Multimedia Tools and Applications (2023) 82:3973–3994
https://doi.org/10.1007/s11042-022-13435-5

Prosody features based low resource Punjabi children ASR and T-NT classifier using data augmentation

Virender Kadyan[1] · Taniya Hasija[2] · Amitoj Singh[3]

Received: 19 April 2021 / Revised: 23 April 2022 / Accepted: 2 July 2022 /
Published online: 20 July 2022

Abstract

Automatic children speech recognition is always challenging due to limited corpus and varying acoustic features. One among those is zero speech corpus and large acoustic variability which limits the power of learning of training dataset. To overcome this issue, an effort has been made to build two types of systems: ASR and Tonal-Non tonal (T-NT) classifiers. Initially, robust features are added into the front phase using prosody embedded feature vectors. Various prosody features are combined with MFCC feature vectors which outperformed conventional Mel Frequency Cepstral Coefficients (MFCC) features only. A small reduction in Word Error Rate (WER) is obtain on the original train and test dataset. To further enhance the recognition rate, training data scarcity is remove through two-level augmentation approach: external prosody modifications (using pitch and time scaling parameters) and internal augmentation using speed perturbation approaches (using 3, 4, and 5 way methods). For that purpose, an original and augmented dataset is pooled to learn more statistical parameters information. Significant improvement in the performance of both systems are observe due to two-level augmentations and prosody embedded features. Finally it achieve a relative improvement of 13.1% and 18.3% for ASR and T-NT classifier systems over the baseline system which are processed on a modified train and original test set respectively.

✉ Amitoj Singh
amitojsingh@psou.ac.in

Virender Kadyan
vkadyan@ddn.upes.ac.in

Taniya Hasija
taniya@chitkara.edu.in

[1] Speech and Language Research Centre, School of Computer Science, University of Petroleum & Energy Studies (UPES), Energy Acres, Bidholi, Dehradun, Uttarakhand, India

[2] Centre of Excellence for Speech and Multimodal Laboratory, Chitkara University Institute of Engineering & Technology, Chitkara University, Rajpura, Punjab, India

[3] School of Science and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala, Punjab, India

Keywords Children Punjabi ASR · Prosody features · T-NT classifier · Data augmentation

Abbreviations

ASR	Automatic Speech Recognition
DNN	Deep Neural Network
T-NT	Tonal -Non Tonal
MFCC	Mel Frequency Cepstral Coefficient
WER	Word Error Rate
VTLN	Vocal Tract Length Normalization
LPCC	Linear Predictive Cepstral Coefficient
RASTA-PLP	Relative Spectral Perceptual Linear Predictive Coding
PLP	Perceptual Linear Prediction
PNCC	Power-Normalized Cepstral Coefficients
GFCC	Gammatone Frequency Cepstral Coefficients
HMM	Hidden Markov Model
DTW	Dynamic Time Warping
DE	Differential Equation
GA	Genetic Algorithm
GMM	Gaussian Mixture Model
VTLP	Vocal Tract Length Perturbation
MMI	Maximum Mutual Information
MPE	Minimum Phone Error
MLE	Maximum Likelihood Equation
DCT	Discrete Cosine Transformation
POV	Probability of Voicing
LDA	Linear Discriminative Analysis
MLLT	Maximum Likelihood Linear Transformation
PS	Pitch Scaling
TS	Time Scaling

1 Introduction

Speech is one of the most powerful and important source of communication among human-human and human-machine. Although the processing of spoken speech signal plays a dominant role. It tried to bridge a gap in understanding, identification, and recognition of various speaker utterances. Nowadays, these utterances are found to be beneficial in various spoken language translators, speech to text converters, emotion recognition, dictation process, and kids youtube application or devices [64]. Success of such applications or devices are only possible when one processes speech as clearly as a human can do. There exist various factors such as speaking styles, surrounding noise, speed of speaking which may sometimes cause skipping of some words, and acoustic variability etc. [4, 11, 29, 46] which largely impact an ASR performance. Tackling of these parameters are a matter of great attention in current state-of-the-art research [4, 36]. Apart, researchers achieved success in order to overcome these variabilities in adult speech systems [68]. Despite in comparison to various well-developed adult ASR [55], children's speech has more variability due to high pitch, lack of proper pronunciation of words, and formant variations. It leads to difficulty in building of an efficient

ASR system [13]. These variabilities are tried to be removed from processed input speech signal and different recognition systems achieved performance improvement [35, 50]. Consequently, they tried to build a child ASR system in different mismatched or matched conditions. It is employed on varying child-adult datasets in train sets and later Vocal Tract Length Normalization (VTLN) approaches is applied on it to further boost its performance [54, 67]. Till date children's ASR systems are well developed in English, Swedish, Italian, Japanese, and German language databases only [40]. These databases mostly employ children speech lies within the age group of 6–15 years only. Whereas very little work has been reported in Asian languages children speakers. Some of these languages are well-spoken but they are still at zero resource stage. On the other hand most of the researchers have focused on American English but a major obstacle is requirement of building ASR systems in native languages of a speaker. In India, there are 22 national languages, from which only Hindi and English adult ASR systems [7, 42] are well developed but very little work has been done on other languages like Punjabi ([15, 26], Kaur and Singh [23, 24], [58]). The other major issue in the development of ASR for such language is due to high cost involved in the collection of corpus and its annotation. Few works have been presented in the past with self-developed adult speech but children's speech is at zero stage. To build a system one needs to collect a large amount of speech corpus. To overcome such challenges various data augmentation approaches are employed in earlier work [28]. These works are found to be beneficial in enhancement of each ASR system performance. It is possible through voice conversion [51], noise augmentation [2], formant modification [52], and prosody modification [50], etc. They tried to artificially enhance training data and tested it on original dataset for general purpose ASR applications and test augmentation for domain-specific ASR systems. To further process these pooled datasets, complex speech information is constructed which further requires attention in extraction of unique feature vectors. The purpose of feature selection over partially labelled data is to pick a subset of accessible features with the least redundancy and the highest relevancy to the target class [47]. It is performed by removal of unwanted or irrelevant information through various conventional front end approaches: Linear Predictive Cepstral Coefficient (LPCC), Relative Spectral Perceptual Linear Predictive Coding (RASTA-PLP), Perceptual Linear Prediction (PLP), MFCC, Power-Normalized Cepstral Coefficients (PNCC), and Gammatone Frequency Cepstral Coefficients (GFCC) (Dua et al. 2019). However, the cepstral feature vectors extracted from these feature extraction techniques are not alone sufficient but some other variability factors are also needed to indulge. Some of these factors are associated particularly with children's speech like vocal tract length variations [66], formant differences [21], or other acoustic and linguistic variations [4]. Likewise, various hybrid prosody features are voice probability, pitch (f0 gradient), intensity (energy), and loudness [62]. To capture large acoustic variations, experiments have been performed by various researchers to build effective children ASR systems [13, 33]. Also, psychological characteristics, speaking style, and inter-speaker variations are well captured through these prosodic features whereas conventional feature extraction approaches alone only capture phonetic features of a speech signal. To further indulge the characteristics of the tonal aspect of a language, pitch features play a crucial role. This tonal information in real scenarios changes the meaning of the word due to the presence of variation in pitch features [26]. Performance of various tonal languages (like Mandarin, Vietnamese and Thai) ASR systems enhanced with inclusion of their tonal features in or with hybrid front end feature vectors using prosodic features like Pitch or fundamental frequency [49]. Till date very few works have been presented in the Punjabi language, so efforts are required to build the Punjabi Children ASR system that can help in the building of Children ASR and T-NT word classifiers.

