Out Domain Data Augmentation on Punjabi Children Speech Recognition using Tacotron

Taniya Hasija¹, Virender Kadyan² and Kalpna Guleria³*

¹Centre of Excellence for Speech and Multimodal Laboratory, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
²Department of Informatics, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India
³Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India.

E-mail: kalpna@chitkara.edu.in

Abstract. The performance of Automatic Speech Recognition (ASR) is directly proportional to the quality of the corpus used and the training data quantity. Data scarcity and more children’s speech variability degrades the performance of ASR systems. As Punjabi is a tonal language and low resource language, less data is available for Punjabi children’s speech. It leads to poor ASR performance for Punjabi children speech recognition. To overcome limited data conditions, in this paper, two corpora of different domains are evaluated for testing the feasibility of ASR performance. We have implemented Tacotron as an artificial speech synthesis system for Punjabi Language. The speech audios synthesized by Tacotron are merged with available speech corpus and tested on Punjabi children ASR using Mel Frequency Cepstral Coefficients (MFCC) + pitch feature extraction, and Deep Neural Network (DNN) acoustic modeling. It is noticed that the merged data corpus has shown reduced Word Error Rate (WER) of the ASR system with a Relative Improvement (RI) of 9-12%.

Keywords: Automatic Speech Recognition (ASR), Data Augmentation, Tacotron, Punjabi Children ASR, Text to Speech

1. Introduction

Speech is the most powerful and important source of communication among people. ASR is the widest field in human-machine interaction with the application area of a spoken language translator, speech to text, or dictation processes, YouTube kids [1]. The main objective of the ASR system is to understand speech as clearly as a human can. Some factors are associated with the speech which makes the ASR difficult and results in poor accuracy [2]. These factors are the expression of the speaker, noise from surroundings, continuous speaking which may cause skipping of some words, and the computer is not able to understand the continuity of the language. Speaker and channel variability is another cause of the decreased performance of ASR [3]. Despite adult speech, children's speech has more variability due to high pitch and improper pronunciation of words, which increase the complexity of the ASR system [4]. Researchers have worked upon removing variabilities of the speech signal and in return performance of the system has been increased. ASR systems are developed for international languages and have great success and applications, but the obstacle is, ASR systems have not developed for native languages. In India, there are 22 national
languages, from which only Hindi, English, and Marathi ASR systems are available and less work has been done on Punjabi speech [5]. All reported work in Punjabi speech is for adult ASR and children’s speech recognition is still a challenging task due to limited resources. Also, Children’s speech has more variability factors [3][4]. It is very difficult to obtain a phonetically balanced corpus in case of children. A quality corpus must have these attributes: consistency of corpus, phonetically rich, less variability, and quality of recording [6]. Temporal and spectral features vary in different age groups. The variability in children’s speech is due to less control on the articulatory or vocal tract geometry [7][8]. Our motivation behind this work is to develop the ASR system for children Punjabi Speech.

The next challenge for developing the ASR system is the collection of ASR speech data, as Punjabi is the native language and the ASR system is not developed yet, therefore no online data available for Punjabi children's speech. A proficient method should be used for enhancing the data of Punjabi speech artificially through data augmentation [9][10]. When the parameter of speech corpus is transformed and new speech corpus is augmented with existing corpus then augmentation is termed as in domain augmentation [11], else when the new corpus is merged with different technique or method, it is termed as out domain data augmentation. Tacotron is a text to speech (TTS) system which is used for producing natural-sounding speech. Tacotron can be used for designing new sounds from the text which will help us in the ASR system [12][13]. Tacotron handles everything as a single model. Text to speech machine has a text front end that extracts linguistic features, also has an acoustic prediction model, duration model, and a vocoder based on signal processing. TTS machines have errors from each component and have complexity [14]. Tacotron provides rich conditioning of parameters or attributes on features. Prosody features can also be varied and pitch can be given to particular utterances using tacotron [15]. The performance of the Tacotron system is measured through the Mean Opinion Score (MOS). Since speech is generated at the frame level, Tacotron is faster than sample level regressive methods.

