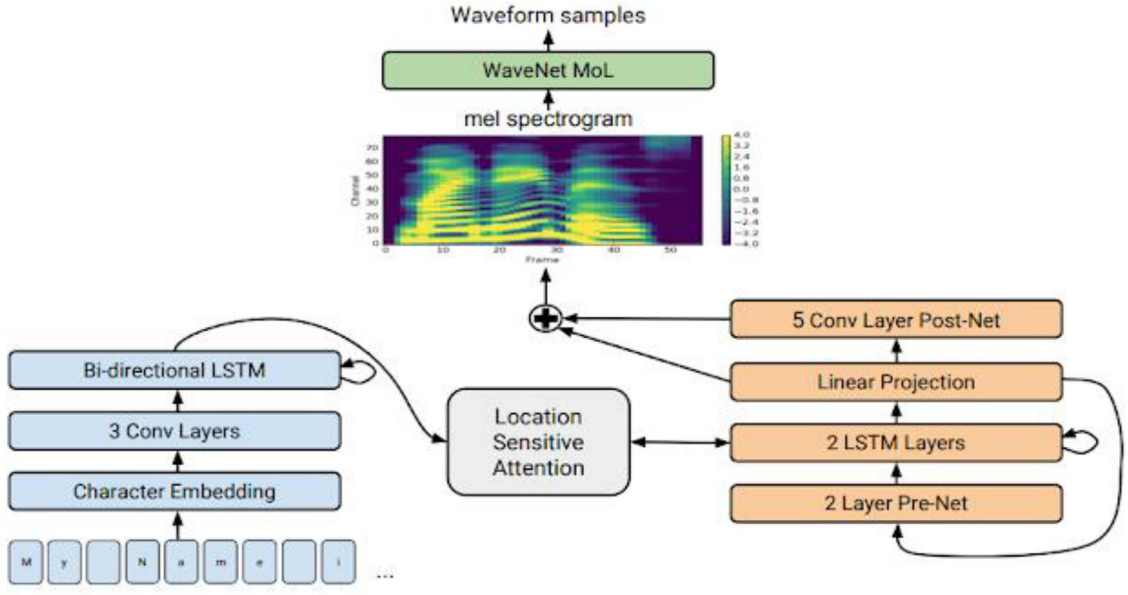

Fig. 1. Tacotron 2 system architecture [40].

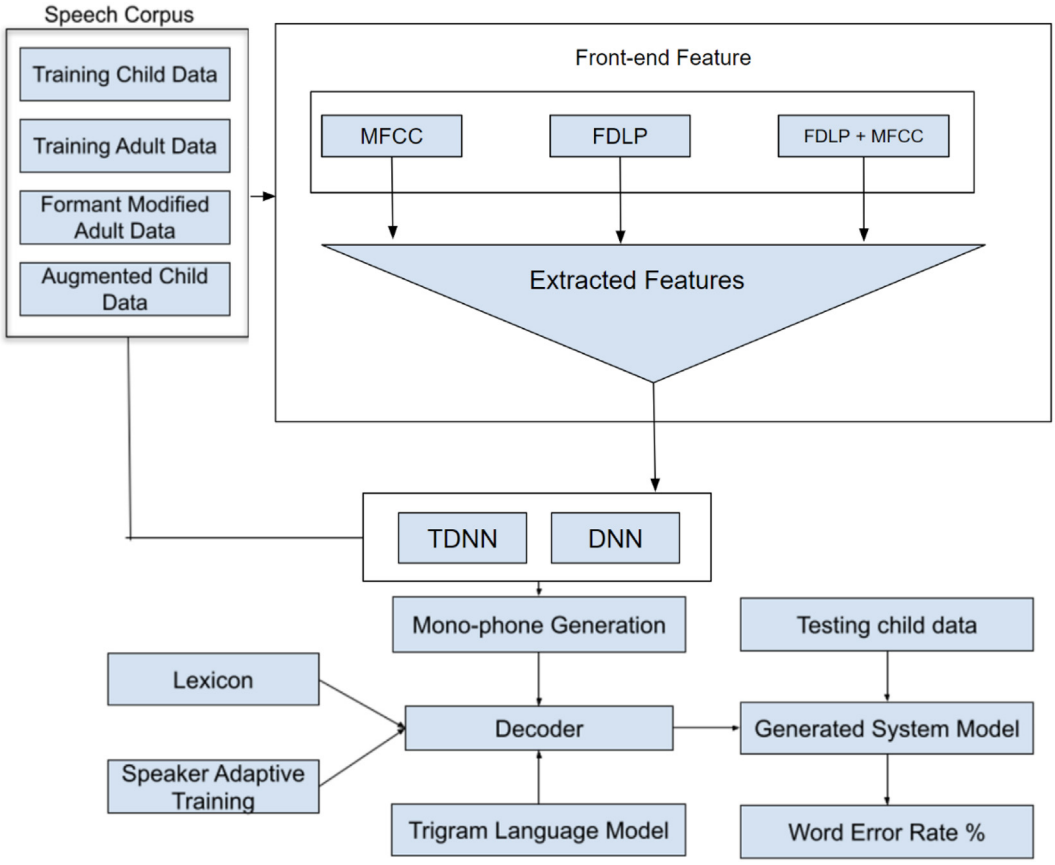

Fig. 2. Proposed ASR system architecture.