In this paper, an effort has been made in building of a Punjabi Children's speech corpus. A self-created limited data has been constructed to overcome zero resource conditions. To build a baseline Punjabi Children system tonal characteristics are also indulge with default MFCC approach using MFCC+prosodic features. It enhances both the systems performance i.e. ASR and Tonal-Non Tonal systems. An external data augmentation method is later introduced on two prosody parameters, i.e. pitch and time variations. Further pooling of original and synthetic data is performed to fix the challenges of data scarcity. The combined dataset is processed to generate robust feature vectors to introduce large train speaker variabilities. It contributed in the reduction of system error to a larger extent than that of the original corpus based ASR and T-NT classifier system.

The remainder of the paper is structured as follows: The motivation behind the work is discussed in Section 2 and also present state of the art related work in Section 3. The technique used for extracting hybrid feature vectors along with methods of data augmentation is given in Section 4. Section 5 further outlines the proposed structure for ASR and T-NT classifiers. In Section 6, experimental setup and discussions are later discussed. Conclusion and future studies are finally discussed in Section 7.

2 Motivation

The development of a native language ASR is an emerging area of research. ASR systems developed so far generally struggle in various Indian languages [57]. There are 22 official languages, and individuals prefer to communicate in their own native language. Only a few ASR systems have been created till date for some languages: Hindi and Marathi. Punjabi was observed as one such language which was spoken by 105 million pupils but still considered as an under-resourced language [3, 19, 26]. Although some work has been reported in adult speech, children's ASR systems are still in their infant stage due to lack of a resources like speech corpus and its high annotation cost. The tonal influence of the language is also act as a barrier in development of an efficient recognition system because tonality lowers recognition performance. The goal of this research is to create a Children Punjabi ASR and T-NT classifier systems that can improve the performance through inclusion of tone and acoustic variability information. Artificial methods such as data augmentation are used to tackle dataset scarcity issues. Later, new features are added to the basic features, which tried to make an easier output hypothesis through recognition of T-NT and natural text words.

3 Literature review

In earlier days, ASR in Punjabi language was in development stage and some of these systems have been developed mostly for isolated, connected, and continuous speech. Initially, an isolated recognition system was implemented by Ravinder [45] using Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) techniques. They achieved an accuracy of 94% with DTW than that of HMM (having an accuracy of 91%) approach. Dua et al. [10] also worked on HMM approach using HTK toolkit and achieved an accuracy of 94–95%. Later, Kadyan et al. (2018) proposed two-hybrid modeling classifiers approaches that outperformed conventional HMM approaches. It was possible through reduction of computational complexity of system training. It was performed with refinement of model parameters using DE + HMM and GA + HMM hybrid classifiers. It was examined from previous work that most of the work discussed in Punjabi only employed adult speech, but for children's speech, ASR systems and other systems: T-NT classifiers were not yet explored. They effectively need comprehensive training data with efficient feature vector extraction (Kadyan et al. 2018; [45]) approaches. Feature extraction is the first stage in any speech recognition system. [48] employed a genetic algorithm through which feature similarities are computed. During the second stage, community discovery algorithms were classified to characterise it into clusters. In the third stage, a genetic algorithm with a new community-based repair operation selects particular characteristics. The performance of the given method was evaluated on nine benchmark classification tasks. Number of machine learning approaches were used for the extraction of robust feature vectors, A similar approach is used by [16]. Different front end techniques were also investigated by Anusuya and Katti [1]. They performed comparative study with multiple filter banks on LPC, LPCC, PLP, RASTA-PLP and MFCC feature extraction techniques. Till date, various clean corpus ASR had reached a high recognition level, however when the device was tried to be recognized in a noisy environment in comparison with other existing systems it has been analysed that its performance has drastically degraded. Zhao and Wang [70] presented a noise-robust ASR architecture through implementation of GFCC feature extraction approaches. They demonstrated how Gammatone Frequency Cepstral Coefficients (GFCC) differ from MFCC based extraction algorithms. Long-term or hybrid characteristics, on other hand, indulge loudness, pitch, and strength information which often have a major effect on the recognition of a spoken utterance. Additional strategies for adequate performance enhancement of ASR systems were explored by Litman et al. [37]. It was feasible through short-term features which were extracted by concatenating prosodic features. Furthermore, analysis between prosodic differences among correct and incorrect recognized words were performed using TOOT information corpus. Kathania et al. [22] were represented the role of prosody features in children's speech systems. They implemented their system in mismatch condition using WSJCAM0 (an adult training set) and performed testing using PF-STAR British English speech corpora. By including prosodic features, class discrimination was also increased and inter speaker variation were reduced with 16% relative improvement using MFCC+prosodic features. The linguistic study also played an important role in building of an ASR and T-NT classifier system. In 2006, Liang Wang et al. [63] demonstrated prosodic information that helped in the classification of tonal and non-tonal languages. It was only possible through 3 layers of feed-forward training along with normalization of feature parameters. Consequently, they represented their work with Gaussian Mixture Model (GMM). The result concluded that the GMM classifiers outperformed Neural Network classifiers with an accuracy of 87.1% [65]. Zhang and Hirose [69] proposed Chinese speech recognition approaches with lexicon tone which sought to introduce hypotheses of pitch anchoring. Lei et al. [34] were presented their work on Mandarin tonal language. These languages had well-developed corpus whereas languages like Punjabi were at infant or zero resource conditions in adult and children speech. Various augmentation strategies were employed till date using Tacotron2 [30], GAN [53], external modification using acoustic variability through formant [21], or prosody features [50]. Tom Ko et al. [28] implemented data augmentation which tried to increase the training dataset by changing the speed of an audio signal. Three versions of speech signals were originated from original train signals by using constant factors of 0.9, 1.0, and 1.1. They tried to achieve a relative improvement of

4.3% by experimented it with all tasks of data augmentation. Later, S Shahnawazuddin et al. [50] were employed prosody modification-based data augmentation and created their speaker-independent ASR system with a relative improvement of 27%. It was only possible through artificial creation of children's train data by modify their pitch and speaking rate parameters only in train sets. Further, Du and Yu [6] presented a novel augmentation approach, in which synthetization of data was performed where an end to end text to speech system was trained and synthetic speech was generated from unseen speakers. They achieved 30% relative improvement compared to without any data augmented systems. In past various data augmentation techniques i.e. Vocal tract length perturbation (VTLP) [3, 17], tempo perturbation [38], and speed perturbation were investigated by Geng et al. [12] and they concluded that speed perturbation were outperformed among all the three augmentation approaches. The similar related work which indulge hybrid or robust front end features with prosody features, tonal-non tonal classifiers on low resource languages is discussed in Table 1.

It is evident from the previous ongoing literature that Punjabi is a tonal language, and till now, little attention has been paid to tonal aspect of the language. The main focused of the researchers was to build an adult ASR system without much focused upon the influence of tonal words. Due to inability of Punjabi children's speech corpus, an effort has been made in this article to utilise speed perturbation capabilities on original Punjabi children's speech. Later experiment is performed in later sub-sections where prosody embedded information are extracted in order to create a robust ASR system along with tonal-non tonal classifier.

