IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 35, No. 3, June, 2021, pp. 235-242
Journal homepage: http://iieta.org/journals/ria

Usage of Prosody Modification and Acoustic Adaptation for Robust Automatic Speech Recognition (ASR) System

Vivek Bhardwaj[1], Vinay Kukreja[1*], Amitoj Singh[2]

[1] Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab 140401, India
[2] School of Sciences and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala, Punjab 147001, India

Corresponding Author Email: vinay.kukreja@chitkara.edu.in

https://doi.org/10.18280/ria.350307

Received: 11 March 2021
Accepted: 13 June 2021

Keywords:
automatic speech recognition, prosody, pitch, duration, acoustic

ABSTRACT

Most of the automatic speech recognition (ASR) systems are trained using adult speech due to the less availability of the children's speech dataset. The speech recognition rate of such systems is very less when tested using the children's speech, due to the presence of the inter-speaker acoustic variabilities between the adults and children's speech. These inter-speaker acoustic variabilities are mainly because of the higher pitch and lower speaking rate of the children. Thus, the main objective of the research work is to increase the speech recognition rate of the Punjabi-ASR system by reducing these inter-speaker acoustic variabilities with the help of prosody modification and speaker adaptive training. The pitch period and duration (speaking rate) of the speech signal can be altered with prosody modification without influencing the naturalness, message of the signal and helps to overcome the acoustic variations present in the adult's and children's speech. The developed Punjabi-ASR system is trained with the help of adult speech and prosody-modified adult speech. This prosody modified speech overcomes the massive need for children's speech for training the ASR system and improves the recognition rate. Results show that prosody modification and speaker adaptive training helps to minimize the word error rate (WER) of the Punjabi-ASR system to 8.79% when tested using children's speech.

1. INTRODUCTION

The process of recognition and translation of the natural language utterances into the text form is called Automatic speech recognition. From the studies, it was found that ASR has a number of modern technologies for recognizing adult speech with higher accuracy, whereas the field of children's speech recognition is straggling behind with poor recognition due to the variabilities in the children's speech [1-3]. Along with this a very limited amount of work is done on children's ASR system especially for the Indian regional language Punjabi as compared to the adult ASR. There are several applications of ASR for children in various fields like education, entertainment, and communication. To improve the recognition rate of the children ASR system, one way is to use a sizable amount of children's speech corpora for training the ASR system. As the speech recognition systems built with deep neural networks (DNN) are data-driven, so with the large amount of speech data recognition rate of the ASR is better. But the issue is the availability of a sufficient amount of Punjabi children's speech corpus for the ASR system and it's also difficult to collect the speech data for children as compared to adults. One another method to improve the recognition rate and performance of the Children ASR system is by minimizing the acoustics mismatches of the children and adult speech with the help of prosody. Acoustic mismatch in ASR means training the system with adult speech corpus, testing with the children's speech corpus, and vice versa. This acoustic mismatch is due to the shorter vocal track of the children as compared to the adults. From the literature, it was also found that the pitch of the children is quite different and higher than the adult's speech. This is one of the factors that make children's speech different from adult speech and causes acoustic mismatch [4, 5]. The range of the pitch frequency mainly lies between 70 Hz to 255 Hz for the adult speakers whereas for children's pitch frequency ranges usually from 200 Hz to 350 Hz [4-6]. The second factor that decreases the recognition rate is the speaking rate of the adult and child speakers. The phoneme duration of the children's speakers is longer as compared to the adults [4]. Thus, the speaking duration of the children's speakers is slower than the adult speakers [7, 8]. Figures 1 and 2 show the pitch and speaking rate of one of the recorded wav files of the children and adults used in experiments. From the figure 1, it is clear that child speakers have taken 2.416 seconds to utter a sentence, whereas Figure 2 shows that the adult speakers have taken 2.112 seconds to speak the same sentence. From figures it is also clear that the pitch of the child speakers is also higher than that of the adult speakers. Thus, prosody modification plays an important role during the speech recognition process.

From the studies, it was found that most of the publicly available ASR system works well with the adult speech, but provides less speech recognition rate when tested using the children speech [9]. It is necessary to build an ASR system for children of Punjabi language under mismatched conditions due to all these reasons.

So, in this research paper, the authors used the second method to increase the speech recognition rate of the children's

ASR system by minimizing the acoustic mismatch between the children's and adult's speech with the help of prosody modification. The prosody modification deals with modifying the pitch and duration or speaking rate of the speech signals without affecting the spectral and temporal features of the speech signal. The results presented in this research work explore the efficiency and effectiveness of pitch and duration prosody modification for improving the recognition rate under mismatched acoustic conditions. Prosody modification was done on the training speech as well as on the testing speech. Remarkable improvements were noticed in the speech recognition rate in both cases. Along with prosody modification, speaker adaptive training (SAT) is also incorporated in the Punjabi-ASR system based on feature-space maximum likelihood linear regression (fMLLR).