In this paper, experiments have been performed related to out domain data augmentation using Tacotron and MFCC+ pitch feature extraction on Punjabi child speech and it is verified that the presented results have better performance. Section 2 will cover the literature review and then proceed with a theoretical background of tacotron in section 3. Section 4 elaborates the system overview which illustrates a block diagram of the ASR system implemented by augmentation of original children’s speech and pre-synthesized speech. Section 5 presents the experimental setup and results. Finally, the conclusion is discussed in the last section.

2. Related Work

ASR systems that were developed in the beginning were focusing on speech recognition of international languages. Punjabi is observed as an unreserved language and the ASR system of the Punjabi language has been in the development state and developed for isolated and continuous adult speech. Initially, an isolated Punjabi ASR system was implemented by Dua et al. [5]. The authors have performed MFCC feature extraction and has utilized Hidden Markov Model (HMM) technique which has shown 94-95% accuracy. Thereafter, the ASR system has been implemented on spontaneous Punjabi speech corpus by Kumar and Singh [16]. The implementation was done in java programming language and on the Sphinx framework. The system is tested in four phases and it has shown accuracy of 90-93%. Kadyan et al. [17] have proposed two-hybrid classifiers which outperformed the result of HMM and the computational complexity was also reduced by these classifiers.

Guglani and Mishra [18] implemented the Kaldi toolkit for Punjabi corpus using MFCC, and PLP feature extraction methods. To include temporal evaluation of MFCC, additional feature delta, and delta-delta values were computed and all are combined to form a single modality. Authors have tested the system on many models and concluded that the triphone model has shown the best accuracy than other models. Kaur and Kadyan [19] have implemented the children Punjabi ASR system using the bMMI (boosted Maximum
Mutual Information) discriminative technique. The authors have shown 22 to 26% relative improvement in accuracy. In 2020, Bhardwaj et al. [20] have implemented Subspace Gaussian mixture model (SGMM) acoustic modelling technique on children Punjabi ASR system. The accuracy of the experimented system is 83%.

It is evident from the abovementioned literature that a lot of work has been done in adult Punjabi speech, however very few work is being done in children Punjabi ASR system. Further, it's worth mentioning that the data used by researchers in case of children speech is quantitatively less.

Focusing on the challenges of data collection, Tacotron is implemented for removing the scarcity of data. Tacotron consists of three stages, these stages are text analysis front end, an acoustic model, and an audio synthesis module. In 2017, Wang et al. [12] have implemented an end-to-end model that produces synthetic speech. They represented key techniques to make frameworks that work sequentially. Their tacotron system got a 3.82 scale mean opinion from 5 on US English. In 2018, Skerry-Ryan et al. [14] presented an extended version of Tacotron which has latent embedding space of prosody. The synthesized speech matches the prosody of the given text’s speech. In 2018, Shen et al. [21] presented a neural TTS system. The system was combined as a sequence to sequence recurrent networks that predict the Mel spectrogram of given text. The system achieved a 4.53 MOS. In 2019, Yasuda et al. [15] suggested Tacotron 2 which outperforms classical systems. Tacotron was extended with self-attention which captured pitch related dependencies, enhanced audio quality.

