sustainability

MDPI

Article

Prosodic Feature-Based Discriminatively Trained Low Resource Speech Recognition System

Taniya Hasija [1], Virender Kadyan [2], Kalpna Guleria [1,*], Abdullah Alharbi [3], Hashem Alyami [4] and Nitin Goyal [1]

[1] Chitkara University Institute of Engineering & Technology, Chitkara University, Rajpura 140401, Punjab, India; taniya@chitkara.edu.in (T.H.); dr.nitingoyal30@gmail.com (N.G.)
[2] Speech and Language Research Centre, School of Computer Science, University of Petroleum and Energy Studies, Dehradun 248007, India; vkadyan@ddn.upes.ac.in
[3] Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; amharbi@tu.edu.sa
[4] Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; hyami@tu.edu.sa
* Correspondence: kalpna@chitkara.edu.in

Abstract: Speech recognition has been an active field of research in the last few decades since it facilitates better human–computer interaction. Native language automatic speech recognition (ASR) systems are still underdeveloped. Punjabi ASR systems are in their infancy stage because most research has been conducted only on adult speech systems; however, less work has been performed on Punjabi children's ASR systems. This research aimed to build a prosodic feature-based automatic children speech recognition system using discriminative modeling techniques. The corpus of Punjabi children's speech has various runtime challenges, such as acoustic variations with varying speakers' ages. Efforts were made to implement out-domain data augmentation to overcome such issues using Tacotron-based text to a speech synthesizer. The prosodic features were extracted from Punjabi children's speech corpus, then particular prosodic features were coupled with Mel Frequency Cepstral Coefficient (MFCC) features before being submitted to an ASR framework. The system modeling process investigated various approaches, which included Maximum Mutual Information (MMI), Boosted Maximum Mutual Information (bMMI), and feature-based Maximum Mutual Information (fMMI). The out-domain data augmentation was performed to enhance the corpus. After that, prosodic features were also extracted from the extended corpus, and experiments were conducted on both individual and integrated prosodic-based acoustic features. It was observed that the fMMI technique exhibited 20% to 25% relative improvement in word error rate compared with MMI and bMMI techniques. Further, it was enhanced using an augmented dataset and hybrid front-end features (MFCC + POV + Fo + Voice quality) with a relative improvement of 13% compared with the earlier baseline system.

Keywords: children Punjabi ASR; discriminative techniques; feature extraction; prosodic features; data augmentation

Citation: Hasija, T.; Kadyan, V.; Guleria, K.; Alharbi, A.; Alyami, H.; Goyal, N. Prosodic Feature-Based Discriminatively Trained Low Resource Speech Recognition System. *Sustainability* **2022**, *14*, 614. https://doi.org/10.3390/su14020614

Academic Editor: Peter Schmidt

Received: 25 November 2021
Accepted: 4 January 2022
Published: 6 January 2022

1. Introduction

In the past, computers were primarily used for human–machine interaction, and communication was performed through keyboard and mouse. Automatic speech recognition (ASR) systems were developed to make them entirely usable for humans to communicate with computers through speech quickly [1]. Real-life applications of ASR systems can be found in Amazon Alexa and Apple's Siri [2]. A variety of techniques have been used to build ASR systems today. In the ASR systems, the accuracy of the words is dependent on a variety of factors. These attributes include the speaker's speech style, emotional state, age, male and female pitch, and the language's international and regional accent [3].

A low resource automatic speech recognition (ASR) system is beneficial for future sustainable native language interface development, emphasizing cultural and social sustainability. The research highlights, which include research questions (RQs) and the outcomes of the proposed work, are as mentioned below:

RQ1: How to develop an ASR system for a low resource database that makes efficient use of a native language-based system?

Outcome: This proposed work has contributed to building an ASR system for a native language (Punjabi) spoken in the northern region of India. The collected corpus helped in the development of a children's ASR system.

RQ2: How to reduce the data scarcity issue that occurs due to less availability of training data?

Outcome: The data scarcity problem was solved by using an artificial data augmentation approach to enrich the training data while preserving default test data. Synthesized speech was augmented with an actual speech corpus. The artificial corpus enhancement is a solution for better computing performance of native language interface systems.

RQ3: How to identify key parameters that generate robust features in developing a children's ASR system?

Outcome: This work employed MFCC feature extraction along with prosodic features to construct an efficient ASR system for children's Punjabi speech using a Mel filter bank. The experiments were performed to verify the performance of the hybrid prosodic features applied on discriminative training models for Punjabi children's speech to overcome the variability in children's speech.

The first step in developing an ASR method is feature extraction, also known as the front-end technique, followed by acoustic modeling, which involves classification using language models as feedback [4]. The fundamentals of ASR focus on the treatment of speech signals as stochastic patterns (after feature extraction) and stochastic pattern recognition techniques to generate hypothesis word sequences with the same probability as the input signal [5]. In this machine learning approach, such as the Hidden Markov Model (HMM) (based on pattern recognition), only reference classes are included in the training processes, and they are treated separately [6]. Despite these approaches, discriminative training methods include competing and reference classes. Class boundaries are more important to consider when optimizing classes. Due to the inherited groups' sequence composition, implementing discriminative training to enforce ASR is a challenging problem. Discriminative approaches boost consistency in execution and analytical comparison [7].