The rest of the research study is organized in the following manner: Section 2, presents the prior research work done to improve the speech recognition rate under mismatched acoustic conditions. The method used for prosody modification of the speech is discussed in the 3rd section, along with the pitch and duration (speaking rate) modification factors used for the alteration of the speech. In section 4, the speech dataset and experimental setup used for developing the Punjabi-ASR system is explained. In section 5, recognition results for the baseline system and system developed using prosody modification are discussed in detail. In the last section, research work is concluded.

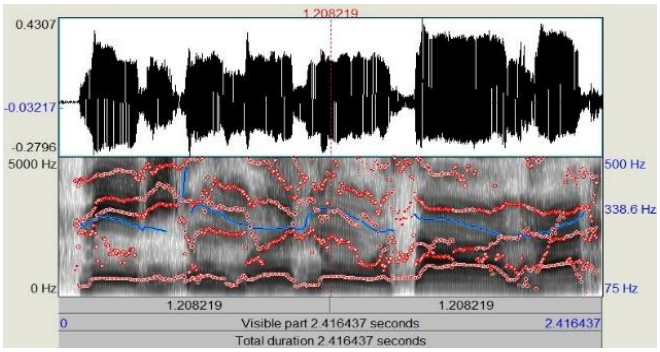

Figure 1. Child waveform

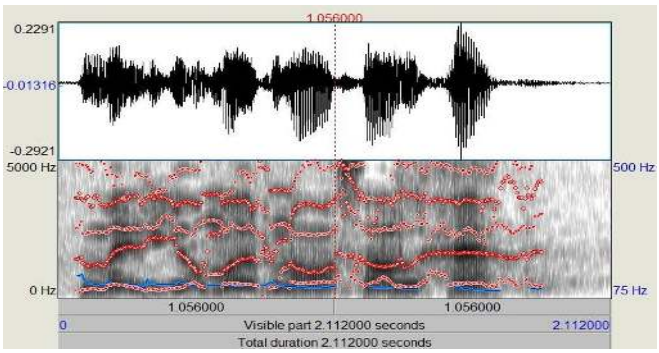

Figure 2. Adult waveform

2. RELATED WORK

This section highlights the recent work in the field of ASR, both under matched and mismatched acoustic environments. According to Prashanth and Panayiotis [9], adult-to-child transfer learning is vital for recognizing children's speech in various mismatched situations. The authors used multiple large vocabulary speech corpora for the development of children's ASR. The system was developed using GMM-HMM and DNN acoustic models. Along with this system was evaluated for the transfer learning versus standard adaptation techniques. The findings confirmed the advantages of age-dependent transfer learning. Guglani and Mishra [10] worked on developing an ASR system for the Punjabi language. The authors compared the use of pitch, Fundamental Frequency Variation (FFV), Yin, and SAcC features. In comparison to the SAcC, FFV, and Yin featured ASR systems, the performance of the ASR system designed with the pitch features is found to be the best. Results show that the pitch features reduce the word error rate (WER) by 1.5%. Mittal and Singh [11] developed a Punjabi ASR system for mobile devices using multiple acoustic models such as context-dependent (CD) untied, context-independent, CD tied, and CD deleted interpolation models. Puneet and Navdeep used 64, 32, 16, and 4 GMMs for evaluating the performance of the ASR system in terms of WER and accuracy. The results demonstrate that the CD untied acoustic model outperforms all other acoustic models in terms of accuracy. Shahnawazuddin et al. [3] examined the use of prosody modification for the development of a speaker-independent ASR system with a higher recognition rate. The main goal of the research work conducted by the authors was to create an ASR system that is less impacted by acoustic changes caused by speakers. The system was trained using the adult speech and tested using the adult and children's speech data. To overcome the adult and children's acoustic mismatches, the authors modified the speaking rate and pitch of the adults. Results show the relative improvements of 27.0% and 11.5% for children's and adult's data sets. In addition, several techniques were proposed by the researchers to improve the acoustic variabilities. Different front-end feature extraction techniques perceptual linear prediction (PLP), spectrum-based feature extraction, and Mel-Frequency cepstral coefficients (MFCC) have been used to extract the acoustic features [1, 9, 12-14]. Researchers have also made minor changes in the feature extraction process implemented in the front-end, these pitch features are also used for improving the speech recognition rate [10, 13, 15-17]. Due to the presence of high variabilities in the pitch and duration (speaking rate) of the adult and children, different studies were found in the literature to overcome these variabilities with the help of pitch and time scaling [6, 18, 19]. In several works, researchers introduced speaker normalization with the help of vocal tract length normalization (VTLN) technique, whereas acoustic model adaptation achieved with the help of maximum likelihood linear regression (MLLR), maximum a-posteriori (MAP), and SAT with fMLLR [1, 3, 9]. To perform the prosody modification, various techniques have been proposed in the literature and prosody parameters were analyzed [20-22]. Research studies included in the literature show that there is very less work done for the regional languages under the mismatched acoustic conditions and the recognition rate is also very poor. Additionally, earlier research has shown that prosody plays a significant influence in correctly recognizing mismatched speech with a better recognition rate.