Domain Linear Prediction) and Mel-frequency cepstral coefficients (MFCC). Speech Corpus is fed to the A/D convertor to digitize it, and the amount of energy is boosted into the high frequencies in the per-emphasis phase. Windowing is responsible for slicing the audio waveform into sliding frames. DFT (Discrete Fourier Transform) and *mel* filter bank are responsible for bringing out information in frequency domain and mapping the measured frequency, respectively. The log of power spectrum obtained from *mel* filter bank is taken, and then in the cepstrum, the glottal source and the filter is separated. On other hand, FDLP is also used for feature extraction where Modulation feature extraction is employed for long term modulation features. Later, the efficiency of these features are explored on proposed formant modification based data augmentation approach. Our proposed methods uses sharpness of the FDLP poles which takes the location of the sub-band temporal envelopes poles into account, the proposed methods mainly focus on the amplitude of the sub-band time-frequency envelopes. In FDLP as well as in MFCC, 39 coefficients are generated by expanding the first 13 coefficients. The first 13 derivatives are static features and the others are dynamic features generated by taking first (Δ) and second order derivatives ($\Delta\ \Delta$) known as delta feature and delta delta feature, respectively. Although individual 39 FDLP features as well as 39 MFCC features convey richer information. In each approach initial 13th parameter is the energy in each frame which is used for identifying phones. The combination of MFCC and FDLP brings improvement in the ASR system.

4.4. Acoustic modeling

The baseline system has been developed using TDNN acoustic modeling. TDNN is a Time Delay Neural Network. It converges fast and is particularly useful when training data is limited. It uses subsampling to exclude the duplicate weights. It is independent of the relationship between the number of sequence steps and the length of input. Since the duplicate updates are reduced, the amount of training drops [41,42].

The nodes and weights in TDNN are updated only when subsampling is used. Some inputs in the hidden layers are not connected, thus, providing space between the frames. If the interval between frames is allowed, the model can learn all input features because TDNN has a long context going up to the upper layer [31].

Acoustic modeling in Kaldi is a pipeline process. Firstly, GMM-HMM are used to form a context independent acoustic model. Acoustic models are trained using the extracted MFCC features including 13 static and other delta and delta delta features. Secondly, this model is used to train another GMM-HMM model called the tri1 model. This tri1 is a stronger model which can be used for training more complex models. This is a context dependent acoustic model. The tri1 model is converted to tri3 using the best alignments. This sets the baseline for training TDNN based models using Kaldi. TDNN produces phone sequences which are given to the decoder for speech recognition.

Triphone based models perform better since articulation depends on phones before and after too. The acoustic realizations of a phoneme can occur as a result of coarticulation beyond the word boundaries.

4.5. Speaker adaptive training

Speaker Adaptive Training (SAT) [43] is used to reduce inter speaker variability. Variability in speaker independent acoustic models is attributed to both phonetic variation and variation