4 Theoretical background

4.1 Feature extraction

4.1.1 MFCC (Mel frequency cepstral coefficient)

The first step in development of an ASR method is to remove non-redundant information and discard odd information which tried to classify the components of an input speech signal (Dua et al. [8, 9]. The following steps are used to perform key functionality:

- Acoustic analogue signals are initially transformed by discrete time stamps to digital data. The amount of energy has been increased at high frequencies in pre-emphasis stage to increase the signal-to-noise ratio. To optimize higher frequencies than lower frequencies, the data is transferred to a filter. The filter for high passes is represented as:

$$\mathrm{X(n)} = \mathrm{x(n)}-\mathrm{a(n}-1) \tag{1}$$

  where X(n) is the output signal when an input signal is x(n) and the filter coefficient is a. These signals are further framed into frames of 20–40 ms with an overlaps of 10 ms on adjacent frames. Signal windowing is performed on each frame to restrict the discontinuous, and hamming window is computed using it.

$$\mathrm{W(n)} = 0.54-0.46\cos\left(\frac{2\pi \mathrm{n}}{\mathrm{N}}-1\right) \qquad 0\leq \mathrm{n} \leq \mathrm{N}-1 \tag{2}$$

Table 1 State-of-the-art work done with inclusion of hybrid front end features on Indian language ASR and T-NT classifier systems

Author's	Dataset	Feature extraction	Acoustic Modeling	Summary
Kathania et al. [22]	WSJCAM0 (an adult training set) and testing set is PF-STAR British English speech corpora.	MFCC/PLP + Prosodic features (voice probability, voice intensity and loudness)	DNN-HMM and GMM-HMM	They implemented ASR system in mismatched conditions and also tried to concatenate three prosodic features with MFCC or PLP feature vectors and showed a R.I. of 16%
Liang wang et al. (2007)	Multi Language CALLFRIEND, OGI-TS and OGI-22 corpus.	Tonal and non-tonal classification using prosodic features.	GMM-HMM	They represented their work on tonal non-tonal classification using Gaussian Mixture Model (GMM). The result concluded that GMM classifier outperformed Neural Network classifiers with an accuracy of 87.1%
Billa [5]	Gujrati, Tamil and Telgu speech corpus	MFCC	end-to end Long Short Term Memory training method having Connectionist Temporal Classification	In this work, author has done monolingual training and multilingual training on Indian low resource languages. Mono lingual system showed 6.5 to 25.5% R.I. and in multilingual training 4.5% to 11.1% R.I. was achieved.
Dua et al. [8, 9]	Hindi Speech database having 100 speakers	MFCC	Discriminative trainings	Maximum mutual information (MMI) and minimum phone error (MPE) discriminative techniques on Hindi corpus showed that MMI and MPE were outperformed than maximum likelihood equation (MLE) techniques.
Lata and Arora [31]	Punjabi Spoken data from natives speakers of Punjab	–	–	To prove that Punjabi is tonal language they have concluded that Punjabi has three tones- high falling, low rising and middle tone, also Punjabi has five tonal characters.
Lata and Arora [32]	Punjabi speech corpus	–	–	They represented that tonal aspects of /h/ sound in Punjabi language was varied with word positions i.e. initial, middle, and final, but it showed tonal behaviour in some cases.

- The previous output is processed to transform the output signal into a frequency domain by filtering operations, i.e. Transformation of Fast Fourier (FFT). It helps to translate the signal from a different time domain to a separate frequency domain [71].
- The human auditory system is modelled by a Mel filter bank which is capable of processing of an input speech signal at various frequencies than that of many which have linear or nonlinear distributions. The Mel function here adjusts the magnitude of the frequencies using FFT to Mel frequency band. It has a broad range of frequencies which are defined as:

$$\mathrm{Mel}(f(t,k)) = 2595\log_{10}\left(1+\frac{f(t,i)}{700}\right) \tag{3}$$

And f(t, i) (FFT) is computed as:

$$F(t,i) = \left|\frac{1}{N\sum_{k=1}^{N-1}\left(e^{-\frac{2\pi jkn}{N}}\right)f_k}\right|\left(S'(n)\right) \tag{4}$$

where $i \rightarrow 0, 1, 2, 3 \ldots\ldots\ldots\ldots \frac{N}{2}-1$

- On the Mel filter bank output, log out is applied and the acoustic variants that are not necessary for speech recognition are removed.
- Finally, Discrete Cosine Transformation, a cepstrum of log output is generated (DCT). Input to DCT is spectrums of log-power Mel and it results into MFCC cepstral characteristics.

$$\mathrm{DCT} = \sum_{k=1}^{K}(\log\log S_k)\left(\cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{k}\right]\right) \tag{5}$$

where S_k is the Mel spectrum output and k = 1,2,3,4…….,K; n = 1, 2, ……, K. DCT decorrelates parameters which are determined by log output transformation. It tried to compute 13 MFCC features.

4.2 Prosody feature vectors

The accuracy of a system gets degraded due to an acoustic and linguistic variability of a language. In many languages, words having same syllabus but varied tones made them lexically different. Therefore, tone recognition is a critical component in a speech recognition tasks. It is an unavoidable requirement for building of an acoustic model in any tonal language. Ittried to utilize knowledge of related sound, which tried to contribute in improvement of performance of an ASR systems. Since auditory interpretation of a tone depends on the pitch of an audio signal, it is possible to perform tonal classification which are based upon pitch speed and pitch level. The inherited property of a periodic signal is calculated by virtual pitch monitoring. It is a fundamental frequency (f0) given by [59]. In northern states of India and northern eastern Pakistan, the spoken language is Punjabi. It is one of the world's 10th most commonly spoken languages which exhibits a tonal form. It has three tones that are phonetically different: high dropping, low rising, and mid-level. It has five tonal characters: (dh) /t/, (bh) /p/, (Jh) /tʃ/, (dh) / /, and (gh) /k/ [32]. Some toneme character where located at three different locations: initial, mid or end which tried to change the meaning of a particular words as:

- kàr (high-falling) house	- kòṛā (high-falling) horse
- kár (low-rising) dandruff	- kóṛā (low-rising) leper
- kar (level) do/hands	- koṛā (level) whip

Tonal effect is observed on one syllable and can recognize it through stress caused on that syllable. However, in Mandarin Chinese language there exists five tones [63] but in Punjabi language, it has only three types of tones. This tone information is captured through pitch where pitch (perceived ‘height’ of the human voice, which depends upon the vocal cord’s rapid variation) is used to differentiate the meaning between words of that language [44]. To further exhibit more acoustic features one needs to examine the impact of various prosody features. It generally carries suprasegmental aspects of every sentence. The cue of a pitch is dependent upon its fundamental frequency (f0), which is computed through vibrations of its vocal folds [56]. As an illustration, samples of fundamental frequencies (f0) contours of Punjabi as well as Hindi languages are shown in Fig. 1. One can observe that Punjabi has more impact on f0 counter variations than that of a non-tonal language like Hindi.

Extraction of robust and extra feature information at syllable levels is always found to be beneficial in getting knowledge of the tonality of a particular syllable. These features indulge a long-term characteristics of utterances which tried to help in providing different context-related information of that utterance. One way of indulging such information is possible through robust prosody features. Various prosodic features are extracted in the past to overcome the poor performance of an ASR system. To capture pitch estimates from its input speech signal [39] an autocorrelation approach is employed. It is possible through similarity calculation between two corresponding waveforms. At various time intervals, waveforms are compared at each interval for measurement of their uniformity. Consequently, dissimilarity also increases with an increase in a time lag. The infinite discrete autocorrelation function x[n] is calculated by:

$$R_x(v) = \sum_{n=-\infty}^{\infty} x[n]x[n+v] \tag{6}$$

Similarly, the autocorrelation function for a finite discrete function x’[n] with N size is calculated as:

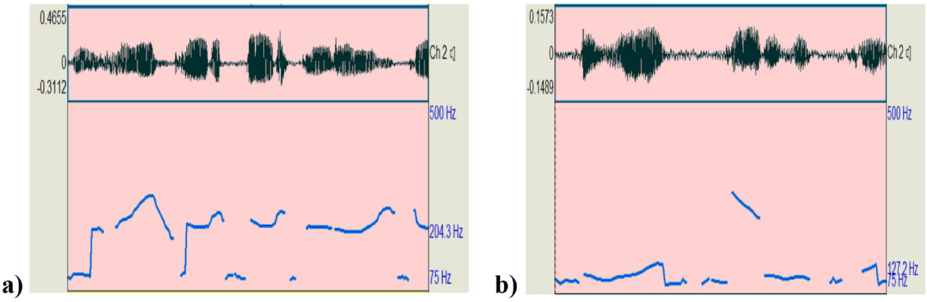

Fig. 1 Variation of f0 contour for a sample utterance in Hindi (**a**) and Punjabi languages(**b**)

$$R_{x'}(v) = \sum_{n=0}^{N-1-v} x'[n]x'[n+v] \tag{7}$$

To calculate the cross-correlation between two functions x[n] and y[n] is given by:

$$R_{xy}(v) = \sum_{n=-\infty}^{\infty} x[n]y[n+v] \tag{8}$$

The highest peak obtained after Eq. 8 results into the generation of fundamental frequency through autocorrelation. After every 25 ms framework with 10 ms overlapping, pitch information is extracted. To further process raw f0 frequency, pitch smoothing and pitch trimmings are employed which results into original f0 information. The calculation of Probability of Voicing (POV) is performed by the percentage of voiced and unvoiced energy present in an input speech signal. POV characteristics are used to prevent unvoiced segments of speech and it results into an increase in ASR system robustness [14]. If c is the value of the cross-correlation function which lies in the range of $-1 < c < 1$, the POV output feature is computed by:

$$f = 2\left((1.0001-c)^{0.15}-1\right) \tag{9}$$

Apart, intensity or stress features are also extracted from variation caused among f0 contour segments. The direction of f0 changes (which is presented with raising and falling of f0 contours) provide intonation information. Stressed words have higher energy and larger f0 movements and longer durations [39]. The voice quality features are extracted to get the long-term data of the speech signal and in speaker verification processes. F0 jitter and shimmer disturbance measurements have also proved to help in identification of vocal characteristics. Jitter is defined as a parameter of frequency variation which varies from one cycle to another, and shimmer refers to the sound wave's amplitude variations. To capture the voice quality features, the time of fundamental period is detected and after evaluating the onset time of the glottal pulses, the jitter, average absolute difference between two consecutive periods, can be calculated for defining measurement shapes.

$$\text{jitter} = \frac{1}{N-1} \sum_{i=1}^{N-1} T_i - T_{i-1} \tag{10}$$

$$\text{jitter(local)} = \frac{\text{Jitter}}{\frac{1}{N-1}\sum_{i=1}^{N-1} T_i} \times 100 \tag{11}$$

where T_i is the duration of each period in seconds and N is the number of periods? Also, the shimmer is derived by measuring the signal's overall peak amplitude. The value of shimmer (local) is the average absolute difference between two consecutive amplitudes. The shimmer is computed as:

$$\text{shimmer(local)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i-1}|}{\frac{1}{N-1}\sum_{i=1}^{N-1} T_i} \times 100 \tag{12}$$

where A_i is the amplitude of each duration and N is the number of periods. Apart, the Harmonic to Noise Ratio (HNR) feature is extracted, which is computed through the ratio of periodic component to the non-periodic component of voiced speech [61]. The employed formula for calculation of HNR is:

$$\mathrm{HNR} = 10 \times \log_{10}\left(\frac{\mathrm{AC_v(T)}}{\mathrm{AC_v(0)}-\mathrm{AC_v(T)}}\right) \tag{13}$$

Here, the peak corresponding to signal's period at an index position is represented by$AC_v(T)$. Finally, 8 prosody features are raw f0, envelope f0, fundamental frequency f0, POV, intensity, voice quality, HNR, and loudness, are further experimented in this paper along with static MFCC features on Punjabi children systems.

4.3 Data augmentation

Data collection involves more cost and time, so the accuracy of an ASR system is dependent upon the stimulation of large training audios [41]. Various approaches like DNN are not able to infer test samples when sufficient training data is not available. So data augmentation is highly required in such cases so that possible test samples can match with the trained dataset. Several techniques have been used for the data augmentation of training data sets. One of them is speed perturbation, where a warped time signal is produced from the original signal. In speed perturbation, the input signal is $y(t)$ and a is a time-warping factor and output signal after the perturbation is given by $y'(t)$ as:

$$\mathrm{y}'(\mathrm{t}) = \mathrm{y(at)} \tag{14}$$

Some changes are done in their frequency domain using:

$$\mathrm{X(f)} \rightarrow \frac{1}{\alpha}\mathrm{X}\left(\frac{1}{\alpha}\mathrm{f}\right) \tag{15}$$

where $X(f)$ and $\frac{1}{\alpha}X\left(\frac{1}{\alpha}f\right)$ represent the Fourier transform of $y(t)$ and $y'(t)$ respectively. When FFT is applied to $y'(t)$, the shift in the frequency component is produced by the warping factor [12]. In speed perturbation, spectral envelope and audio duration both are changed. As the duration of the signal is affected by speed perturbation, therefore the number of frames are also varied during MFCC feature extraction. Sox speed function is used for speed perturbation through sox manipulation tool [27].

Pitch scaling is another approach for data augmentation, when a Spectral content or perceived pitch of the audio signal is modified, without affecting its time duration and evaluation. The resulting audio signal has the same time duration but having a modified pitch. The vibration rate of the vocal fold is changed for modifying the pitch scale of the input audios. During voicing excitation of the vocal tract, frequencies are scaled using:

$$\mathrm{y'_k} = \beta \mathrm{y_k} \tag{16}$$

with associated change in the pitch, the period is given by:

$$\mathrm{P}'(\mathrm{t}) = \frac{\mathrm{P(t)}}{\beta} \tag{17}$$

The modified excitation function of the resultant model is as follows:

$$\mathrm{e}'(\mathrm{t}) = \sum_{\mathrm{k}=1}^{\mathrm{N}} \mathrm{a'_k(t)} \cos\left[\Omega'(\mathrm{t})\right] \tag{18}$$

where

$$\Omega'(\mathrm{t}) = \left(\mathrm{t}{-}\mathrm{t}_0{'}\right)\beta \mathrm{y}_\mathrm{k} \tag{19}$$

where $t_0{'}$ is modified onset time. The excitation amplitude is the original excitation amplitude. The new amplitude and phase functions are given as:

$$\mathrm{M'}_\mathrm{k}(\mathrm{t}) = \mathrm{M}(\beta\mathrm{y}_\mathrm{k};\mathrm{t}) \tag{20}$$

and

$$\psi'_\mathrm{k}(\mathrm{t}) = \psi(\beta\mathrm{y}_\mathrm{k};\mathrm{t}) \tag{21}$$

The final pitch information contains only voiced speech data which results into a synthetic pitch augmented speech signal.

5 Corpus collection

Building a corpus in zero resource conditions is always challenging. An effort has been made to collect speech data from school pupils who lie in the age group of 7–13 years. The corpus is collected through microphones at a 16 kHz sampling rate. The baseline ASR system is framed on 4 hours 20 minutes data which is manually transcribed through native Punjabi speakers. Table 2 showed a detailed overview of the train and test corpus employed in all experimental studies. For the language model, 5 k unique lexicon has been involved in system building. All the experiments are performed using Kaldi toolkit [43] on Ubuntu operating system.