3. PROSODY MODIFICATION FACTORS AND METHODS

Prosody modification of the acoustic or speech signal deals with alteration of the pitch and duration (time) without

affecting the spectral, temporal distortions, naturalness, and message of the signal [19, 20]. Prosody modification is done either at the training time or during the testing of the ASR system. The pitch and duration parameters of the speech signal are modified using the Eqns. (1) and (2)

$$\begin{aligned}&\text{Prosody (Speech after pitch modification)}\\&= \text{Prosody (Children or Adult speech)}\\&* \text{Pitch Modification Factor } (\alpha)\end{aligned} \quad (1)$$

$$\begin{aligned}&\text{Prosody (Speech after duration modification)}\\&= \text{Prosody (Children or adult speech)}\\&* \text{Duration Modification Factor } (\beta)\end{aligned} \quad (2)$$

where, prosody $\in$ pitch, duration, α= Pitch Modification Factor, and β= Duration Modification Factor.

The values of the pitch and duration modification factors used during the experimental process are shown in Table 1. Pitch scale modification is done to change the fundamental frequency (f0) to squeeze or stretch the harmonic spaces present in the spectrum whereas duration modification is done to alter the speaking rate of the speech signal without changing the characteristics of the speech signal. For this pitch contour to be compressed or stretched to enlarge or reduce the time duration of the signal. Values of the modification factors $\alpha, \beta > 1$ indicate the pitch or time expansion and $\alpha, \beta < 1$ indicates the pitch or time compression. For all the files present in the training as well as testing datasets, pitch period transposition or modification was done by using the modification factors (α) 1.5, 1.30, 1.20, 1.10, 0.90, 0.80, 0.70, and 0.60. The corresponding duration modification factors (β) used for the time scaling are 0.65, 0.75, 0.85, 0.95, 1.10, 1.25, 1.35 and 1.50.

Table 1. Pitch and duration modification factors for speech prosody alteration

Pitch Modification Factor (α)	Duration Modification Factor (β)
1.5	0.65
1.30	0.75
1.20	0.85
1.10	0.95
0.90	1.10
0.80	1.25
0.70	1.35
0.60	1.50

From the literature, it was found that diverse techniques are present for modification of the prosody [19, 20, 23, 24]. The following techniques synchronous overlap and add (SOLA), pitch-SOLA (PSOLA), overlap and add (OLA), harmonic plus noise model (HNM), STRAIGHT, and discrete cosine transform (DCT) [24] are used for prosody modification. Researchers can use PSOLA modification for both pitch and time scale prosody modification whereas SOLA and OLA techniques can be used for time-scale prosody modification. PSOLA technique further divided into different types: 1) Linear-Predictive-PSOLA (LP-PSOLA), 2) Frequency Domain-PSOLA (FD-PSOLA), and 3) Time-domain - PSOLA (TD-PSOLA).

LP-PSOLA technique uses the concept of residual excited vocoders for both pitch and time scale prosody modification whereas FD-PSOLA technique can be used only for scaling the pitch of the speech signal. TD-PSOLA can be used for the moderation of both pitch and time scale prosody but it causes phase and spectral distortions because of the direct synchronization of the speech sounds. In this work, LP-POSLA digital signal processing technique is used for the alteration of the speech signal.

3.1 Determination of time instants of significant excitation for LP-PSOLA and prosody modification

The modified speech files' quality depends upon the position of the glottal closure instants (GCI) or significant excitation. Pitch synchronization is accomplished by determining the GCI and fundamental frequency (F0), utilizing zero frequency filtering (ZFF). Excitation source information is used by the LP-PSOLA method for prosody modification. In the linear prediction (LP) analysis, the residual signal is thus used as an excitation signal. The residual signal is obtained by a time-varying zeros filter using LP coefficients (LPCs) linked with each time instant. By doing residual manipulation, limited distortion is present in the speech signal synthesized with the help of altered LP residual and LPCs.

The steps involved in processing the speech signal to track the instants locations of significant excitation GCIs are

(1) The first step is to differentiate the input speech signal (S[n]) to remove the low frequency (time-varying) influences present in the S[n].

$$X[n] = S[n] - S[n-1] \quad (3)$$

where, X[n] is the differenced speech.

(2) Pass the speech signal X[n] obtained from step 1 (Eq. (3)) twice through a zero frequency ideal resonator. The primary purpose for using a zero-frequency resonator is that the time-varying characteristics of the vocal tract system are unaffected by the resonator's output discontinuities, i.e. by sending the signal via the cascade of the two resonators, the influence of vocal tract system resonance is reduced.

$$Y_1[n] = -\sum_{k=1}^{2} a_k Y_1[n-k] + X[n] \quad (4)$$

where, $Y_1[n]$ = resonator output, a_1 = -4 and a_2 = 6.

$$Y_2[n] = -\sum_{k=1}^{2} a_k Y_2[n-k] + X[n] \quad (5)$$

where, $Y_2[n]$ = resonator output, a_1 = -4 and a_2 = 1. This process is zero frequency signal filtering and equivalent to signal integration four times.