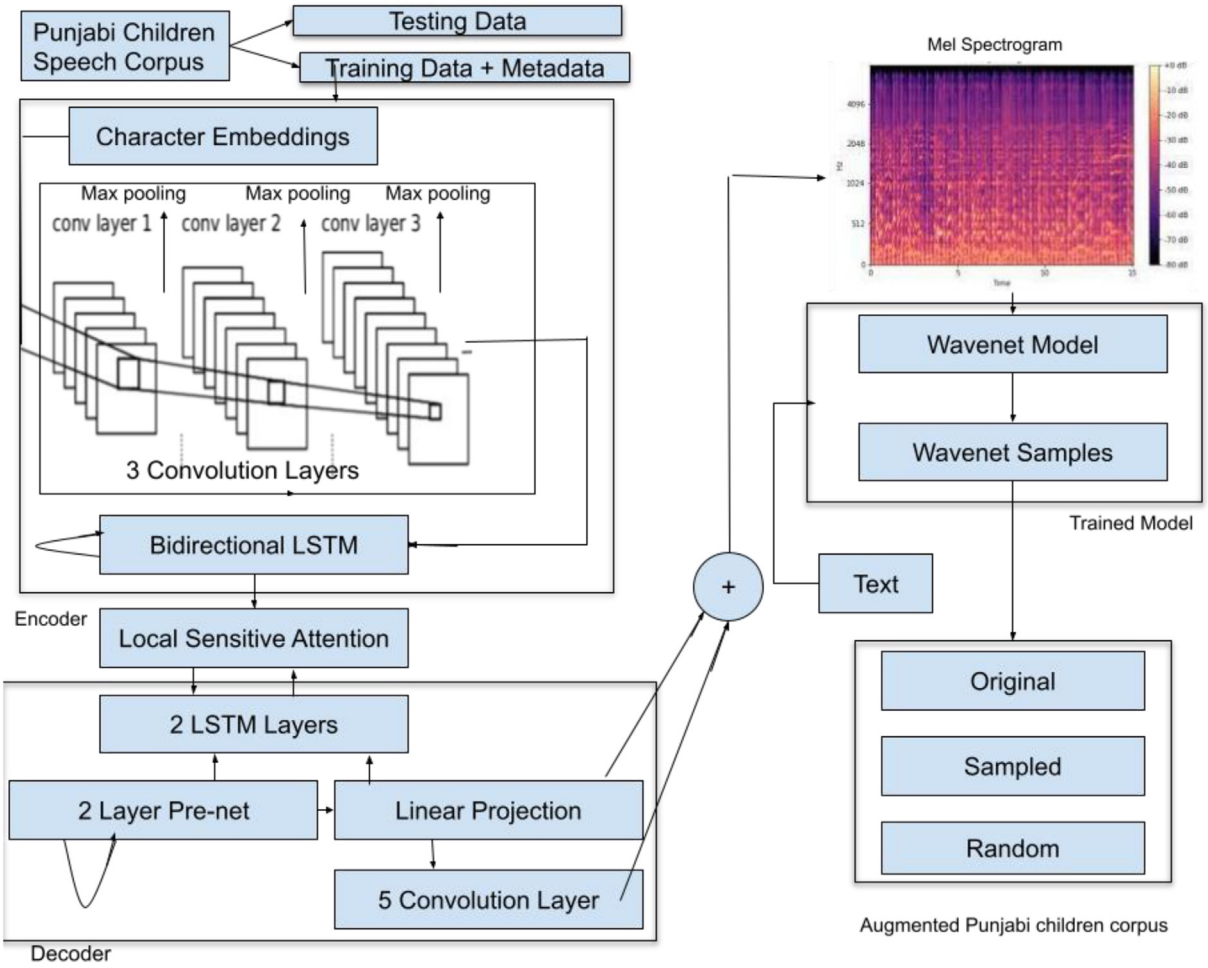

Fig. 3. Block diagram represents generation of TTS through external data augmentation approach.

among the speakers of the training population, thus, independent of the information content of the speech signal. These two variation sources are decoupled that helps in making the model suitable for unseen speakers as well.

4.6. Lexicon

Statistical modeling requires a sufficient number of examples to get a good estimate of the relationship between speech input and the parts of words. Pronunciation lexicon models the sequence of phones of a word. Phone is a basic sub-word unit that makes up a word.

Pronunciation model uses Markov chains. The HMM model aligns phones with the observed audio frames using self-looping. This provides flexibility in handling time-variance in pronunciation. The pronunciation dictionary is written by human experts. The pronunciation of words is typically stored in a lexical tree, a data structure that allows us to share histories between words in the lexicon. Phones are not homogeneous. The amplitudes of frequencies change from the start to the end. Also, variations in gender, pitch, accent, age etc. bring a change in the way a person utters. Therefore, this model gives the likelihood of the words to the decoder but cannot detect the exact word. In our case, we have collected Punjabi speech data from the children of age group 6–13 years with a mix of female and male children. The training data is expanded using data augmentation so that all phonemes of the language are covered. The speech data is digitally recorded with 16 KHz sampling frequency. To enhance lexical diversity, we have supplied additional words in our pronunciation dictionary and compared its performance.

4.7. Language model and decoder

There is an important role of the language model in speech recognition. It assigns a probability estimate to word sequences and defines what the speaker may say, the vocabulary, the probability over possible sequences, by training on some texts. The trigram model is used in our current system. The idea behind the trigram model is to truncate the word history to the last 3 words, and therefore approximate the history of the word. Decoder receives all the outputs from the acoustic, lexicon and language model. It then uses these outputs to recognize the word spoken.

5. Experimental setup & result analysis

Kaldi toolkit is used to implement the modeling classifiers. Table 1 describes the details of the dataset used. The adult data set consists of isolated words, continuous, and long-contextual sentence types spoken by 42 speakers of age groups 17–28 years. The child data set is composed of continuous sentence types spoken by 66 speakers of age groups 5–13 years in clean environment. Both data set have been recorded in same recording *.wav* format. The total data duration of adult data set recording is of 14 h and 36 min, whereas duration of child data set recording is of 11 h

Table 1
Database Description.