Precise extraction of speech recognition is another important task for speech recognition. A number of techniques have been used for extracting features from speech signals. Gupta and Gupta [22] have represented the comparative studies of Linear Predictive Coding (LPC), RelAtiveSpecTrAl (RASTA), and MFCC. They concluded that LPC is not working well as the words with the same vowel sounds were not recognized. Performance of RASTA is enhanced when RASTA features are combined with perceptually based Linear Predictive (PLP) features. It is observed that MFCC is performing well but not in a noisy environment. Hachker et al. [23] have done a comparison of MFCC and PLP on the Arabic alphabet speech set and concluded that MFCC is performing slightly well as compared to PLP. The comparison of LPC, MFCC, PLP, and RASTA-PLP is done by Dave [24]. The author concluded that PLP and MFCC parameters carry the nature of speech whereas, on the basis of previous features, LPC predicts future features. LPC will not be chosen as the human voice is filled with nonlinearity. MFCC has better results. After feature extraction, the ASR system is processed for acoustic modeling. Gerosa et al. [25] have presented their work to a better understanding of temporal and spectral variability of children’s speech of different age groups and HMM is used. Further vocal tract length normalization (VTLN) has been adopted. Hinton et al. [26] have presented the results of HMM, Gaussian Mixture Model (GMM), and Deep Neural Network (DNN) acoustic modeling techniques and concluded that DNN outperforms among all. In 2020, Taniya et al. [27] have represented their work on DNN trained children Punjabi speech ASR. In their work, a number of hidden layers and dimensions are tweaked in DNN acoustic modelling, and authors have concluded that DNN with four hidden layers and 1024 hidden units is outperforming other existing techniques.

In this research paper, we have implemented a Tacotron system to generate synthesis speech. Thereafter, the synthesized corpus and original corpus pooled together to capture the high accuracy of the Children Punjabi ASR system.

3. Tacotron
The Tacotron model takes sequences of characters and induces waveforms of that sequence of characters. The Tacotron model consists of an encoder, a decoder, a post-processing net, and waveform synthesis. During encoding, the CBHG module is implemented [12][14].
3.1. CBHG module

The CBHG module has three layers which include 1-D Convolution filter Bank, Highway network, and bidirectional Gated recurrent unit (GRU). The CBHG module takes input sequences that are convolved with K sets using 1-D convolution filter banks. The output of the convolution layer is fed to highway networks. Highway networks extract high-level features. Sequential features are extracted through a third layer i.e. GRU Recurrent Neural Networks (RNN), using backward and forward context. The idea of the CBHG module is inspired by work in machine translation [28].

3.2. Encoder

Encoder extracts feature representation of text with high robustness. Input text embedded into a continuous vector matrix. A nonlinear transformation called pre-net is applied to each embedding. After this step, the CBHG module is applied to the pre-net, which produces a robust representation of text.

3.3. Decoder

Output generated from GRU RNN works as an input to the decoder. Stack of GRUs is used for decoding which has residual connections set up vertically. Convergence is high when we use residual connections. A post-processing network is used to convert the sequence-to-sequence target to the waveform. Decoder predicts non-overlapping output frames at each step of decoding. If r frames are predicted, then decoding steps are divided by r, reducing training time and model size. The reason behind that is each character represents multiple frames and neighboring speech frames are correlated. During the first decoding step, frames are all-zero, represented as a <GO> frame. On step t, the output of the t step is the input of the t+1 decoding step. The input frame is passed to pre-net (nonlinear transformation) as done in encoding.

3.4. Post Processing Net and Waveform synthesis

Postprocessor converts sequence to sequence target to synthesis waveform by using the Griffin-Lim algorithm [29]. The output of the post-processing is a sequence that is fully decoded. Post-processing runs in forward and backward direction and corrects the prediction error from each individual frame. Post-processing predicts alternative targets.

4. Experimental Setup and Procedure

As Punjabi is a low resource language, there is less corpus available for the Punjabi language. 4 hours 20 minutes of data was manually collected for developing continuous Punjabi children speech corpus. The audio recording was done in schools with the help of a microphone at a 16 kHz sampling rate. The age group of children for recording data is 8-14 years. Experiments were performed on the Kaldi toolkit using the Ubuntu operating system [30]. There are 1885 utterances in a training data set from 39 speakers which has 18 female and 21 male speakers. The test set has 485 utterances spoken by 6 speakers, having 3 male and 3 female speakers. The language model has a 5k unique lexicon that has been given to the system. At first, MFCC features are extracted from the system. The system is then tested on monophones and triphones by using DNN+HMM classifiers. A total of 13 feature coefficients are extracted using the MFCC technique which is the energy parameters of each frame. The size of each frame is 25 ms and the frame-shift is 10 ms. Hamming window has been used for frame extraction and a 23 channel Mel filter-bank was employed. Thereafter, logarithm and DCT are applied and feature coefficients are computed. MFCC features describe the instantaneous and spectral envelope shape of the speech signal. Speech has also some dynamic features i.e. trajectories of MFCC features over time. Thereafter, these trajectories are calculated and merged with 13 coefficients of MFCC and concatenation is done to enhance the ASR performance. Trajectories features are known as delta features (tri 1) and are computed as:
\[ d_t = \frac{\sum_{n=1}^{N} n(c_t - c_{t-n})}{2 \sum_{n=1}^{N} n^2} \]