6 System overview

The proposed system for enhancing the accuracy of Punjabi children's speech and T-NT classifier is summarized in Fig. 2. Initially, an effort has been made to build ASR and T-NT classifier systems on original dataset by training model parameters with conventional front-end vectors or on concatenated individual prosody feature vectors. Later, all extracted features are combined with default 13 energy parameters where each frame is extracted using MFCC technique which is generally derived from an energy parameters of each frame as depicted in Eqs. 1 to 5. The size of each frame is 25 ms and the frame shift is 10 ms. After computation of frame information, a hamming window is applied along with a 23 channel mel filterbank. To enhance the performance of an ASR system time-variant dynamic features i.e. trajectories are

Table 2 An overview of Punjabi train and test children speech corpus

Term	Train	Test
No. of Speakers	39 (21 male, 18 female)	6 (3 male, 3 female)
No. of Unique Sentences	1885	485
Type of Corpus	Continues	Continues
Age Group	7 to 13	7 to 13
No. of words	24,536	2845
No. of Unique words	4643	1002

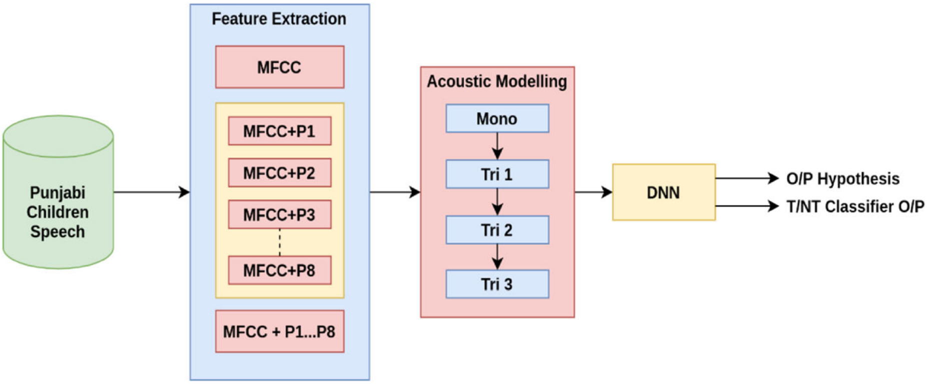

Fig. 2 Block Diagram of ASR system implemented by concatenation of MFCC features and Prosody Features for enhancing the performance of overall Punjabi ASR and T-NT classifier Systems

calculated and merged with 13 coefficients of MFCC. Trajectories features are also known as delta features (tri 1) which are computed by:

$$\mathrm{d_t} = \frac{\sum_{n=1}^{N} n(c_t - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{22}$$

Where d_t is delta coefficients and it is computed on frame t, c_t and c_{t-n} which are static coefficients and the value of N is kept as constant 2. After computation of delta features, 26 total features are computed. Further delta–delta features (tri2) are also extracted, which are the trajectories of delta features and are computed through:

$$\mathrm{dd_t} = \frac{\sum_{n=1}^{N} n(d_t - d_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{23}$$

where dd_t are delta-delta coefficient. These features are also known as the first and second derivatives of a signal. There arises further need in reduction of coefficients which are multiplied after delta and double delta. For converting it into smaller amounts of acoustically distinct units, Linear Discriminative Analysis (LDA) is implemented on the output of tri 2, which tried to reduce the required coefficients into 40 feasible dimensions. By estimating the probability of a new set of inputs which are belong to each class and the class with which the highest probability is evaluated in an output class. Consequently, Maximum Likelihood Linear Transformation (MLLT) is estimated over utterances, and exclusion of speaker-specific information is performed. Later LDA + MLLT is called for tri 3 modeling of the system where the word error rate is generated using DNN-HMM classifiers [60]. After implementation of MFCC, the next step is to embed prosody features along with 13 static MFCC feature vectors. Eight prosody features are extracted with the help of an OPENSMILE toolkit. The first prosody features are computed on an input speech signal is the voice probability feature (P1), which is concatenated with 13 MFCC features and then 14 aggregate features are given for computation of delta, delta-delta information. Similarly, f0 (P2), intensity (P3), loudness(P4), voice quality (P5), f0 raw (P6), f0 envelope (P7), and harmonic to noise ratio (P8) features are concatenated individually with MFCC features and then their corresponding

WER is computed. Finally, all the prosodic features are pooled together which resulted into a total of 21 feature vectors as shown in Fig. 2.

The step by step procedure for implementation of prosodic features concatenated with MFCC features on ASR system and T-NT classifier systems are as follows:

Step 1: Collection of original children Punjabi Speech data (male/female of age group 7 to 13 years) corpus.
Step 2: Initialize:

```
Segmentation and transcription of audios.
// if implementing ASR system on prosody features then transcription include text written format of audios
// if implementing T-NT classifier system then transcription includes the T keyword if spoken word is tonal, otherwise NT keyword for non-tonal words.
training_data = 1885 utterances
testing_data = 485 utterances
```

Step 3: Extract MFCC and prosody features from training and testing datasets as:

```
mfcc(training_data) // using Eq. (1), (2), (3), (4) and (5).
mfcc(testing_data) // using Eq. (1), (2), (3), (4) and (5).
prosody(training_data) // using Eq. (6), (7), (8), (9), (10), (11), (12), (13).
prosody(testing_data) // using Eq. (6), (7), (8), (9), (10), (11), (12), (13)
```

Step 4: Combine one to one prosody feature with MFCC features and repeat step 5 to step 9 for every combination.
Step 5: Do monophone training (mono) and aligning of monophone results. //HMM training
Step 6: Do delta training (tri 1) and align their phones. //using Eq. (22)
Step 7: Perform delta+delta training (tri 2) and also align their triphones. //using Eq. (23)
Step 8: Training of LDA + MLLT training (tri 3) on tri 2 output and aligning of its phones.
Step 9: Perform DNN-HMM model and further calculate Word Error Rate (WER) of the system.

Finally performance is analysed after comparing the results of an ASR system and T-NT classified system on number of prosody varied features combination with MFCC.

For evaluating the performance of an ASR system WER is computed where substitution of words (S) or deletion of words (D) or insertion of new words (I) is evaluated through:

$$\mathrm{WER\%} = \frac{\mathrm{S+I+D}}{\mathrm{N}} \times 100 \tag{25}$$

Another parameter used for performance evaluation is Relative Improvement (RI). RI is the absolute increase corresponding to a new value (N) with respect to old value (O):

$$\mathrm{RI\%} = \frac{\mathrm{N{-}O}}{\mathrm{O}} \times 100 \tag{26}$$

To further analyse the T-NT classifier system each tonal word mapping is done with T and non-tonal words with NT in manually method of transcription. The features embedded with

ASR are also tested with T-NT classifier system. The issue arises with both the system is of handling large feature vector information which is tackled using LDA approach in tri3 model.

Motivated by these facts we also further tried to increase the performance of both the systems by artificially increasing the training dataset. It is performed by modifying the original train corpus with respect to internal (using pitch modification) and external augmentation approach (using perturbation). The test utterances are kept in their original form. Since resampling of the original signal is performed using sox command. 3-way perturbation is evaluated by creating two additional copies of the original corpus using 0.9 and 1.1 scaling factor than that of the original rate and merged with original one. Similarly in 4-way and in 5 way 0.8, 0.9, 1.1, and 1.2 scaling factor has been used. Pitch sampling has also been used as data augmentation and 0.85 scaling factor is used for sampling. Augmented corpus statistically enhanced training data scarcity issue. Consequently, significant performance improvement is obtained as demonstrated in experimental analysis studies. Further, the training complexity increases system training time but keeps original processing time in the testing phase as represented by shown in Fig. 3.

7 Results and discussions

7.1 Performance analysis of baseline ASR and T-NT system using conventional and varying prosody features

Initially development test is investigated using 5-fold cross validation on both the systems then finally the proposed test set is evaluated on it. Different speakers and spoken dataset is involved in training and testing of the systems. In Table 3, the baseline Children ASR WER is computed by employing original train and test speech on conventional MFCC feature vectors using DNN-HMM approach. Further different WER is enlisted with MFCC+embedded prosodic feature vectors. On performing the comparison between different WER as shown in Table 3 major differences are noted. To alleviate the performance of the further systems' best output obtained on each embed, MFCC features lower WER-based prosody features are considered. Furthermore, the efficiency of the system is computed in consecutive sections. The explored results showed that a R.I. of 9.12% and 12.39% are obtained than that of baseline ASR and T-NT classifier systems.