Specification	Adult Speech Corpora	Children Speech Corpora
No. of speakers	42	66
Recording Parameters	.wav file using mono-channel	.wav file using mono-channel
Recording Environment	Studio, Microphone	Dictaphone, Microphone with both Open and Closed Environment
Sentences Types	Isolated Words, Continuous, and Long-Contextual	Continuous
Age group	17–28 years	6–13 years
Total Hours	14 h and 36 min	11 h and 54 min
Gender	19 male and 23 females	32 male and 34 females

and 54 min. It has already been estimated by some of the earlier proposed research works that using speech data size between 10 and 40 h to train a baseline Tacotron produces good synthesis [44]. The proposed system is using K-fold cross validation technique. The children speech corpus is divided into train and test with 80% in training i.e Train_Child, and rest 20% in testing i.e., Test_Child. Formant modification is applied on the adult speech Corpus i.e. Original_Adult and the resultant dataset is given the alias as Formanted_Adult.

The system performance is calculated by using Word Error Rate (WER) metric that uses the concept of "Percentage Correct (PC)" and "Percentage Accuracy (PA)". Percentage correction (PC) gives word correction rate and Percentage Accuracy (PA) gives word accuracy rate. Eqs. (10)–(12) define these metrics where, $N = numberofwordsintestset$, $D = numberofdeletions$, $S = numberofsubstitutions$, $I = numberofinsertions$.

$$PC = (N - D - S)/N * 100 \tag{10}$$

$$PA = (N - D - S - I)/N * 100 \tag{11}$$

$$WordErrorRate(WER) = 100\% - Percentageaccuracy \tag{12}$$

In the proposed approach, a combination of feature vectors is created before classification. To deal with inter-speaker variability, the features are processed in multiple phases. In the first phase, mono-phone (mono) models are produced for the corresponding training samples. In the second phase, tri-phone models are used for the computation of delta features (tri1) and delta delta features(tri2) which helps in the production of 13-dimensional features across 4 frames. As a result, 117 dimensional vectors are generated. Linear discriminant analysis (LDA) [45] and Maximum likelihood linear transformation (MLLT) [46] estimation (tri3) is applied to reduce the dimensions from 117 to 30. To normalize inter speaker variability, global fMLLR [47] (Feature space Maximum Likelihood Linear Regression) is used so that reduced dimensions are aligned.

5.1. Performance analysis of baseline system on varying front-end and modeling approaches

The baseline system uses the original adult speech corpus along with the training part of the child speech corpus as training data and TDNN based acoustic model. The comparative analysis of DNN-based acoustic models and TDNN-based acoustic models is demonstrated in Table 2. In addition to this, the effect of mismatched conditions is tested with different combinations of adult speech and children speech.

It can be observed from the results that, when we train our ASR system on adult speech and test it on children speech, the ASR system performs poorly. Using children's data for training improves the system performance by a huge margin. When we combine adult as well as children's data for training, we saw an increase in accuracy of the system. This is mainly attributed to the availability of a good amount of training data since using only children's data for training is not sufficient and causes overfitting, thus, degrading the performance. Further, TDNN based acoustic models prove to be better on every combination of training data than the DNN based acoustic model. A relative gain of approximately 32% on using TDNN instead of DNN while training on Original_Adult and Train_Child.

5.2. Performance analysis using data augmentation

In the proposed work, augmented data is added to the adults' and children's data for improving performance. This is done for controlling the speaker diversity in the synthesized data. The proposed system performs sampling over Tacotron_Child to generate original, random and sampled datasets, where Tacotron_Child is the dataset of synthesized utterances that is received from Tacotron 2 model. In sampled data augmentation case, d-vector is generated from some utterance seen during training but the source utterance and synthesized one have some changes in speaker characteristics. These changes in the acoustic characteristics along with increase in dataset leads to better performance.

Table 3 describes the performance of our ASR system after data augmentation is applied. It is evident from the results that data generated using sampling technique gives the better results. WER is reduced to 6.94% from 7.98% using data augmentation through sampled data. This gives a relative improvement of 13% compared to our baseline system.

5.3. Performance analysis using formant modification

Formant modification has been performed to further enhance the baseline system's performance, Original_Adult and Train_Child corpus in training process is considered as the baseline system. Formant modification is done on adult speech to mitigate it's difference with children speech.

As shown in Table 4, we experimented with different values of warping factor in formant modification. The application of formant modification improves the efficiency of the ASR model as WER is

Table 2
Performance Analysis of Baseline System.