where \( d_t \) is delta coefficients, computed on frame \( t \). \( c_t \) and \( c_{t-n} \) are static coefficients and the value of \( N \) is 2. After the computation of the delta feature, the total features are 26. Delta–delta features (tri2) are also extracted, which are the trajectories of delta features (tri 2) and are computed from the same formula:

\[ dd_t = \frac{\sum_{n=1}^{N} n(d_t - d_{t-n})}{2 \sum_{n=1}^{N} n^2} \]

where \( dd_t \) are delta-delta coefficient. These features are also known as the first and second derivative of a signal. There is a need for a reduction of coefficients, as coefficients are multiplied after delta and double delta. To convert smaller amounts of acoustically distinct units, Linear Discriminative Analysis (LDA) is implemented on the output of tri 2, which reduces the coefficient into feasible 40 dimensions. After estimating the probability, it is decided that a new set of inputs belong to the new class and the class with the highest probability will be considered as the output class. Next Maximum Likelihood Linear Transformation (MLLT) is obtained which is estimated over utterances and exclusion of speaker-specific information was done. LDA+MLLT is the tri 3 modeling of the system. Finally, the DNN–HMM classification is done. 1024 dimensions are utilized by 4 hidden layers and on input features, tanh nonlinearity function has been applied. Learning rate is used in the range of 0.005 to 0.0005 and 20 epochs are initialized by DNN. 10 extra epochs are fixed for constant learning rate (0.0005). The system is evaluated for performance enhancement. Evaluation of the ASR performance is done using Word Error Rate (WER) term. The error can be the substitution of words (S) or deletion of words (D) or can be the insertion of new words (I) [31]. WER is computed as:

\[ \text{WER} \% = \frac{S+I+D}{N} \times 100 \]

WER is computed by the system after DNN classification. Now, out domain augmentation is applied to the Punjabi children corpus. Pre-synthesized speech is computed from Tacotron and augmented with children’s speech and given to the ASR system. Augmentation leads to a speech corpus which has 2032 utterances. Augmented data again fed to the ASR system as shown in fig. 1 and MFCC features are extracted. Monophone and triphone models are also computed on features and finally DNN is implemented.
The efficiency of the system will be compared on both outputs; system without augmented data and system with augmented data. Another parameter used for performance evaluation is Relative Improvement (RI). RI is the absolute increase corresponding to a new value (N) with respect to the old value (O):

$$RI \% = \frac{N - O}{O} \times 10$$

Here, old value represents the result of baseline corpus and new value is the result of system having augmented data, and evaluation is done on RI that how much improvement has taken place after augmentation.

5. Results and Discussions
Initially, only children’s data has been given to the system, which acts as baseline data. Corpus is fed to the system for feature extraction using the MFCC technique, thereafter delta and double delta features are computed and LDA+MLLT are applied on features, then output is fed to DNN-HMM model. Further, the tacotron system is implemented and audios generated from tacotron are merged with children’s corpus. Two techniques of feature extraction have been implemented which are MFCC and MFCC with Pitch features. Robustness of the system has been seen when pitch features are used along with MFCC features. Table 1 shows the baseline result of children’s speech corpus using MFCC and MFCC + Pitch feature extraction and DNN acoustic modeling.