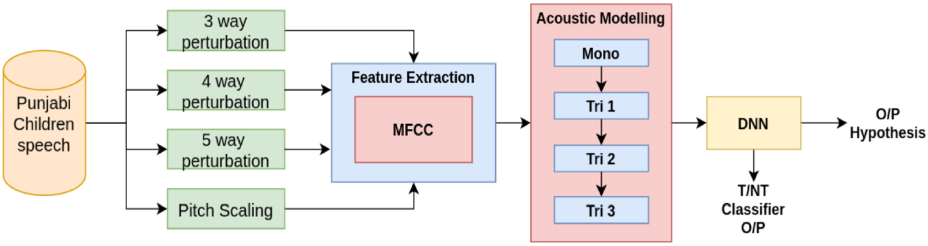

Fig. 3 Block Diagram summarized the steps involved in the implementation of enhanced Punjabi Children ASR and T-NT classifier system

Table 3 WER(%) obtained on concatenated MFCC features with varying prosodic features

Feature Extraction	ASR output	T-NT classifier output
MFCC (Baseline system)	14.91	21.62
MFCC+Voiceprob_sma(P1)	**13.55**	19.7
MFCC+F0_sma(P2)	13.78	**18.94**
MFCC+ pcm_intensity_sma(P3)	14.19	21.1
MFCC+pcm_loudness_sma(P4)	14.11	20.15
MFCC+voiceQuality_sma(p5)	14.28	19.61
MFCC+ HNR_smap(p6)	14.6	19.45
MFCC+ F0raw_sma(p7)	14.14	19.23
MFCC+ F0env_sma(p8)	15	19.37

7.2 Performance analysis of the system using concatenated prosodic features

Effect of performance of individual features on ASR systems has been shown in Table 3, later evaluation of the performance is done by combining these features or some of best output generated feature vectors. First, all features are combined with 13 static MFCC features and which does not show significant improvement. Later few selected prodigy features are combined on the basis of their output achieved in the previous section result. A significant improvement is obtained with a small fall in WER as shown in Table 4. Compared to baseline output, a significant relative improvement of 11.87% and 16.2% is obtained in ASR and T-NT classifier systems. On reducing the combination of two best-selected prosody features the WER is increased.

7.3 Performance analysis with augmented train dataset

To further overcome the issue of data scarcity, an artificial method of enhancement of training data is employed to augment only training data by keeping default test data. The original train corpus is modified using two key parameters: pitch and time scaling. The best optimal value of these parameters has been taken into consideration from our previous studies mentioned in [50, 51, 53]. It can be observed that both systems performed well when pooling of PS, TS, and original speech data has been performed but out of PS and TS augmented dataset only TS has more influence in the reduction of WER. After applying external augmentation using prosody modification, Kaldi-based internal augmentation using perturbation has been performed. Three types of internal augmentation using 3,4, and 5 way has been performed where 5 way showed a large reduction in WER with a R.I. of 13.1% and 18.3% in comparison to that of baseline systems as depicted in Table 5. These augmented datasets are later investigated with prosody embedded front-end features to enhance system accuracy. It can be observed from Table 5.

Table 4 WER(%) obtained on embedded prosodic features with default MFCC approach in ASR and T-NT classifier systems

Feature Extraction	ASR output	T-NT classifier output
MFCC+8 prosody	15.1	19.12
MFCC+ 4 prosody (f0+voice probability+intensity+loudness)	**13.14**	18.94
MFCC+2 prosody(F0+voice probability)	14	**18.10**

Table 5 WER(%) obtained by augmentation of data with pitch scaling, time scaling, and speed perturbation

Data augmentation	ASR output	T-NT classifier output
O (Original)	14.91	21.29
O+PS	14.35	21.26
O+TS	14.29	20.15
O+PS+TS	13.98	**17.56**
O+ 3 Way	**12.95**	21.16
O+4 Way	13.35	21.6
O+ 5 Way	13.42	20.9

7.4 Comparative analysis of proposed work with existing state of the art study

The majority of ASR research has been focused on various languages of different countries. Developing ASR systems for low resource is always act as a tedious task. Few attention has been paid to one such Punjabi language based ASR system. Where training data has been evolved from isolated to continuous, then to spontaneous words. In previous study only Punjabi adult speech was utilised, and much effort was put to achieve excellent performance in it. Children's speech corpora are still in their infant stage. External and internal augmentations are used on it to overcome human cost. It provided us natural like speech corpus which eliminate the zero children's speech corpus. To capture additional information from speech, robust features such as prosodic features are combined with baseline MFCC features, and further two systems, ASR and T-NT classifier, are used which performed better than that of other similar type of system but unlikely it performed in limited data scenario. It also outperformed with T-NT systems which is the first work presented for zero resource speech corpus. Table 6 showed the comparative study of the proposed work to the current state of the art work.

8 Conclusion

In this paper, we have explored two types of Punjabi Children speech systems: ASR and T-NT classifiers. In this regard, an effort has been made initially by building own speech corpus to overcome zero resource condition issues. Further, it focused on indulging robust feature vectors to capture relevant information of an input speech signal using conventional MFCC and 8 embedded prosodic features. Consequently, two-level augmentation has been performed on the training dataset using prosody modification (by pitch and time scaling) through external methods and internal speed perturbation approach (using 3,4, and 5 ways). In the first approach, optimal value through prosody modification has been performed by pooling of original + external augmented dataset. In the second approach, internal augmentation has been performed on the pooled augmented dataset. While analysing the system performance a small increased in WER performance is evident in ASR and T-NT classifier systems than that of each baseline systems. In both the systems, training complexity has been increased by keeping no change in the time of test set. Finally system achieved a R.I. of 13.1% and 18.3% for ASR system and T-NT classifier system along with data augmentation processes. In future, we wish to explore the effectiveness of formant modification in children's utterances by exploring formant modification approaches. Further to that, tactron2 and spectrogram augmentation needs to be performed which tries to address the shortcomings of data scarcity of the proposed implemented system.

Table 6 Comparative analysis with previous studies and proposed ASR/T-Nt Classifier systems

Author	Data set and its type	Feature Extraction Technique	Acoustic Modeling Technique	Performance
M. Dua et al. [10]	2760 Distinct words are employed (Isolated word dataset)	MFCC	HMM	In classroom environment it obtained WER of 4.37% and in open environment it obtained WER of 5.92%.
Kadyan et al. [18]	45,000 utterances (Isolated Word dataset)	MFCC, PLP, RASTA-PLP	HMM+GA (Genetic Algorithm), HMM+DE (Differential evolution)	An average word accuracy of 67.38%, 61.17% and 58.67% is achieved on MFCC, PLP and RASTA-PLP front end approaches.
Kadyan [20]	3611 sentences in train set, 422 sentences in test set (phonetically rich sentences)	MFCC and GFCC	GMM-HMM and DNN-HMM	On MFCC feature extraction, WER is 5.22% with DNN+HMM approach and for GMM+HMM it resulted into 7.01% using GFCC approach. Where GFCC is found to be effective.
Kaur and Kadyan [25]	1575 utterances in train and 584 in test (Continuous Children Speech)	MFCC	Discriminative techniques i.e. Maximum Mutual Information (MMI). bMMI, fMMI	On fMMI technique a RI of the system is 22%.
Taniya et al. [60]	1887 utterances in train, 485 in test (Continuous Children Speech)	MFCC	DNN+HMM	A baseline Children Speech recognition system is proposed with higher WER of 14.46%.
Proposed Work	1887 utterances in train, 485 in test sets (Continuous Children Speech)	MFCC and Prosody features	DNN+HMM	When individual prosody features are merged with MFCC, R.I. of 9% is observed on ASR and 12% is observed on T-NT classifier System on limited train dataset, however when four prosody features are combined, RI of 8% and 13% is observed. Later on, the use of augmentation has been proven to be more effective with improvement in systems performance.