Training Type	Testing TYPE	WER (%) using DNN	WER (%) using TDNN
Train_Child	Test_Child	12.73	9.18
Original_Adult		38.75	36.21
Original_Adult + Train_Child		**11.76**	**7.98**

Table 3
Performance analysis using data augmentation.

Dataset	Augmentation	WER (%)
Original_Adult + Train_Child + Tacotron_Child	Original	7.02
	Random	7.56
	Sampled	6.94

Table 4
Performance analysis using formant modification on MFCC front end approach.

Formant Modification Dataset	WER (%)
Original_Adult + Train_Child	7.98
Original_Adult + Formanted_Adult F1 (0.15) + Train_Child	6.78
Original_Adult + Formanted_Adult F2 (0.20) + Train_Child	6.67
Original_Adult + Formanted_Adult F3 (0.25) + Train_Child	6.72

reduced from 7.98% to 6.67%. We tweaked the warping factor (in the range between −1 to 1) out of which 0.20 outperforms the other values. It results in lowest WER i.e., 6.67%. Hence, we have used 0.20 as the value of the warping factor for further experimentation. A relative gain of 16.4% is observed on using formant modified adult speech corpus along with the Original_Adult and Train_Child dataset.

5.4. Performance analysis using formant modification with data augmentation

In order to further validate the effectiveness of formant modification, we combine it with data augmentation. These two techniques tackle the issue of data scarcity. Table 5 shows the results after applying augmentation as well as formant modification. We are using the three augmented data (original, random and sampled) along with formant modified adult data with a warping factor of 0.2 to test the performance of the system.

It can be seen that using the sampled dataset is the most suited. WER dropped to 6.31% using formant modification as well as sampled augmented dataset. Keeping Original_Adult, Train_Child and Sampled data fixed, we observed a relative gain of 9% by adding formant modified adult data in training our system.

To further analyse the system performance, the augmented audios obtained on original signals are combined which are processed with different combination of front end feature vectors approaches. The role of these features are to produce robust feature vector, it is only possible by analysing the efficiency of each individual features which are later combined with MFCC feature vectors. The final WER obtained on pooled dataset is as shown in Fig. 4 where MFCC combined FDLP features performed better in comparison to that of earlier MFCC based system only.

Here, we are expanding the system's vocabulary by introducing lexically diverse words i.e. we supplied additional words in our pronunciation dictionary and compared its performance.

This empowers the pronunciation model to match phones to words. Table 6 shows different result scenarios. It can be observed that after adding 10 k extra words, WER is at its lowest. Introducing lexical diversity made our model perform better with WER dropping to 5.87% from 5.63%. Thus, a relative improvement of 4% is noticed. Overall in comparison to initial baseline system which achieved a WER of 7.98% which is later improved by pooling original and formanted augmented dataset using hybrid front end features (FDLP + MFCC). It achieved a final WER of 5.63% with a relative improvement of 29.44% respectively.

5.5. Discussion

Initially, the performance of the proposed continuous ASR system for Punjabi language has been analysed by varying front-end and modeling approaches, and it can be observed from the results that TDNN based acoustic models outperforms DNN with relative gain of approximately 32% while training on Original_Adult and

Table 5
Performance analysis using formant modification with data augmentation.

Augmentation + Formant Modification using MFCC approach	WER (%)
Original_Adult + Train_Child + Sampled	6.94
Original_Adult + Formanted_Adult F2 (0.20) + Train_Child + Original	6.52
Original_Adult + Formanted_Adult F2 (0.20) + Train_Child + Random	6.79
Original_Adult + Formanted_Adult F2 (0.20) + Train_Child + Sampled	6.31

Table 6
Performance analysis with lexically diverse words.

No. of words added	WER (%) Using Adult + Child + Adult Formanted F2 (0.20) + Sampled
0	**5.87**
5 k	5.71
10 k	5.63
20 k	5.79