| Modeling of features | MFCC  | MFCC + Pitch |
|----------------------|-------|--------------|
| Mono                 | 19.3  | 20           |
| Tri 1                | 27.75 | 26.91        |
| Tri 2                | 27.19 | 27           |
| Tri 3                | 18.04 | 16.64        |
| DNN                  | 14.91 | 14.21        |

In Table 1, the WER generated by the ASR system is 14.21% after tri 1, tri 2 and tri 3 modeling of features. These results are termed as baseline results and are compared with augmented system’s results. After the synthesis of artificial speech through tacotron, synthesized speech is augmented with original speech corpus, and results after augmentation is shown in Table 2. The output WER of the system with augmentation of data is 12.87%.
Table 2. Result of Punjabi Children ASR after augmentation

| Modeling of features | MFCC  | MFCC + Pitch |
|----------------------|-------|--------------|
| Mono                 | 17.76 | 18.68        |
| Tri 1                | 16.88 | 16.25        |
| Tri 2                | 17.02 | 15.62        |
| Tri 3                | 13.86 | 14.49        |
| DNN                  | 13.33 | 12.87        |

Augmentation has shown an enhancement in the accuracy of the system by decreasing the WER%. Thereafter, augmentation results are compared with the baseline results and it is evaluated that the new system has shown 10% RI. It has been concluded after implementation of both systems, that data augmentation using tacotron has shown improved results and data scarcity is also improved by synthesis speech.

6. Conclusion

This paper represents the children Punjabi ASR system with improved performance. Novelty of our presented work lies in implementation of children Punjabi ASR system with improved performance using out domain data augmentation whereas the earlier works focus on English language. In this work, data augmentation is implemented by producing pre-synthesized speech. Two types of the corpus that are clean and synthesized are merged, thereafter a corpus with greater quantity is implemented on the ASR system. After the implementation, the system has shown improved accuracy than the baseline. The Accuracy of the Punjabi Children ASR system is nearly 88% and RI is 9 to 12% after extending data using pre-synthesized speech. Further work can be extended using in domain and out domain data augmentation on corpus for evaluating the enhanced performance of the ASR system.

References

[1] Wang L, Ambikairajah E and Choi E HC 2007 Automatic language recognition with tonal and non-tonal language pre-classification 15th European Signal Processing Conference pp 2375-2379.
[2] Forsberg M 2003 Why is Speech Recognition Difficult? Technology pp 1–10.
[3] Hagen A, Pellom B and Cole R 2003 Children’s speech recognition with application to interactive books and tutors IEEE Work ASRU pp 186–191.
[4] Gurunath Shivakumar P and Georgiou P 2020 Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations Comput. Speech Lang. 63 pp 1–15.
[5] Dua M, Aggarwal R, Kaday V and Dua S 2012 Punjabi Automatic Speech Recognition Using HTK IJCSI Int. J. Comput. Sci. Issues 9 4 pp 359–364.
[6] Watts O, Yamagishi J, Berkling K and King S 2008 HMM-based synthesis of child speech Challenge.
[7] Potamianos A and Narayanan S Robust Recognition of Children’s Speech IEEE Trans. Speech Audio Process. 11 6 pp 603–616.
[8] Potamianos A, Narayanan S and Lee S 1997 Automatic Speech Recognition for Children Fifth Eur. Conf. Speech Commun. Technol. pp 1–4.
[9] Shahnawazuddin S, Adiga N, Kathania and Sai B T 2020 Creating speaker independent ASR system through prosody modification based data augmentation Pattern Recognit. Lett. 131 pp 213–218.
[10] Dahl G E, Yu D, Member S, Deng L and Acero A 2012 Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition 20 1 pp 30–42.
[11] Ratner A J, Ehrenberg H R, Hussain Z, Dunnmon J and Ré C 2017 Learning to compose domain-specific transformations for data augmentation Adv. Neural Inf. Process. Syst. 32:3237–3247.

[12] Wang Y, Skerry-Ryan R J, Stanton D, Wu Y, Weiss R J and Jaitly N 2017 Tacotron: Towards end-to-end speech synthesis arXiv (preprint arXiv:1703.10135).