Along with discriminative techniques, improving the efficiency of the ASR systems, the front end is vital, as feature extraction is a necessary step that requires informative word parameters [8]. In ASR, various feature-extraction techniques are available [9]. Different speakers have different spoken utterance styles, or the language can be tonal where the tone on a syllable can change the meaning of the whole word [10]. Prosodic features are extracted in addition to Mel Frequency Cepstral Coefficient (MFCC) features to catch pitch and tone-based features. The hybrid prosody features are voice probability, pitch (F0 gradient), intensity (energy), and loudness. Researchers have constructed children's ASR systems to catch huge acoustic fluctuations. These prosodic features also capture psychological qualities, speaking style, and inter-speaker variances, whereas traditional feature-extraction algorithms capture phonetic components of a speech signal. Pitch features are essential in extracting the tonal aspects of a languages' characteristics [11].

While ASR systems for international languages are well established, establishing ASR systems for native languages remains a difficult challenge due to limited resources and a lack of corpora for these languages. Punjabi is a native language whose ASR system is in the development phase, and adult speech has been implemented for the Punjabi ASR system. No data are available for children's Punjabi speech. The collection of manual data is a time-consuming and challenging task. Hence, a competent approach for artificial enhancement of the Punjabi speech corpus should be employed [12]. In-domain augmentation occurs when the parameters of a speech corpus are altered using pitch modulation or time modulation

and a new speech corpus is integrated with an existing corpus. The out- domain data augmentation occurs when speech is generated using a different technique or method and merged with the existing corpus [13,14]. Tacotron is a text-to-speech (TTS) system that takes text as input and produces synthesized speech that seems natural. Tacotron can be used to create a new speech corpus, enabling us to develop an augmented corpus [15]. Tacotron is a single model that manages everything. Tacotron has four parts: the first is a front end for extracting linguistic parameters, the next is an acoustic prediction model, the part following is a duration model, and the last is a signal processing vocoder [16].

Our Contribution: Much research has been conducted on English, but only Tamil and Hindi ASR systems have been developed for national languages. Since the speech corpus is still in its infancy, native language ASR systems are underdeveloped. Although Punjabi is one of India's 22 national languages, and since the people of northern India speak Punjabi, there is a need for a Punjabi ASR system so that a more significant number of ASR applications can be available in Punjabi. In recent years, researchers have actively proposed adult ASR systems. The adult Punjabi speech corpus has been submitted to the research community. However, the children's speech data corpus is still in the early stages of development; implementing and improving the performance of children's ASR systems is a difficult task due to the variability in children's speech. This work has contributed to the collection of a Punjabi children's speech corpus to deploy Punjabi ASR for children. The collection of the speech corpus was carried out at various schools, which is a time-consuming process, and calculating utterances is a challenging task.

Further, the artificial corpus enhancement was performed using data augmentation techniques. This work employed MFCC feature extraction with prosodic features to construct an ASR system for children's Punjabi speech. The MFCC and prosodic features were obtained, and the retrieved features were subsequently subjected to discriminative techniques. In order to satisfy the scarcity of speech corpus, new speech was synthesized using Tacotron, which was augmented with the children's Punjabi speech corpus, and experiments were carried out by extracting prosodic features from the new augmented corpus.

The state-of-the-art on ASR and discriminative methodology is included in Section 2. The theoretical history of prosodic features and discriminative strategies is covered in Section 3. In Section 4, the experimental setup is defined. Section 5 describes the system overview, followed by the results, discussion, and comparison with the state-of-the-art in this domain in Section 6. Section 7 finally concludes.

2. Literature Review

In speech recognition, Dreyfus Graf of France represented the output of a six-band pass filter, and for determining transcription of the input signal, he traced the band filter output [17]. Later in 1952, the Bell laboratory of the USA constructed the first ASR system. The system recognized telephonic digits when spoken regularly [16]. In the 1960s, Japanese laboratories were fully active in speech recognition and constructed vowel recognition, phoneme recognizer, and digit recognizer systems [18]. The implementations of word recognition were pushed off the rails in the 1980s, and people started to emphasize machine learning algorithms. The Defense Advanced Research Project Agency (DARPA) financed the research on speech interpretation in the United States. Carnegie Mellon University (CMU) developed a speech recognition technology in 1973 that could identify 1011 vocabulary words in a dataset. Algorithms of numerous forms were formulated and applied, such as template-based pattern recognition or explicit pattern recognition and later statistical modeling architectures in the 1980s. HMM models were used to perform rigorous statistical simulations. HMM has a double stochastic process, which includes several stochastic processes; hence, the term hidden is used in HMM name. In the 1990s, the HMM technique became popular.