(3) The trend removal in speech signal $Y_2[n]$ is done by average subtraction over 20ms at each sample.

$$Y[n] = Y_2[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} Y_2[n+m] \quad (6)$$

Here, 2N+1 in the above equation represents the sample numbers used for trend removal in the window.

(4) The resulting signal Y[n] in Eq. (6) is called a filtered or ZFF signal.

Following the desired prosody modification, a new

excitation signal (LP residual) is produced from the residual signal utilising the desired pitch and duration modifications in the speech signal. There are mainly four steps used for prosody modification of the speech signal using instants of significant excitation. Figure 3 illustrates the steps to perform pitch and duration modification of the speech signal and also mentioned in Ref. [20]:

1) Finding the significant excitation (GCI) instants from the input speech signal.
2) Generation of the new GCI's sequences is done according to required pitch and duration prosody.
3) Using modified GCI sequence for the generation of prosody modified LP residual signal.
4) The speech signal is synthesized with the help of LP residual signal obtained after step 3 and LPCs.

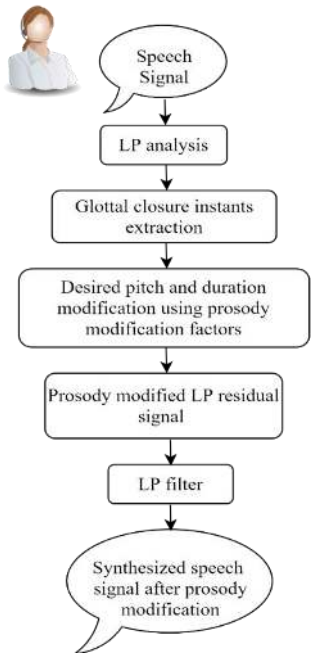

Figure 3. Block diagram for pitch and duration prosody modification

4. EXPERIMENTAL SETUP FOR PUNJABI-ASR SYSTEM

In this section, information is divided into two subsections. The details of the training and testing speech dataset used for the development of the Punjabi-ASR system are presented in the first subsection. The second subsection presents the ASR model shown in Figure 4 and the experimental setup used for conducting the experiments.

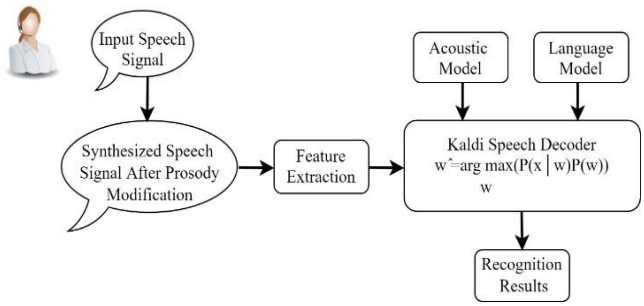

Figure 4. ASR model for recognizing speech with acoustic variabilities using prosody modification

4.1 Punjabi speech dataset

This research paper mainly focuses on building a Punjabi-ASR system and improving the speech recognition rate of the system using prosodic features. Three different Punjabi language speech datasets were used for evaluating the performance of the Punjabi-ASR system under mismatched acoustic conditions. Punjabi speech recordings of adults and children were sampled at a 16 kHz rate and stored in a wav file. These speech dataset wav files were recorded in the speech lab, noise-free, and noisy environmental conditions in the Malwa and Majha Punjab region. A total of 14.5-hour hand-transcribed Punjabi adult speech dataset is used for training the Punjabi-ASR system. The training speech dataset contains 8345 utterances spoken by 59 native speakers (35 male and 24 female) of Malwai and Majhi Punjabi dialects. This collected speech data is named as PunjCorA_Train Punjabi adult speech dataset. Two speech datasets were used to test the ASR system. The first testing dataset was used to check the performance of the ASR system using adult speech and was named as PunjCorA_Test. The adult testing dataset PunjCorA_Test contains 3.2 hours of speech files collected from 18 Malwai and Majhi speakers (12 male and 6 female) with 1455 utterances. The second testing dataset was used to check the performance of the developed ASR system by utilizing children's speech and named as PunjCorCh_Test. The children's testing dataset PunjCorCh_Test built with the help of 30 (6-17 years) child speakers. PunjCorCh_Test speech dataset consisted of 4.5 hours of speech data with a total of 720 utterances of children's speech. Along with the speech dataset, authors also have to prepare meta-data for the Kaldi toolkit. This metadata includes the following files:

- utt2spk (utterance to speaker) – file contains the utterance-speaker mapping.
- spk2utt (speaker to utterance) – file contains speaker-utterance lists.
- wav.scp – file contains the absolute path of the recorded speech file with utterance-id.
- Text – file contains utterance-id and transcription (written text) of the utterances.
- corpus.txt – file contains transcription (written text) of each utterance.