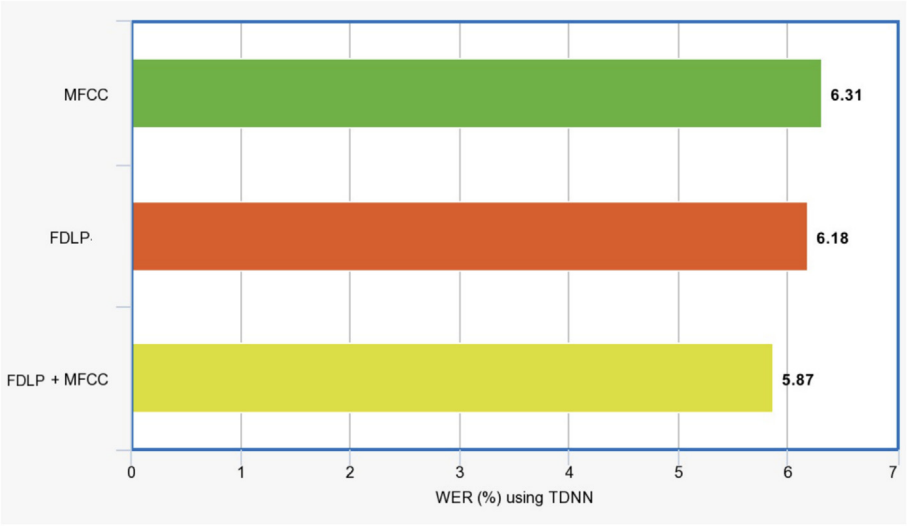

Fig. 4. Performance analysis using formant modification on hybrid front end approaches.

Table 7
Analysis and comparison of proposed approach with existing works.

Approach	Feature extraction	Acoustic modeling	Data-set	Performance rate (WER)	Remarks
Kadyan et al. (2021) [48]	MFCC	DNN	Punjabi Child Speech Corpus	R.I: 50.10%	Children speech corpus is augmented using Prosody and tactron2 augmentation techniques
Kaur and Kadyan (2020) [49]	MFCC	boosted Maximum Mutual Information	Punjabi Child Speech Corpus	R.I. 22–26%	A small corpus has been framed for Punjabi children's speech. System has been processed using discriminative approaches
Kadyan et al. (2021) [50]	MF-GFCC + pitch + VTLN	DNN-HMM, GMM-HMM	Punjabi Child Speech Corpus	RI: 20.59% on noisy and 19.39% on clean environment	Sentence level medium size Punjabi children ASR system. Testing has been performed in clean and noisy conditions
Bawa and Kadyan (2021) [51]	GFCC + Pitch + VTLN	DNN-HMM	Punjabi Adult Speech Corpus, Punjabi Children Speech Corpus	R.I. 30.94%	Gender based selection with medium size Sentence level Punjabi children ASR system is employed using mismatched and varying environment conditions.
Proposed system	MFCC + FDLP	TDNN	Punjabi Adult Speech Corpus, Punjabi Children Speech Corpus	R.Is 29.44%	Punjabi ASR system for children speech by using data augmentation and formant modification.

Train_Child. Secondly, the performance of this ASR system is improved by application of data augmentation and formant modification, separately. It has been observed that data augmentation reduces WER to 6.94% from 7.98% giving a relative improvement of 13% compared to our baseline system, and formant modification results in lowest WER i.e., 6.67%. and relative gain of 16.4%. Then, analysis has been carried out using Formant modification with Data Augmentation, which resulted in WER dropped to 6.31%. and relative gain further increased to 9%. Finally, this improved system has been tested with MFCC + FDLP hybrid feature set and expanded the system's vocabulary, which helped in achieving a final WER of 5.63% with a relative improvement of 29.44% respectively.

5.6. Discussion and comparative analysis with earlier proposed techniques

Most of the research works in ASR have been around high resource languages like English, Spanish, Mandarin etc., because of technological advancements in these language speaking regions. Hence, ample amount of speech data is available for studies in these languages. However, such state of the art dataset is not available for Indian languages such as Hindi, Punjabi, Dogri etc. Hence, data scarcity remains a big challenge in developing state-of-the-art ASR systems for these languages. Table 7 gives the comparative analysis of the proposed work with some existing state of the art ASR systems implemented for Punjabi language. The research woks proposed in [48-50] use Punjabi child speech corpus, and the work proposed in [54] uses both Punjabi child and adult speech corpus. It can be clearly observed from the given comparison that the proposed work of this paper i.e. combining data augmentation with formant modification and hybrid MFCC-FDLP-M features, outperforms existing works.

6. Conclusion

A novel approach to improve the performance of ASR systems for children targeting low resource languages has been proposed. Data augmentation has been done using Tacotron 2 to tackle the issue of data scarcity. Sampled data augmentation on children speech corpus has proved to give best results. Formant modification is applied on adult speech corpus to mitigate the acoustic and linguistic variabilities. The combined dataset includes training part of children speech corpus, sampled augmented children corpus, original adult speech corpus as well as formant adult speech corpus. Feature extraction is done using multiple front end approaches: MFCC, FDLP-S, FDLP-M and later best output of these approaches are combined as MFCC + FDLP-M to generate robust front end features. We have used TDNN as the acoustic model generating mono-phones, trigram language model along with supplying 10,000 words to the pronunciation model to make it lexically diverse. Speaker Adaptive Training is also done to make the model more robust for unseen speakers. The proposed hybrid front end feature based children ASR system gives WER of 5.63% when tested on testing part of only children speech corpus whereas, using training part of children speech corpus with adult speech corpus gives WER of 7.98% using TDNN acoustic model. Overall, we achieved relative gain of 29.44% from the baseline model using our proposed approach. Further work can be extended by employing spectrogram augmentation and deep conversion method to artificially enhance training data and accordingly increases system efficiency.