B.H. Juang and L.R. Rabiner [19] (1991) reviewed the statistical HMM model, and a consistent statistical framework was provided. The authors highlighted several aspects

of the general HMM approach. In contrast, it demands further consideration to improve results in various applications, such as modeling parameters, especially the issue of minimal classification error, integration of new features and prior linguistic awareness, modeling of state durations, and their usage in speech recognition. A continuous-based and speaker-independent ASR system was introduced on SPHINX by Kai-Fu Lee et al. [20]. The authors utilized the TRIM dataset provided by Texas Instruments, which utilized 80 teaching speakers and 40 research speakers, with 85 men and 35 women. The authors employed LPC function extraction and HMM acoustic simulation techniques. Word-based phone modeling and triphone modeling was performed. This work achieved an accuracy of 71%, 94%, and 96% on grammar, word pair grammar, and bigram grammar, which can be improved further. Xuedong Huang et al. [21] designed SPHINX-II to cope with speaker and environment heterogeneity. SPHINX-II extracted speaker-normalized features from the corpus along with dynamic features. Authors utilized between-word triphone models, semi-continuous HMM models, and senons, and the overall model achieved better accuracy than SPHINX.

In [22], feature extraction was performed in three stages: static feature extraction, normalization, and temporal information inclusion. The cepstral unconstrained monophony test revealed that MFCC outperformed PLP, cepstral mean subtraction. A comparative study of different feature-extraction techniques has been presented by Gupta and Gupta [23]. The authors presented MFCC, Relative Spectral (RASTA), and Linear Predictive Coding (LPC), where MFCC outperformed the others. After MFCC feature extraction, Wang et al. [24] used prosodic details and normalized feature parameters for tone- and pitch-related features to train a 3-layer feed-forward neural network and introduced the Parallel Phoneme Recognition followed by Language Modeling (PPRLM) system. The PPRLM system achieved an 86 percent classification rate. Furthermore, in [25], the authors used the Gaussian Mixture Model (GMM) for classification and utilized various levels of speech features, such as phonetic, acoustic, and prosody. The authors presented tonal and non-tonal classifiers, including pitch extraction, pitch trimming, pitch smoothening, pitch shifting, and pitch speed measuring. The work was focused on the data collection problems; Tacotron, a text-to-speech synthesizer, was deployed to reduce data scarcity. Wang et al. [15] proposed an end-to-end methodology that generated synthetic speech. The authors represented critical approaches for generic strategies, and the system obtained a 3.82 Mean Opinion Score (MOS) on a scale of 5. Skerry-Ryan et al. [26] introduced an expanded Tacotron with latent embedding space of prosody. The output produced by the Tacotron represented prosodic information such as pitch and loudness. Shen et al. [27] demonstrated a Tacotron system, a neural model. The authors used a recurrent network and predicted the Mel spectrogram of a given text. The system scored 4.53 MOS. Tacotron 2 was proposed by Yasuda et al. [28], and it outperformed traditional systems. Self-attention was added to Tacotron 2, which captured pitch-related dependencies and improved the audio quality. Later in 2021, Hasija et al. [14] presented the work on the Punjabi ASR system for children by extending the corpus of children's speech by pre synthesizing new speech using a Tacotron text-to-speech model. The original corpus was combined with pre-synthesized speech, and it was fed into the ASR system, which exhibited a RI of 9% to 12%.

2.1. Discriminative Techniques Based ASR Systems

Researchers have developed ways to help ASR systems perform better in the recent past. Povey and Woodland [29] researched discriminative techniques using a large vocabulary dataset. The authors defined and compared Lattice-based Maximum Mutual Information Estimation (MMIE) training to Frame Discrimination (FD). The effectiveness of MMIE and Maximum Likelihood Estimation (MLE) were also evaluated, and MMIE outperformed MLE. Povey and Woodland again [30] conducted a study on the Minimum Word Error (MWE) and Minimum Phone Error (MPE) criteria for discriminative HMM training after the publication of MMIE. Further, the authors used I-smoothing and per-

formed discriminative training. The Switchboard/Call Home telecommunications corpora were used in the experiments. The proposed method described a relative improvement of 4.8 percent.

Further, Povey et al. [31] proposed a new approach called feature MPE (fMPE), which applied various functions on a feature to train millions of parameters. The authors implemented fMPE process in various phases, such as generation of high-dimensional features, an extension of the acoustic context, projection of features, and training of the feature matrix. The authors proved that it is a unique method for training feature matrixes. In [31], the authors provided further improvements by releasing a new version of the MMI feature, which improved accuracy [32] and probability routes with higher phone error owing to proper transcription enhanced in lattices. Additionally, it led to I-smoothing, which replaced I-smoothing to the maximum probability estimate to the preceding iteration's frequency. These derived features were subjected to Vocal Tract Length Normalization (VTLN) and feature–space maximum likelihood linear regression (FMLLR), which improved the performance. The authors proved that the enhanced MMI approach produced more accurate results than the MPE technique.