4.2 ASR model and experimental setup

Figure 4 shows the proposed Punjabi speech recognition system using prosody modification. Povey [25] speech recognition toolkit is used to conduct all the experiments.

4.2.1 Feature extraction

MFCC feature extraction is used in the front end for extracting the acoustic features. The pre-emphasis is given by equation_ and the value of the pre-emphasis factor α used during the feature extraction was 0.97.

$$H(z) = 1 - \alpha z^{-1} \quad (7)$$

The signal is passed through the hamming window after it has been framed into shorter frames.

$$w(n) = 0.54 - 0.46 \cos(2\pi n/N\text{-}1)\ 0 \le n \le N-1 \quad (8)$$

Here the hamming window is represented by w(n). The first and second-order derivatives of the 13 mel-cepstrum

coefficients were extracted using a 40-channel Mel-filter bank and a frame-length of 25ms with a frame-shift overlap of 10ms for frame-blocking. These 13 mel-cepstrum coefficients are time spliced and converted to 117-dimensional vectors with the frame's context size of nine. As a result, 117-D acoustic feature vectors were converted to 40-D acoustic feature vectors utilising the front-end adaption methods linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT). Authors have also used adaptation techniques MLLR, fMLLR for speaker independent, and SAT. For handling the inter and intra-speaker variability in children's, vocal tract length normalization (VTLN) is used [26].

4.2.2 GMM-HMM and DNN-HMM acoustic model

The interrelation between the phonemes and a speech signal units that make up articulation or speech is represented by an acoustic model in ASR. In this work, the Kaldi toolkit was configured on the Linux-Ubuntu operating system for statistically modeling of each phone. 3-state left-to-right hidden Markov models (HMMs) were used for phone modeling. A GMM-HMM Punjabi-ASR system was employed as the baseline Punjabi speech recognizer. The baseline system consists of 100,124 gaussians. Along with the GMM-HMM system, a hybrid DNN-HMM based Punjabi-ASR system was developed. The posterior probabilities of the GMM based system were replaced by the DNN-HMM based Punjabi-ASR system [27]. Our DNN based system is built using 4 - 8 hidden layers with 1K-2K hidden units in each hidden layer.

4.2.3 Language model (LM)

In ASR to predict the next word spoken, a LM is utilized to provide a probability for a word sequence. The mathematical expression shown in Figure 4 is the base of the ASR system. This expression is derived from Eq. (9). In the equation, W signifies the sequence of words, and x denotes the observations that reflect the feature vectors generated from the acoustic data. Mathematical expression in Eq. (9) is used by the speech recognizer to find more appropriate words [13, 28, 29].

$$\widehat{w} = arg \max_{w} P(w \mid x) \tag{9}$$

Preceding equation is rewritten using Bayes' rule as

$$\widehat{w} = \arg \max_{w} \frac{P(x|w)P(w)}{P(x)} \tag{10}$$

In Eq. (10), the LM probability is represented by P(w) and it gives word sequences a probability estimate and defines:

- the speaker's possible words
- by training on texts, the likelihood across all potential sequences

An ARPA-based trigram model was trained using the transcription (Text file) of the training data for decoding the speech.

4.2.4 Decoder

After training the ASR system, the final step is the decoding process. In this work, weighted finite-state transducers (WFSTs) [30, 31] are employed in our ASR system for decoding the system. WFSTs help in the simple representation of the language models, statistical models (HMMs), dictionaries, and other ASR outputs [32].

5. RECOGNITION RESULTS

In this section results of the baseline system and ASR system developed after prosody modification of the speech were discussed.

5.1 Baseline Punjabi-ASR system results

The results of the Punjabi-ASR system were evaluated with reference to WER and presented in Table 2. WER was calculated using Eq. (11).

$$WER = \frac{S + D + I}{TW} \tag{11}$$

where, TW indicates the total words used during testing, S is the number of substitutions, I is used to indicate the number of insertion errors and D is the deletion error in the test.

The baseline system was trained using the PunjCorA_Train dataset and tested using the adult dataset (PunjCorA_Test) as well as the children dataset (PunjCorCh_Test). From the results shown in Table 1, it is clear that the children's speech is difficult to recognize as compared to the adult's speech due to various acoustic variabilities present in the children's speech. In this work, some more speech files have been added to the existing speech dataset used by Bhardwaj and Kukreja [13] for training the system. The acoustic models were trained using mono phone, triphone, and a combination of DNN, and HMM. In our experiments, acoustic adaptation techniques MMLT and SAT were used. Authors have also performed experiments with SAT with fMLLR transform in our system.

From Table 2 it is clear that the ASR system provides the best results (7.89% WER) for adult speech when trained using the DNN-HMM model with MLLT, SAT, and VTLN.