CRediT authorship contribution statement

Mohit Dua: Supervision, Software, Validation, Investigation. **Virender Kadyan:** Writing – original draft, Visualization. **Neha Banthia:** Conceptualization, Validation. **Akshit Bansal:** Methodology, Software. **Tanya Agarwal:** Data curation, Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Lopatovska I, Rink K, Knight I, Raines K, Cosenza K, Williams H, et al. Talk to me: Exploring user interactions with the Amazon Alexa. J Librarianship Inf Sci 2019;51(4):984–97.
[2] Sharma AS, Bhalley R. ASR—A real-time speech recognition on portable devices. In *2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Fall)*: IEEE; 2016. pp. 1-4.
[3] Janssen CP, Donker SF, Brumby DP, Kun AL. History and future of human-automation interaction. Int J Hum Comput Stud 2019;131:99–107.
[4] Sheridan TB, Parasuraman R. Human-automation interaction. Rev. Human Factors Ergon. 2015;vol. 1:41.
[5] Bachate RP, Sharma A. Automatic Speech Recognition Systems for Regional Languages in India. Int J Recent Technol Eng 585-592.

[6] Moore RK. A comparison of the data requirements of automatic speech recognition systems and human listeners. Eighth European Conference on Speech Communication and Technology, 2003.
[7] Antoniou A, Storkey A, Edwards H. Data augmentation generative adversarial networks; 2017. *arXiv preprint arXiv:1711.04340*.
[8] Kathania H, Singh M, Grósz T, Kurimo M. Data augmentation using prosody and false starts to recognize non-native children's speech; 2020. *arXiv preprint arXiv:2008.12914*.
[9] Park DS, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk ED, Le QV. Specaugment: A simple data augmentation method for automatic speech recognition; 2019. *arXiv preprint arXiv:1904.08779*.
[10] Jaitly N, Hinton GE. Vocal tract length perturbation (VTLP) improves speech recognition. In Proc. ICML Workshop on Deep Learning for Audio, Speech and Language (vol. 117); 2013.
[11] Cui X, Goel V, Kingsbury B. Data augmentation for deep neural network acoustic modeling. IEEE/ACM Trans Audio Speech Lang Process 2015;23 (9):1469–77.
[12] Ittichaichareon C, Suksri S, Yingthawornsuk T. Speech recognition using MFCC. In: International conference on computer graphics, simulation and modeling. p. 135–8.
[13] Hermansky H, Tsuga K, Makino S, Wakita H. Perceptually based processing in automatic speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE; 1986. p. 1971–4.
[14] Athineos M, Ellis DP. Frequency-domain linear prediction for temporal features. In 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721) (pp. 261-266). IEEE; 2003.
[15] Zhang Z, Geiger J, Pohjalainen J, Mousa A-D, Jin W, Schuller B. Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Trans Intell Syst Technol 2018;9(5):1–28.
[16] Guglani J, Mishra AN. Continuous Punjabi speech recognition model based on Kaldi ASR toolkit. Int J Speech Technol 2018;21(2):211–6.
[17] Kathania HK, Kadiri SR, Alku P, Kurimo M. Study of Formant Modification for Children ASR. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7429-7433). IEEE; 2020.
[18] Sunil Y, Prasanna SRM, Sinha R. Children's Speech Recognition Under Mismatched Condition: A Review. IETE J Educ 2016;57(2):96–108.
[19] Huber JE, Stathopoulos ET, Curione GM, Ash TA, Johnson K. Formants of children, women, and men: The effects of vocal intensity variation. J Acoust Soc Am 1999;106(3):1532–42.
[20] Dua M, Aggarwal RK, Biswas M. GFCC based discriminatively trained noise robust continuous ASR system for Hindi language. J Ambient Intell Hum Comput 2019;10(6):2301–14.
[21] Adda G, Stüker S, Adda-Decker M, Ambouroue O, Besacier L, Blachon D, et al. Breaking the unwritten language barrier: The BULB project. Proc Comput Sci 2016;81:8–14.
[22] Potamianos A, Narayanan S, Lee S. Automatic speech recognition for children. Fifth European Conference on Speech Communication and Technology, 1997.
[23] Shahnawazuddin S, Adiga N, Kathania HK. Effect of prosody modification on children's ASR. IEEE Signal Process Lett 2017;24(11):1749–53.
[24] O'Shaughnessy D. Linear predictive coding. IEEE Potentials 1988;7(1):29–32.
[25] Cooper EL. Text-to-speech synthesis using found data for low-resource languages (Doctoral dissertation). Columbia University; 2019.