2.2. Hybrid Front End Approach-Based Discriminative Techniques

In [33], McDermott et al. applied the benefits of discriminative approaches, utilized the Minimum Classification Error (MCE) discriminative technique on HMM models, and proved a reduction of 7% to 20% in the error rate. Later, Vesely et al. [34] developed frame-based cross-entropy and sequence discriminative MMI on DNN models [29]. It was proved that the system was improved by 8% to 9% compared with prior studies. In [35], Dua et al. reported their work on heterogeneous feature vectors, wherein the two feature-extraction approaches (MFCC and PLP) were hybridized, and signal features were retrieved using the MF-PLP methodology. MMIE and MPE methods were used to train acoustic models. The authors concluded that MF-PLP combined with MPE outperformed the other heterogeneous features and discriminative combinations. In [36], Dua et al. investigated Differential Evaluation (DE) on Gammatone Frequency Cepstral Coefficient (GFCC) features and used discriminative approaches on acoustic models of datasets. The outcomes of discriminative approaches in clean and noisy environments were compared using MFCC and GFCC features. The authors concluded that the DE-based GFCC feature-extraction method combined with MPE training methodology produced better results in clean and noisy situations. After successfully using these methods for ASR in the Hindi language, the researchers were inspired to study discriminative methods on the Punjabi corpus. In [37], Kaur and Kadyan presented their work on the Punjabi speech corpus and the implemented BMMI, fMMI, and fBMMI on a corpus, which resulted in a relative improvement of 26%. The work on discriminative methods for ASR systems is summarized in Table 1.

The research in this paper aimed at elevating the infancy of children's Punjabi speech corpus. A four-hour speech corpus was manually gathered, and artificial methods were used to enhance the corpus. Later, prosodic features were extracted to improve the system's performance, as MFCC features alone were insufficient for capturing the variety of variations in children's speech. Seven prosodic features were retrieved, and their impacts were investigated using discriminative techniques on a speech corpus. Integrated prosodic features were tested on a speech corpus afterward, results were analyzed, and performance was evaluated.

Table 1. Related work based on discriminative techniques for ASR system.

Author	Year	Data Set	Feature Extraction	Discriminative Technique	Results
Povey and Woodland [29]	2001	North American Business News (NAB) corpus, Telephone switchboard dataset	MFCC	MMIE, MLE	The authors propose the MMI approach. A 16.3 percent relative improvement (RI) in NAB and a 5.5 percent RI in switchboard, indicating that MMIE outperforms MLE.
Povey and Woodland [30]	2002	"HubS" from the Switchboard and Call Home English	MF-PLP	MPE and MWE	I-smoothing was used to make discriminative approaches more generic. With a RI of 4.8 percent, I-smoothened MPE surpassed MLE and MWE.
Povey et al. [31]	2005	Conversational telephone speech, Broadcast News, Call center and Malach corpus	PLP and MFCC	fMPE, MPE	fMPE was introduced, which indicated that MPE objective functions were also applied to the feature. When employing fMPE instead of MPE, there was a 6.5 percent reduction in RI.
McDermott et al. [33]	2007	Corpus of Spontaneous Japanese (CSJ) dataset was used. 186k utterances in training data set of about 230 hours. 130 min speech in the testing dataset.	MFFC, delta, and delta-delta feature extraction	Minimum Classification Error (MCE)	The author presented an MCE framework for discriminative training, which improved the HMM's performance. The ASR system performed better and showed a 7% to 20% relative reduction in the word error rate.
Povey et al. [32]	2008	Arabic Broadcast news corpus, conversational telephone news, English broadcast news, TC-STAR corpuses	Vocal tract length normalization (VTLN) + feature-space maximum likelihood linear regression (FMLLR)	MMI, BMMI, fMMI, fBMMI	The MMI function was changed, resulting in boosted MMI function and feature space MMI. In addition, I—smoothening was used. When compared with MPE, the system revealed a RI of 0.5% to 0.7%. When compared with MMI, fMMI outperformed it.
Vesely et al. [34]	2013	Switchboard-1 Release 2 (LDC97S62) training dataset of 300 hours	MFCC, LDA + STC (semi-tied covariance), FMLLR	Frame-based cross-entropy, MMI sequence-discriminative training of DNNs	The authors proposed a DNN–HMM hybrid system that uses a sequence discriminative method and frame-based entropy and achieved an 8% to 9% relative improvement.
Dua et al. [35]	2017	Hindi speech corpus	MFCC, PLP, MF-PLP	MMI and MPE	The RI of 25.9% was obtained using the MF-PLP feature-extraction approach and the MPE discriminative methodology.
Dua et al. [36]	2018	Hindi speech corpus	GFCC	MMI, MLE, and MPE	The accuracy rate of MPE with DE optimized GFCC features was 86.9% in a clean environment and 86.2% in a noisy environment, according to a differential equation (DE) optimization on GFCC features.
Kaur and Kadyan [37]	2020	Children Punjabi speech corpus	MFCC	BMMI, fMMI, and fBMMI	With a RI of 26% from baseline results, the fBMMI discriminative approach outperformed both BMMI and fMMI.