Table 2. Results of the baseline speech recognition system trained with adult speech and tested with both adult and children speech

Acoustic Model	WER (%)	
	PunjCorA_Test Adult Dataset	PunjCorCh_Test Children Dataset
Monophone	22.23	73.50
Triphone	13.15	65.33
Triphone (LDA + MLLT)	12.98	61.26
Triphone (LDA + MLLT+SAT)	9.76	58.12
Triphone (LDA+MLLT+SAT+VTLN)	9.20	56.84
DNN-HMM	14.28	60.20
DNN-HMM (LDA + MLLT)	10.55	53.39
DNN-HMM (LDA + MLLT + SAT)	8.45	47.53
DNN-HMM (LDA+MLLT+SAT+VTLN)	7.89	41.30

5.2 Prosody modification results

In this subsection, the results of the Punjabi ASR system using prosody modification were discussed. Prosody modification was done on training as well as on testing speech. Steps to perform the prosody modification are shown in Figure

3 and modification factors used for pitch and time scales are presented in Table 1.

5.2.1 Recognition results after prosody modification on training speech

Table 3 shows the WER for the Punjabi-ASR system tested using adult's and children's speech. Pitch and duration modification factors used during the training phase for pitch and time scaling of the adult speech are α (1.50, 1.30, 1.20, and 1.10) and β (0.65, 0.75, 0.85, and 0.95). As the pitch of the children is higher than the adults, so pitch enhancement of the adult speech is done by $\alpha > 1$. Whereas, the speaking rate of the children is slower as compared to the adults, so duration modification is done by $\beta < 1$. After performing the prosody modification, the Punjabi-ASR system was trained using the prosody modified speech and PunjCorA_Train dataset. Results obtained after the alteration of the prosody of the training speech signal using different values of α and β are shown below in Table 3. While recognizing the children's speech, the system provides the best WER of 19.79% with modification factors α=1.20 and β=0.85. In the case of adult speech recognition, best results were obtained with α=1.10 and β=0.95.

Table 3. WER (%) results for the Punjabi- ASR system trained with the prosody modified speech and tested using adults, children's speech

		WER (%)			
Modification Factors	α	1.50	1.30	1.20	1.10
	β	0.65	0.75	0.85	0.95
Speech Dataset (Testing)	Adult	8.23	7.89	7.31	**7.15**
	children	29.20	24.89	**19.79**	23.56

5.2.2 Recognition results after prosody modification on testing speech

After prosody modification on the testing speech, Table 4 illustrates the recognition results for the Punjabi-ASR system. When prosody modification was done on the testing speech, the pitch of the children's speech was decreased by using $\alpha < 1$ and increased the duration by using $\beta > 1$. The Recognition system was trained using the PunjCorA_Train dataset and tested using prosody-modified adult's and children's speech. Results obtained after the alteration of the prosody of the testing speech signal using different values of α and β are shown below in Table 4.

From the results, it is clear that prosody modification during the testing increases the WER for adult speech. This happens due to the increase in the acoustic variabilities of the prosody modified adult speech and original adult speech dataset (PunjCorA_Train). On the other hand, there is a small degradation in the speech recognition rate while recognizing the children's speech. The system provides a WER of 22.59% with modification factors α=1.25 and β=0.80.

Table 4. WER (%) results for the Punjabi- ASR system trained with the adult's dataset (PunjCorA_Train) and tested using prosody modified adults, children's speech

		WER (%)			
Modification Factors	α	0.90	0.80	0.70	0.60
	β	1.10	1.25	1.35	1.50
Speech Dataset (Testing)	Adult	**14.12**	18.90	20.05	24.65
	children	24.56	**22.59**	27.95	32.78

5.2.3 Recognition results for ASR system trained using PunjCorA_Train, prosody-modified speech, and children's speech

Tables 3 and 4 present the findings., which indicate that the Punjabi ASR system provides a better recognition rate for the adult speakers as well as for child speakers when pitch and duration modification was done on training speech. So, to increase the speech recognition rate, the experiments were also conducted by adding children's speech dataset (PunjCorCh_Train) used along with prosody modified speech and PunjCorA_Train speech dataset during the training of the system. Table 5 shows the recognition results when the system was trained with PunjCorCh_Train, PunjCorA_Train, and prosody modified speech. The system provides WER of 8.79% and 7.15% for children and adult speech recognition.

Table 5. WER (%) results for the Punjabi- ASR system trained with PunjCorCh_Train, PunjCorA_Train, and prosody modified speech and tested using adults, children's speech

		WER (%)			
Modification Factors	α	1.50	1.30	1.20	1.10
	β	0.65	0.75	0.85	0.95
Speech Dataset (Testing)	Adult	8.23	7.89	7.31	**7.15**
	children	11.57	10.15	**8.79**	9.68