[26] Oord AVD, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kavukcuoglu K. Wavenet: A generative model for raw audio; 2016. arXiv preprint arXiv:1609.03499.
[27] Li J, Gadde R, Ginsburg B, Lavrukhin V. Training neural speech recognition systems with synthetic speech augmentation; 2018. arXiv preprint arXiv:1811.00707.
[28] Ragni A, Knill KM, Rath SP, Gales MJ. Data augmentation for low resource languages. In: INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association. p. 810–4.
[29] Rosenberg A, Zhang Y, Ramabhadran B, Jia Y, Moreno P, Wu Y, et al. Speech recognition with augmented synthesized speech. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE; 2019. p. 996–1002.
[30] Gerosa M, Giuliani D, Brugnara F. Acoustic variability and automatic recognition of children's speech. Speech Commun 2007;49(10-11):847–60.
[31] Kadyan V, Mantri A, Aggarwal RK, Singh A. A comparative study of deep neural network based Punjabi-ASR system. Int J Speech Technol 2019;22(1):111–9.
[32] Kadyan V, Shanawazuddin S, Singh A. Developing children's speech recognition system for low resource Punjabi language. Appl Acoust 2021;178:108002. https://doi.org/10.1016/j.apacoust.2021.108002.
[33] Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 4779–83.
[34] Jia Y, Zhang Y, Weiss RJ, Wang Q, Shen J, Ren F, Wu Y. Transfer learning from speaker verification to multispeaker text-to-speech synthesis; 2018. arXiv preprint arXiv:1806.04558.
[35] Deng Y, He L, Soong F. Modeling multi-speaker latent space to improve neural tts: Quick enrolling new speaker and enhancing premium voice; 2018. arXiv preprint arXiv:1812.05253.
[36] Herre J, Johnston JD. Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS). In *Audio Engineering Society Convention 101*. Audio Engineering Society; 1996.
[37] Thomas S, Ganapathy S, Hermansky H. Phoneme recognition using spectral envelope and modulation frequency features. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2009. p. 4453–6.
[38] Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2015. p. 5206–10.
[39] Russell M. The pf-star british english childrens speech corpus. The Speech Ark Limited; 2006.
[40] Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, Jaitly N, Saurous RA. Tacotron: Towards end-to-end speech synthesis; 2017. arXiv preprint arXiv:1703.10135.
[41] Park H, Lee D, Lim M, Kang Y, Oh J, Kim JH. A Fast-Converged Acoustic Modeling for Korean Speech Recognition: A Preliminary Study on Time Delay Neural Network; 2018. arXiv preprint arXiv:1807.05855.
[42] Waibel A, Hanazawa T, Hinton G, Shikano K, Lang KJ. Phoneme recognition using time-delay neural networks. IEEE Trans Acoust Speech Signal Process 1989;37(3):328–39.
[43] Anastasakos T, McDonough J, Schwartz R, Makhoul J. A compact model for speaker-adaptive training. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96 (vol. 2, pp. 1137-1140). IEEE; 1996.
[44] Chung YA, Wang Y, Hsu WN, Zhang Y, Skerry-Ryan RJ. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2019. p. 6940–4.
[45] Haeb-Umbach, R., & Ney, H. (1992, March). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In Proc. ICASSP (Vol. 1, pp. 13-16). USA: ICASSP.
[46] Gales MJF. Maximum likelihood linear transformations for HMM-based speech recognition. Comput Speech Lang 1998;12(2):75–98.
[47] Parthasarathi SHK, Hoffmeister B, Matsoukas S, Mandal A, Strom N, Garimella S. fMLLR based featurespace speaker adaptation of DNN acoustic models. In Sixteenth annual conference of the international speech communication association, 2015.
[48] Kadyan V, Kathania H, Govil P, Kurimo M. Synthesis Speech Based Data Augmentation for Low Resource Children ASR. In: International Conference on Speech and Computer. Cham: Springer; 2021. p. 317–26.
[49] Kaur H, Kadyan V. April). Feature Space Discriminatively Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit. Proceedings of the International Conference on Innovative Computing & Communications (ICICC), 2020.
[50] Kadyan V, Bawa P, Hasija T. In domain training data augmentation on noise robust Punjabi Children speech recognition. J Ambient Intell Hum Comput 2021:1–17.
[51] Bawa P, Kadyan V. Noise robust in-domain children speech enhancement for automatic Punjabi recognition system under mismatched conditions. Appl Acoust 2021;175:107810. https://doi.org/10.1016/j.apacoust.2020.107810.