Declarations

Conflict of interest We have no conflict of interest to declare.

References

1. Anusuya MA, Katti SK (2011) Front end analysis of speech recognition: a review. Int J Speech Technol 14(2):99–145. https://doi.org/10.1007/s10772-010-9088-7
2. Balam J, Huang J, Lavrukhin V, Deng S, Majumdar S, Ginsburg B (2020) Improving noise robustness of an end-to-end neural model for automatic speech recognition. https://arxiv.org/abs/2010.12715
3. Bawa P, Kadyan V (2021) Noise robust in-domain children speech enhancement for automatic Punjabi recognition system under mismatched conditions. Appl Acoust 175:107810
4. Benzeghiba M, De Mori R, Deroo O et al (2007) Automatic speech recognition and speech variability: a review. Speech Comm 49(10–11):763–786. https://doi.org/10.1016/j.specom.2007.02.006
5. Billa J (2018). ISI ASR system for the low resource speech recognition challenge for Indian languages. In INTERSPEECH 3207–3211
6. Du C, Yu K (2020) Speaker augmentation for low resource speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE 7719–7723. https://doi.org/10.1109/ICASSP40776.2020.9053139
7. Dua M, Aggarwal RK, Biswas M (2018) Performance evaluation of Hindi speech recognition system using optimized filterbanks. Engineering Science and Technology 21(3):389–398. https://doi.org/10.1016/j.jestch.2018.04.005
8. Dua M, Aggarwal RK, Biswas M (2019a) Discriminatively trained continuous Hindi speech recognition system using interpolated recurrent neural network language modeling. Neural Comput & Applic 31(10): 6747–6755
9. Dua M, Aggarwal RK, Biswas M (2019b) GFCC based discriminatively trained noise robust continuous ASR system for Hindi language. J Ambient Intell Humaniz Comput 10(6):2301–2314. https://doi.org/10.1007/s12652-018-0828-x
10. Dua M, Aggarwal RK, Kadyan V, Dua S (2012) Punjabi automatic speech recognition using HTK. Int J Comput Sci Issues (IJCSI) 9(4):359
11. Forsberg M (2003) Why is speech recognition difficult. Chalmers University of Technology
12. Geng M, Xie X, Liu S, Yu J, Hu S, Liu X, Meng H (2020) Investigation of data augmentation techniques for disordered speech recognition. Proc. Interspeech 696–700. https://doi.org/10.21437/Interspeech.2020-1161
13. Gerosa M, Giuliani D, Brugnara F (2007) Acoustic variability and automatic recognition of children's speech. Speech Comm 49(10–11):847–860. https://doi.org/10.1016/j.specom.2007.01.002
14. Ghahremani P, BabaAli B, Povey D, Riedhammer K, Trmal J, Khudanpur S (2014) A pitch extraction algorithm tuned for automatic speech recognition. In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE 2494–2498. https://doi.org/10.1109/ICASSP.2014.6854049
15. Goyal K, Singh A, Kadyan V (2021) A comparison of laryngeal effect in the dialects of Punjabi language. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-021-03235-4
16. Hakak S, Alazab M, Khan S, Gadekallu TR, Maddikunta PKR, Khan WZ (2021) An ensemble machine learning approach through effective feature extraction to classify fake news. Futur Gener Comput Syst 117: 47–58
17. Jaitly N, Hinton GE (2013, June) Vocal tract length perturbation (VTLP) improves speech recognition. In Proc. ICML workshop on deep learning for audio, speech and language (Vol. 117).
18. Kadyan V, Mantri A, Aggarwal RK (2017) A heterogeneous speech feature vectors generation approach with hybrid hmm classifiers. Int J Speech Technol 20:761–769. https://doi.org/10.1007/s10772-017-9446-9
19. Kadyan V, Mantri A, Aggarwal RK, Singh A (2019) A comparative study of deep neural network based Punjabi-ASR system. Int J Speech Technol 22(1):111–119. https://doi.org/10.1007/s10772-018-09577-3
20. Kadyan V (2018) Acoustic features optimization for Punjabi automatic speech recognition system. PhD diss. Chitkara University
21. Kathania HK, Kadiri SR, Alku P, Kurimo M (2020) Study of formant modification for children ASR. In ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE 7429–7433. https://doi.org/10.1109/ICASSP40776.2020.9053334

22. Kathania HK, Shahnawazuddin S, Adiga N, Ahmad W (2018) Role of prosodic features on children's speech recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE 5519–5523. https://doi.org/10.1109/ICASSP.2018.8461668
23. Kaur A, Singh A (2016a) Power-normalized cepstral coefficients (PNCC) for Punjabi automatic speech recognition using phone based modelling in HTK, second international conference on applied and theoretical computing and communication technology. IEEE Explore, ICATCCT2016, Bengaluru.
24. Kaur A, Singh A (2016b) Optimizing feature extraction techniques constituting phone based modelling on connected words for Punjabi automatic speech recognition, communicated in 5th International Conference on Advances in Computing, Communications and Informatics, IEEE Explore, ICACCI-2016, Jaipur
25. Kaur H, Kadyan V. (2020) Feature space discriminatively trained Punjabi children speech recognition system using Kaldi toolkit. Available at SSRN 3565906.
26. Kaur J, Singh A, Kadyan V (2020) Automatic speech recognition system for tonal languages: state-of-the-art survey. Archives of Computational Methods in Engineering:1–30. https://doi.org/10.1007/s11831-020-09414-4
27. Ko T, Peddinti V, Povey D, Khudanpur S (2015) Audio augmentation for speech recognition. In Sixteenth Annual Conference of the International Speech Communication Association.
28. Ko T, Peddinti V, Povey D et al (2017) A study on data augmentation of reverberant speech for robust speech recognition. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 5220–5224. https://doi.org/10.1109/ICASSP.2017.7953152
29. Kumar Y, Singh N, Kumar M, Singh A (2021) AutoSSR: an efficient approach for automatic spontaneous speech recognition model for the Punjabi language. Soft Comput 25:1617–1630. https://doi.org/10.1007/s00500-020-05248-1
30. Kwon O, Jang I, Ahn C, Kang HG (2019) Emotional speech synthesis based on style embedded Tacotron2 framework. In 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC). IEEE, 1–4. https://doi.org/10.1109/ITC-CSCC.2019.8793393
31. Lata S, Arora S (2012, May) Exploratory analysis of Punjabi tones in relation to orthographic characters: a case study. In Workshop on Indian Language and Data: Resources and Evaluation Workshop programme 76
32. Lata S, Arora S (2013, August) Laryngeal tonal characteristics of Punjabi—an experimental study. In 2013 International Conference on Human Computer Interactions (ICHCI). IEEE, 1–6 https://doi.org/10.1109/ICHCI-IEEE.2013.6887793
33. Lee S, Potamianos A, Narayanan S (1999) Acoustics of children's speech: developmental changes of temporal and spectral parameters. The Journal of the Acoustical Society of America 105(3):1455–1468. https://doi.org/10.1121/1.426686
34. Lei X, Siu M, Hwang MY et al (2006) Improved tone modeling for mandarin broadcast news speech recognition. In Ninth International Conference on Spoken Language Processing
35. Li C, Qian Y (2019) Prosody usage optimization for children speech recognition with zero resource children speech. In Interspeech 3446–3450. https://doi.org/10.21437/Interspeech.2019-2659
36. Li X, Wu X (2015) Modeling speaker variability using long short-term memory networks for speech recognition. In Sixteenth Annual Conference of the International Speech Communication Association.
37. Litman DJ, Hirschberg JB, Swerts M (2000) Predicting automatic speech recognition performance using prosodic cues, Proc. 1st North Am. Chapter Assoc. Comput. Linguist. Conf. 218–225 [Online]. Available: http://dl.acm.org/citation.cfm?id=974305.974334.
38. Long Y, Li Y, Zhang Q, Wei S, Ye H, Yang J (2020) Acoustic data augmentation for mandarin-English code-switching speech recognition. Appl Acoust 161:107175. https://doi.org/10.1016/j.apacoust.2019.107175
39. Mary L, Yegnanarayana B (2008) Extraction and representation of prosodic features for language and speaker recognition. Speech Comm 50(10):782–796. https://doi.org/10.1016/j.specom.2008.04.010
40. Milde B, Köhn A (2018) Open source automatic speech recognition for German. In Speech Communication; 13th ITG-Symposium 1–5 VDE
41. Nguyen TS, Stueker S, Niehues J, et al (2020) Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 7689–7693 https://doi.org/10.1109/ICASSP40776.2020.9054130
42. Passricha V, Aggarwal RK (2020) A comparative analysis of pooling strategies for convolutional neural network based Hindi ASR. J Ambient Intell Humaniz Comput 11(2):675–691. https://doi.org/10.1007/s12652-019-01325-y
43. Povey D, Ghoshal A, Boulianne G et al(2011) The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society