3. Theoretical Background

3.1. Prosodic Features

The MFCC features alone are not enough to extract all the informatics parameters from the input signal. The performance of the features can be affected by speaking variability and accent tone. Prosodic features were paired with MFCC features to address this problem, resulting in increased system robustness and ASR device accuracy. Extraction of robust and extra feature details at the syllable level helped determine the tonality of a given syllable. These features were long-term characteristics of utterances that aided in presenting various context-related details about that utterance. Various prosodic features have been extracted in the past to address an ASR system's low efficiency. The prosodic features are F0, voicing probability, intensity, loudness, voice quality, harmonic-to-noise ratio (HNR), F0 raw, and F0 envelope [11]. An autocorrelation technique extracts pitch predictions from the input speech signal [38]. Fundamental frequency (F0) is the cue of the pitch. It is possible by calculating the similarity of two corresponding waveforms. The uniformity of waveforms is determined by comparing them at various time intervals. The F0 raw is captured using the autocorrelation function. The autocorrelation function for infinite discrete function x[n] is computed as shown in Equation (1):

$$R_x(v) = \sum_{n=-\infty}^{\infty} x[n]x[n+v] \tag{1}$$

The autocorrelation function of the finite discrete function x′[n] of size N is derived as per Equation (2):

$$R_{x'}(v) = \sum_{n=0}^{N-1-v} x'[n]x'[n+v] \tag{2}$$

The following is the cross-correlation function between the x[n] and y[n] functions as shown in Equation (3):

$$R_{xy}(v) = \sum_{n=-\infty}^{\infty} x[n]y[n+v] \tag{3}$$

Then, on F0 raw, pitch trimming and smoothing methods are applied, and F0 is obtained. Later, the pitch means subtraction algorithm is added to the F0 function, and the probability of voicing is captured, as it indicates the percentage of the signal's unvoiced and voiced data [11]. Let the normalized cross-correlation function value be 'a', which must be absolute on a particular frame. The POV is calculated as follows in Equation (4):

$$L = -5.2 + 5.4\exp(7.5(a-1)) + 4.8a - 2\exp(-10a) + 4.2\exp(20(a-1)) \tag{4}$$

where L is approximation of log-likelihood ratio log(p(voiced)/p(unvoiced)). The approximation value of p is calculated in that frame as mentioned in Equation (5):

$$p = \frac{1}{(1+\exp(-L))} \tag{5}$$

The jitter and shimmer algorithm is used to capture the voice quality [39]. The essential property of tonal words is lexical stress, also known as intensity. Stressed words are more energetic, have more prominent F0 movements, and have long durations. The F0 contour reflects the amplitude of a change in log energy in the voiced area of a syllable. Boundary identification is a method for calculating stress [40].

3.2. Discriminative Training Technique

The main objective of the discriminative technique is to reduce the mismatch between incorrect and correct word sequences in testing and training modeling [41]. The discriminative techniques are discussed below:

3.2.1. Maximum Mutual Information (MMI)

Until the language model is tweaked or updated, conditional machine learning is identical to MMI. MMI was first used in an isolated word recognition system, and after proving its worth in isolated word ASR, it was successfully applied to continuous speech recognition. The MMI criterion is used to optimize the posterior for the spoken words in the decoding process. The MMI criterion for each sentence is shown in Equation (6):

$$\mathrm{MMI}(\theta;S) = \sum_{m=1}^{M} \log \frac{p\left(O^m|S^m\right)^k P(w^m)}{\sum_w p\left(O^m|S^w;\theta\right)^k P(w)} \quad (6)$$

The observation sequence is represented by O^m, while the accurate word transcription of the m-th utterance is represented by w^m. M is the total number of words in the utterance's transcription. θ is a deep neural networks model parameter that incorporates biases and weight matrices. S^m represents the sequence of states about w^m. The letter k denotes the acoustic scaling factor. The total in the denominator should theoretically be calculated across all possible word sequences [42]. When MMI is used, there is a reduction in Bayes risk decoding and a boost in system efficiency. The acoustic and language models in MMI are scaled such that the class posterior is smoothed [7].

3.2.2. Boosted MMI (BMMI)

A boosting parameter is transferred to the function in BMMI. This boosting parameter produces more confusing results by increasing the probability of sentences with more errors. Boosted MMI can be thought of as an attempt to impose a soft margin proportional to the number of errors in a hypothesized statement. When using the forward–backward algorithm on the denominator lattice, BMMI needs a little more computation than MMI. The BMMI is calculated as shown in Equation (7):

$$\mathrm{BMMI}(\theta;S) = \sum_{m=1}^{M} \log \frac{p\left(O^m|S^m\right)^k P(w^m)}{\sum_w p\left(O^m|S^w\right)^k P(w)^{e^{-bA(w,w^m)}}} \quad (7)$$

where b is the boosting factor having 0.5 value, and $A(w, w^m)$ measures accuracy between word sequence w and w^m. The sole difference between MMI and BMMI is adding a boosting component to the BMMI equation's denominator [42].