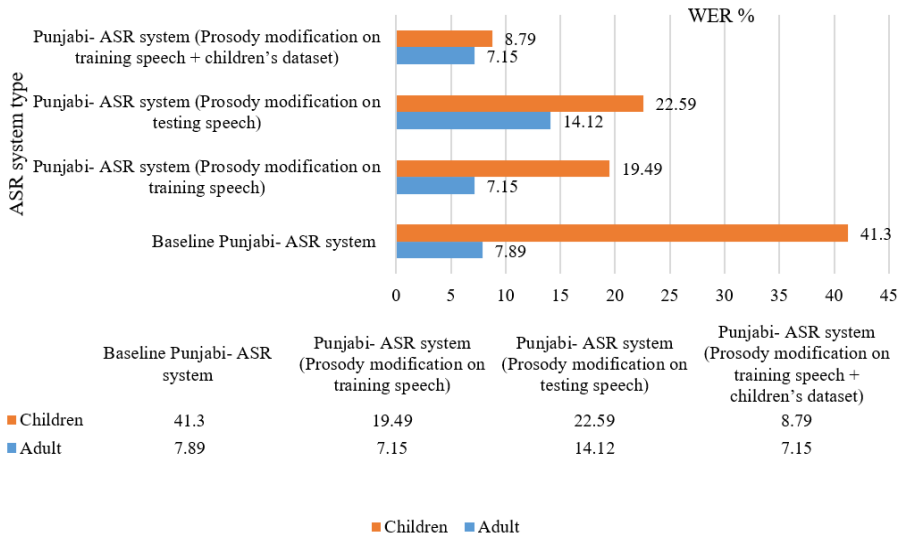

Figure 5. Comparison of baseline ASR and ASR system using prosody modification

5.2.4 Comparison of baseline ASR system and Punjabi - ASR system developed using prosody modification

In this subsection, bassline system results and results obtained after prosody modification on training and testing speech are presented together in Figure 5. Best results for the adults and children speech recognition were observed for the case when the Punjabi-ASR system was trained using both adult's and children's speech. When prosody modification was done during the testing phase, there was a degradation in the adult speech recognition rate. Hence, pooling of the children's speech was done only at the training phase for enhancing the speech recognition rate.

6. CONCLUSION

In this work, the prosody modification and acoustic model adaptation methods were used to enhance the speech recognition rate of the Punjabi-ASR system under mismatched acoustic conditions. Pitch and duration prosody modification factors are used for the alteration of the speech signal to reduce acoustic mismatches. For developing the ASR system Kaldi toolkit is used with MFCC in the front end. Results presented in this paper provide supporting evidence that the ASR system developed with the help of prosody provides a better recognition rate than the baseline system. When the system is trained with the help of adult speech and prosody modified adult speech, the WER for children is reduced by 21.81%. The system provides the best results of 7.15%, 8.79% WER for adults and children speech when trained using adult, prosody modified adult and children speech. A massive reduction in the WER for children validates the use of prosody modification and acoustic adaptation.

REFERENCES

[1] Shivakumar, P.G., Potamianos, A., Lee, S., Narayanan, S. (2014). Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In WOCCI, 15-19.

[2] Nagano, T., Fukuda, T., Suzuki, M., Kurata, G. (2019). Data augmentation based on vowel stretch for improving children's speech recognition. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 502-508. https://doi.org/10.1109/ASRU46091.2019.9003741

[3] Shahnawazuddin, S., Adiga, N., Kathania, H.K., Sai, B. T. (2020). Creating speaker independent ASR system through prosody modification based data augmentation. Pattern Recognition Letters, 131: 213-218. https://doi.org/10.1016/j.patrec.2019.12.019

[4] Shahnawazuddin, S., Adiga, N., Kathania, H.K. (2017). Effect of prosody modification on children's ASR. IEEE Signal Processing Letters, 24(11): 1749-1753. https://doi.org/10.1109/LSP.2017.2756347

[5] Lee, S., Potamianos, A., Naraanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J. Acoust. Soc. Amer, 105(3): 1455-1468. 10.1121/1.426686

[6] Li, C., Qian, Y. (2019). Prosody usage optimization for children speech recognition with zero resource children speech. Proc. Interspeech 2019, pp. 3446-3450. https://doi.org/10.21437/Interspeech.2019-2659

[7] Martins, I.P., Vieira, R., Loureiro, C., Santos, M.E. (2007). Speech rate and fluency in children and adolescents. Child Neuropsychology, 13(4): 319-332. https://doi.org/10.1080/09297040600837370

[8] Robb, M., Gilbert, H., Reed, V., Bisson, A. (2003). A Preliminary study of speech rates in young australian english-speaking children. Contemporary Issues in Communication Science and Disorders, 30: 84-91. https://doi.org/10.1044/cicsd_30_s_84

[9] Gurunath Shivakumar, P., Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. Computer Speech and Language, 63: 101077. https://doi.org/10.1016/j.csl.2020.101077

[10] Guglani, J., Mishra, A.N. (2020a). Automatic speech recognition system with pitch dependent features for Punjabi language on KALDI toolkit. Applied Acoustics, 167: 107386. https://doi.org/10.1016/j.apacoust.2020.107386

[11] Mittal, P., Singh, N. (2019). Development and analysis of Punjabi ASR system for mobile phones under different acoustic models. International Journal of Speech Technology, 22(1): 219-230. https://doi.org/10.1007/s10772-019-09593-x