44. Rafi MS (2010) Semantic variations of Punjabi Toneme. Lang India 10(8):56–65 http://hdl.handle.net/123456789/543
45. Ravinder K (2010) Comparison of hmm and dtw for isolated word recognition system of Punjabi language. In Iberoamerican Congress on Pattern Recognition. Springer, Heidelberg. 244–252 https://doi.org/10.1007/978-3-642-16687-7_35
46. Rose R, Yin SC, Tang Y (2011) An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 4508–4511. https://doi.org/10.1109/ICASSP.2011.5947356
47. Rostami M, Berahmand K, Forouzandeh S (2020) A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty. J Big Data 7(1):1–21
48. Rostami M, Berahmand K, Forouzandeh S (2021) A novel community detection based genetic algorithm for feature selection. J Big Data 8(1):1–27
49. Shahnawazuddin S, Adiga N, Kathania HK (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
50. Shahnawazuddin S, Adiga N, Kathania HK, Sai BT (2020a) Creating speaker independent ASR system through prosody modification based data augmentation. Pattern Recogn Lett 131:213–218. https://doi.org/10.1016/j.patrec.2019.12.019
51. Shahnawazuddin S, Adiga N, Kumar K et al (2020b). Voice conversion based data augmentation to improve Children’s speech recognition in limited data scenario. Proc. Interspeech 2020, 4382–4386. https://doi.org/10.21437/Interspeech.2020-1112
52. Shahnawazuddin S, Adiga N, Sai BT, Ahmad W, Kathania HK (2019) Developing speaker independent ASR system using limited data through prosody modification based on fuzzy classification of spectral bins. Digital Signal Processing 93:34–42. https://doi.org/10.1016/j.dsp.2019.06.015
53. Shahnawazuddin S, Ahmad W, Adiga N, Kumar A (2020c,) In-domain and out-of-domain data augmentation to improve Children’s speaker verification system in limited data scenario. In ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). 7554–7558. IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053891
54. Shahnawazuddin S, Kathania HK, Dey A, Sinha R (2018) Improving children’s mismatched ASR using structured low-rank feature projection. Speech Comm 105:103–113. https://doi.org/10.1016/j.specom.2018.11.001
55. Shivakumar PG, Georgiou P (2020) Transfer learning from adult to children for speech recognition: evaluation, analysis and recommendations. Comput Speech Lang 63:101077
56. Shriberg E, Ferrer L, Kajarekar S et al (2005) Modeling prosodic feature sequences for speaker recognition. Speech Commun 46(3–4):455–472. https://doi.org/10.1016/j.specom.2005.02.018
57. Singh A, Kadyan V, Kumar M, Bassan N (2019) ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. Artif Intell Rev 53:1–32. https://doi.org/10.1007/s10462-019-09775-8
58. Singh A, Kaur N, Kukreja V et al (2022) Computational intelligence in processing of speech acoustics: a survey. Complex Intell Syst 8(2623):2661 https://doi.org/10.1007/s40747-022-00665-1
59. Talkin D, Kleijn WB (1995) A robust algorithm for pitch tracking (RAPT). Speech coding and synthesis 495:518
60. Taniya, Bhardwaj V, Kadyan V (2020) Deep neural network trained Punjabi children speech recognition system using Kaldi toolkit. In 2020 IEEE 5th international conference on computing communication and automation (ICCCA) (pp. 374-378). IEEE
61. Teixeira JP, Oliveira C, Lopes C (2013) Vocal acoustic analysis–jitter, shimmer and hnr parameters. Procedia Technology 9:1112–1122. https://doi.org/10.1016/j.protcy.2013.12.124
62. Ten Bosch L (2003) Emotions, speech and the ASR framework. Speech Comm 40(1–2):213–225. https://doi.org/10.1016/S0167-6393(02)00083-3
63. Wang L, Ambikairajah E, Choi EH (2006) Automatic tonal and non-tonal language classification and language identification using prosodic information. In International Symposium on Chinese Spoken language Processing. (ISCSLP) 485–496
64. Wang L, Ambikairajah E, Choi EH (2007a,) A novel method for automatic tonal and non-tonal language classification. In 2007 IEEE International Conference on Multimedia and Expo. IEEE. 352–355. https://doi.org/10.1109/ICME.2007.4284659
65. Wang L, Ambikairajah E, Choi EH (2007b) Automatic language recognition with tonal and non-tonal language pre-classification. In 2007 15th European Signal Processing Conference 2375–2379. IEEE.
66. Yadav IC, Shahnawazuddin S, Pradhan G (2019) Addressing noise and pitch sensitivity of speech recognition system through variational mode decomposition based spectral smoothing. Digital Signal Processing 86:55–64. https://doi.org/10.1016/j.dsp.2018.12.013
67. Yeung G, Alwan A (2018) On the difficulties of automatic speech recognition for kindergarten-aged children. In INTERSPEECH 1661–1665. https://doi.org/10.21437/Interspeech.2018-2297

68. Zehra W, Javed AR, Jalil Z et al (2021) Cross corpus multi-lingual speech emotion recognition using ensemble learning. Complex and Intelligent Systems 7:1–10
69. Zhang JS, Hirose K (2000) Anchoring hypothesis and its application to tone recognition of Chinese continuous speech. In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). IEEE. 3:1419–1422. https://doi.org/10.1109/ICASSP.2000.861859
70. Zhao X, Wang D (2013) Analyzing noise robustness of MFCC and GFCC features in speaker identification. In 2013 IEEE international conference on acoustics, speech and signal processing 7204–7208. IEEE. https://doi.org/10.1109/ICASSP.2013.6639061
71. Zhu W, O'Shaughnessy D (2004) Incorporating frequency masking filtering in a standard MFCC feature extraction algorithm. In Proceedings 7th International Conference on Signal Processing, 2004. Proceedings. ICSP'04. 2004. IEEE. 1:617–620. https://doi.org/10.1109/ICOSP.2004.1452739

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.