3.2.3. Feature Space MMI (fMMI)

fMMI gives a significant improvement in the field of other discriminative techniques. A single modification is performed in the function of MMI to change it into an fMMI function. In fMMI learning, the rate is reduced to 0.01 or 0.015 to compensate for the lower range of the MMI objective function [32]. If the boosting parameter is passed to the fMMI function, the term used is fBMMI.

4. Experimental Setup

This research was carried out to see how effective prosodic features are on acoustic modeling discriminative techniques. The speech corpus of Punjabi children is still in its infancy stage. A concerted effort was made to set a Punjabi children speech corpus, and the whole study was based on this corpus. The recordings were made in classrooms with microphones tuned to a frequency of 16 kHz. The age range was 7 to 13 years old. Following the collection of audios, the audios were segmented to obtain utterances aligned with suitable meaning. In the segmentation procedure, the Praat toolset was employed. The utterances were then transcribed for the language model. The total number of utterances was 2370. A total of 1885 utterances were used in the training data set and 485 utterances in the testing data set. The details of children's Punjabi corpora are explained in Table 2.

Table 2. Description of Children Punjabi Speech Corpora.

Term	Train	Test
No. of Speakers	39	6
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

The procedure was carried out with the Kaldi toolkit [43]. In the training data set, there were 39 speakers, while in the testing data set, there were six speakers. The Word Error Rate (WER) evaluated the ASR system. The substituted words (S), the elimination of words (D), or the addition of new words (I) were the error words. WER was calculated as mentioned in Equation (8):

$$\text{WER\%} = \frac{\text{S} + \text{I} + \text{D}}{\text{N}} \times 100 \tag{8}$$

where S is the substitution of words, I is the number of inserted words, D is deleted words, and N is the total number of words in the dataset [44]. Relative Improvement (RI) is another performance evaluation metric. The absolute rise related to a new value (N) in comparison with an old value (O) is called RI, as shown in Equation (9):

$$\text{RI\%} = \frac{\text{O} - \text{N}}{\text{O}} \times 100 \tag{9}$$

5. System Overview

In the proposed system initially, the children's speech corpus was fed into the feature-extraction module, which extracted MFCC features from the input signal. Using the MFCC technique, each frame's energy parameters were collected, yielding 13 function coefficients. Each frame took 25 milliseconds to complete, with a 10-millisecond frameshift. Frame extraction used a Hamming window and a 23-channel Mel filter bank. After that, function coefficients were calculated using logarithm and DCT. MFCC features defined the instantaneous and spectral envelope shapes of a speech signal. Since MFCC features were insufficient, prosodic features were extracted as well. OpenSmile Toolkit and MATLAB were used to extract prosodic features. The extracted prosodic features were the probability of voicing (POV) (P1), F0 (P2), intensity (P3), loudness (P4), voice quality (P5), F0 raw (P6), and F0 envelope (P7). The extracted prosodic features were merged with MFCC features one by one and in combinations. These were fed to the ASR system, which employed discriminative techniques. MMI, BMMI, and FMMI are some of the discriminative methods used. Monophone (HMM) modeling was achieved when features were initially fed into the ASR framework. In Monophone modeling, acoustic vectors were incoming sequences; a word sequence W had to be found using probability P(W | A) and had to include constraints imposed by grammar. The Bayesian theorem was applied, and probability was calculated as shown in Equation (10):

$$\text{P}(\text{W}|\text{A}) = \frac{\text{P}(\text{A}|\text{W}) \cdot \text{P}(\text{W})}{\text{P}(\text{A})} \tag{10}$$

The probability of a sequence of acoustic vector A to the given word sequence (referred to as an acoustic model trained on training input data) is represented by P(A | W). The second P(W) is the probability of word sequence given by the language model (LM). The language model is a text corpus that includes 1000 or 10,000 words. P(W | A) is a pattern recognition approach, and a number of techniques were used to compute the acoustic model of given input training data [6]. Dynamic features of speech include MFCC feature trajectories over time. These trajectories were then estimated and combined with MFCC and prosodic coefficients to improve ASR results. Delta features were characteristics of

trajectory paths. Following Monophone, delta features were utilized to achieve triphone modeling (tri 1). Delta features were computed as shown in Equation (11):

$$d_t = \frac{\sum_{n=1}^{N} n(c_t - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (11)$$

where d_t stands for delta coefficients computed on frame t, c_t and c_{t-n} for static coefficients, and N equals 2. Triphone simulation is achieved again for delta features, this time with delta-delta features, which are the time variant component of delta features (tri 2) and formula of computation, as mentioned in Equation (12):

$$dd_t = \frac{\sum_{n=1}^{N} n(d_t - d_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (12)$$