[12] Guglani, J., Mishra, A.N. (2021). DNN based continuous speech recognition system of Punjabi language on Kaldi toolkit. International Journal of Speech Technology, 24: 41-45. https://doi.org/10.1007/s10772-020-09717-8

[13] Bhardwaj, V., Kukreja, V. (2021). Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions. Applied Acoustics, 177: 107918. https://doi.org/10.1016/j.apacoust.2021.107918

[14] Vegesna, V.V.R., Gurugubelli, K., Vuppala, A.K. (2019). Application of emotion recognition and modification for emotional Telugu speech recognition. Mobile Networks and Applications, 24(1): 193-201. https://doi.org/10.1007/s11036-018-1052-9

[15] Umesh, S., Sinha, R. (2007). A study of filter bank smoothing in MFCC features for recognition of children's speech. IEEE Transactions on Audio, Speech, and Language Processing, 15(8): 2418-2430. https://doi.org/10.1109/TASL.2007.906194

[16] Shahnawazuddin, S., Dey, A., Sinha, R. (2016). Pitch-adaptive front-end features for robust children's ASR. Proc. Interspeech 2016, pp. 3459-3463. https://doi.org/10.21437/Interspeech.2016-1020

[17] Ghai, S., Sinha, R. (2015). Pitch adaptive MFCC features for improving children's mismatched ASR. International Journal of Speech Technology, 18(3): 489-503. https://doi.org/10.1007/s10772-015-9291-7

[18] Shahnawazuddin, S., Adiga, N., Sai, B.T., Ahmad, W., Kathania, H.K. (2019). Developing speaker independent ASR system using limited data through prosody modification based on fuzzy classification of spectral bins. Digital Signal Processing: A Review Journal, 93: 34-42. https://doi.org/10.1016/j.dsp.2019.06.015

[19] Vegesna, V.V.R., Gurugubelli, K., Vuppala, A.K. (2018). Prosody modification for speech recognition in emotionally mismatched conditions. International Journal of Speech Technology, 21(3): 521-532. https://doi.org/10.1007/s10772-018-9503-z

[20] Rao, K.S., Yegnanarayana, B. (2006). Prosody modification using instants of significant excitation.

IEEE Transactions on Audio, Speech, and Language Processing, 14(3): 973-980. https://doi.org/10.1109/TSA.2005.858051

[21] Prasanna, S.R.M., Govind, D., Rao, K.S., Yegnanarayana, B. (2010). Fast prosody modification using instants of significant excitation. In Speech Prosody 2010-Fifth International Conference.

[22] Govind, D., Mahadeva Prasanna, S.R. (2013). Dynamic prosody modification using zero frequency filtered signal. International Journal of Speech Technology, 16(1): 41-54. https://doi.org/10.1007/s10772-012-9155-3

[23] Quatieri, T.F., McAulay, R.J. (1992). Shape invariant time-scale and pitch modification of speech. IEEE Transactions on Signal Processing, 40(3): 497-510. https://doi.org/10.1109/78.120793

[24] MuraliSankar, R., Ramakrishnan, A.G., Rohitprasad, A.K., Anoop, M. (2001). Baced Pitch modification. Proc. SPCOM 2001 6th Biennial Conference, pp. 114-117.

[25] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Vesely, K. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (No. CONF). IEEE Signal Processing Society.

[26] Dendani, B., Bahi, H., Sari, T. (2021). Self-supervised speech enhancement for Arabic speech recognition in real-world environments. Traitement du Signal, 38(2): 349-358. https://doi.org/10.18280/TS.380212

[27] Kukreja, V., Dhiman, P. (2020). A Deep Neural Network based disease detection scheme for Citrus fruits. 2020 International Conference on Smart Electronics and Communication (ICOSEC), pp. 97-101. https://doi.org/10.1109/ICOSEC49089.2020.9215359

[28] Kaur, H., Bhardwaj, V., Kadyan, V. (2021). Punjabi children speech recognition system under mismatch conditions using discriminative techniques. Innovations in Computer Science and Engineering: Proceedings of 8th ICICSE, pp. 195-203.

[29] Bhardwaj, V., Bala, S., Kadyan, V., Kukreja, V. (2020). Development of robust automatic speech recognition system for children's using Kaldi Toolkit. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 10-13. https://doi.org/10.1109/icirca48905.2020.9182941

[30] Mohri, M., Pereira, F., Riley, M. (2002). Weighted finite-state transducers in speech recognition. Computer Speech & Language, 16(1): 69-88. https://doi.org/10.1006/csla.2001.0184

[31] Kumar, D., Kukreja, V. (2021). N-CNN based transfer learning method for classification of powdery mildew wheat disease. 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 707-710. https://doi.org/10.1109/ESCI50559.2021.9396972

[32] Houari, H., Guerti, M. (2020). Study the influence of gender and age in recognition of emotions from Algerian dialect speech. Traitement du Signal, 37(3): 413-423. https://doi.org/10.18280/ts.370308