[13] Rosenberg A, Zhang Y, Ramabhadran B, Jia Y, Moreno P, Wu Y and Wu Z 2019 Speech Recognition with Augmented Synthesized Speech IEEE ASRU (Singapore) pp 996-1002.

[14] Skerry-Ryan R J, Battenberg E, Xiao Y, Wang Y, Stanton D, Shor J, Weiss R J, Clark R and Saurous R A 2018 Towards end-to-end prosody transfer for expressive speech synthesis with tacotron 35th Int. Conf. Mach. Learn. ICML 11 pp 7471–7480.

[15] Yasuda Y, Wang X, Takaki S and Yamagishi J 2019 Investigation of Enhanced Tacotron Text-to-speech Synthesis Systems with Self-attention for Pitch Accent Language ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc pp 6905–6909.

[16] Kumar Y and Singh N 2017 An automatic speech recognition system for spontaneous Punjabi speech corpus Int. J. Speech Technol. 20 2 pp 297–303.

[17] Kadyan V, Mantri A and Aggarwal R K 2018 Refinement of HMM Model Parameters for Punjabi Automatic Speech Recognition (PASR) System IETE J. Res. 64 5 pp 673–688.

[18] Guglani J and Mishra A N 2018 Continuous Punjabi speech recognition model based on Kaldi ASR toolkit Int. J. Speech Technol. 21 2 pp 211–216.

[19] Kaur H and Kadyan V 2020 Feature Space Discriminatively Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit SSRN Electron. J. pp 1–5.

[20] Bhardwaj V, Bala S and Kadyan V 2020 Development of Robust Automatic Speech Recognition System for Children's using Kaldi Toolkit Second International Conference on Inventive Research in Computing Applications (ICIRCA) pp 10-13.

[21] Shen J, Pang R, Weiss R J, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerry-Ryan RJ, Saurous R A, Agiomvrgiannakis Y and Wu Y 2018 Natural TTS synthesis by conditioning WaveNet on MEL Spectrogram Predictions ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc pp 4779–4783.

[22] Gupta K and Gupta D 2016 An analysis on LPC, RASTA and MFCC techniques in Automatic Speech Recognition system Proc. 2016 6th Int. Conf. - Cloud Syst. Big Data Eng. Conflu. pp 493–497.

[23] Hachkar Z, Mounir B, Farchi A and Abbadi J El 2011 Comparison of MFCC and PLP Parameterization in pattern recognition of Arabic Alphabet Speech Can. J. Artif. Intell. Mach. Learn. Pattern Recognit. 2 3 pp 56–60.

[24] Dave N 2013 Feature Extraction Methods LPC , PLP and MFCC In Speech Recognition Int. J. Adv. Res. Eng. Technol. 1 pp 1–5.

[25] Gerosa M, Giuliani D and Brugnara F 2007 Acoustic variability and automatic recognition of children’s speech Speech Commun. 49(10–11) pp 847–860.

[26] Hinton G, Deng L, Yu D, Dahl G E, Mohamed A R, Jaitly N et al 2012 Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups IEEE Signal processing magazine 29(6) pp 82-97.

[27] Taniya, Bhardwaj V and Kadyan V 2020 Deep Neural Network Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit IEEE 5th International Conference on Computing Communication and Automation (ICCCA) pp 374-378.

[28] Lee J, Cho K and Hofmann T 2017 Fully Character-Level Neural Machine Translation without Explicit Segmentation Trans. Assoc. Comput. Linguist. 5 pp 365–378.

[29] Griffin D and Lim J 1984 Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2) pp 236-243.

[30] Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Silovsky J, Stemmer G and Vesely K 2011 The Kaldi speech recognition toolkit Workshop on automatic speech recognition and understanding IEEE Signal Processing Society.

[31] Morris A C, Maier V and Green P 2004 From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition Eighth International Conference on Spoken Language Processing J. Manuf. Processes A 56 75-86.