Delta and delta-delta features were extracted, also known as the first and second derivatives of the speech signal. Following that, triphone simulation using the Maximum Likely Linear Transform (MLLT) and Linear Discriminative Analysis (LDA) was performed (tri 3). Linear Discriminative Analysis (LDA) was applied to the output of tri 2 to transform smaller volumes of acoustically distinct units, reducing the coefficient to a manageable 40 dimensions. Following the likelihood estimation, a new set of inputs was assigned to the new class, and the output class with the highest probability was chosen. The scatter matrix is computed in LDA as shown in Equation (13):

$$S = \frac{1}{n}\sum_{i} n_i(m_i - m)(m_i - m)^T \quad (13)$$

where sample class is represented by n_i, m_i represents the mean of ith class, m is the global mean, and T is the transpose of $(m_i - m)$ [38,45]. The next step was to calculate the Maximum Likelihood Linear Transformation (MLLT), which was computed over utterances and excluded speaker-specific information. The system's tri 3 modeling is LDA + MLLT [46]. The tri 3 output was then used to evaluate MMI, BMMI, and fMMI. The dataset was trained using the MMI function, and then decoding was performed for two rounds. The BMMI implementation was achieved by adding a 0.5 boosting factor to the MMI. The learning rate factor was given a value of 0.0025, and the boosting factor was given a value of 0.1 during the implementation of fMMI. Discriminatively trained modules were then sent for decoding. Decoding is the process of identifying recorded test samples based on auditory features of words. For precise voice recognition, it uses training, acoustic, and language models. It decodes the feature vectors, same as the training module, to determine the most likely word sequence. Figure 1 shows a block diagram of the implementation of discriminative techniques on the Punjabi children's speech corpus using prosodic features.

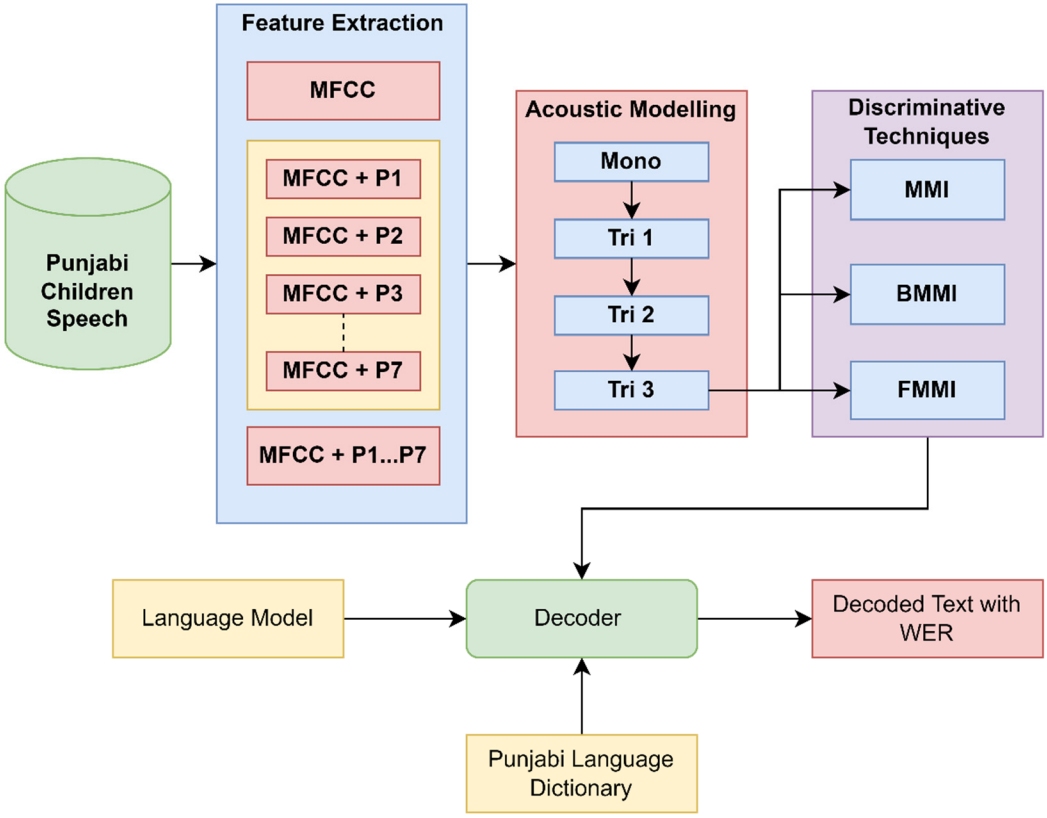

Figure 1. Block diagram of an implementation of discriminative techniques on the Punjabi children's speech corpus using prosodic features.

The step-by-step procedure for the implementation of prosodic features on the discriminative technique is explained below, and the illustration of this methodology is shown in Figure 2.

Step 1: Collection of original children's Punjabi speech data (male/female of age group 7 to 13 years) corpus.

Step 2: Initialize: Segmentation and transcription of audios. training_data = 1885 utterances from 2370 utterances testing_data = 485 utterances from 2370 utterances

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

A mfcc(training_data) and mfcc(testing_data)

The Mel filter bank can process speech signals with linear or nonlinear distributions at various frequencies.

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

Additionally, f(t,i) (Fast Fourier Transformation) is computed as:

$$F(t,i) = \left| \frac{1}{N \sum_{k=1}^{N-1} \left(e^{-\frac{2\pi jkn}{N}}\right) f_k} \right| (S'(n)) \quad (14)$$

where i 0, 1, 2, 3, . . . , $(N/2) - 1$.

B prosody(training_data)and prosody(testing_data) Applying prosodic feature extraction on the utterances of training and testing data set:

i To calculate the cue of F0 feature, autocorrelation function was used in the discrete function x[n] of speech signal on each 25 ms frame.

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

where $R_x(v_i)$ is the feature value of F0 raw at i frame of given speech utterance.

ii Cross-correlation function of two consecutive discrete function x[n] and y[n] of speech signal at each 25 ms frame is F0.

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

$R_x(v_i)$ is the feature value of fundamental frequency at i frame of given speech utterance.

iii The POV is calculated from the pitch mean subtraction formula.

$$L = -5.2 + 5.4\exp(7.5(a-1)) + 4.8a - 2\exp(-10a) + 4.2\exp(20(a-1))$$

where 'a' is the F0 value of the frame. Approximation value of POV is:

$$p = \frac{1}{(1+\exp(-L))}$$

Step 4: The extracted prosodic features are in matrix form and stored in a .xls file. There is one .xls file for each utterance, and features are extracted at every 25 ms frame. The MFCC features are combined with Prosodic features to form a single matrix using MATLAB. Later on, this single matrix is converted in .htk format for the Kaldi toolkit to proceed further.

```
[ml,f]=MFCC(ado,fs,'z0Mp',12,23,20e-3*fs,10e-3*fs, 0,0.5,0.97);
//MFCC feature Matrix

PROm=dlmread(pro_f_name);  // prosodic feature matrix where pro_f_name
is prosody file name extracted using OpenSmile toolkit

MFPro=[ml Pros];     //New matrix having all features of MFCC and prosody

writehtk(output,MFPro,0.010,8198);     // writing the file in .htk format
which is kaldi supportive.
```

Step 5: Conduct monophone training (mono) and align monophone results using kaldi toolkit.

```
steps/train_mono.sh --nj $num_jobs --cmd $train_decode_cmd $taining_
directory $language_directory exp/mono
where $num_jobs is the number of jobs of training data set,
$train_decode_cmd is run.pl file
$taining_directory is the training where the training utterances
and training transcription is stored
$language_directory is the language model directory
exp/mono is the directory where the training model and the results
of recognition are saved.
```

Step 6: Conduct delta training (tri 1) and align their phones.

```
steps/train_deltas.sh --cmd $train_decode_cmd 600 7000 $taining_directory
$language_directory exp/mono_ali exp/tri1
```

```
where 600 is the number of senons and 7000 is the number of leaves used
by train_delta.sh file for tri 1 modeling.
```

Step 7: Perform delta + delta training (tri 2) and also align their triphones.

```
steps/train_deltas.sh --cmd $train_decode_cmd 500 5000
$taining_directory $language_directory exp/tri1_ali exp/tri2
```

Step 8: Training of LDA + MLLT training (tri 3) on tri 2 output and aligning of their phones.

```
steps/train_lda_mllt.sh --cmd $train_decode_cmd 600 8000 $taining_directory
$language_directory exp/tri2_ali/ exp/tri3steps/align_fmllr.sh --nj
"$train_nj" --cmd "$train_decode_cmd" $taining_directory $language_
directory exp/tri3 exp/tri3_ali || exit 1
```

Step 9: Perform MMI training on tri 3 output.

```
steps/train_mmi.sh $taining_directory $language_directory exp/tri3_ali
exp/tri3_denlats exp/tri3_mmi
```

Step 10: Perform BMMI training on tri 3.

```
steps/train_mmi.sh --boost 0.5 $taining_directory $language_directory r
exp/tri3_ali exp/tri3_denlats exp/tri3_bmmi_0.5
```

Step 11: Conduct fMMI training on tri 3.

```
steps/train_mmi_fmmi.sh --learning-rate 0.0025 --boost 0.1 --cmd $train_
decode_cmd $taining_directory $language_directory exp/tri3_ali exp/dubm3b
exp/tri3_denlats exp/tri3b_fmmi_b
```

Step 12: Repeat steps 4 to 11 for each prosodic feature.
Step 13: Finally, performance is analyzed after comparing the results of three discriminative techniques on the number of prosody features combined with MFCC.

Further, the speech corpus was enhanced by artificially expanding the training dataset to improve the performance of both systems. The Punjabi children corpus was subjected to out-domain data augmentation. The ASR system received pre-synthesized speech from Tacotron, which was augmented with children's speech. After augmenting, a speech corpus of 2032 utterances was produced. Prosodic features and MFCC features were also retrieved from the new enhanced corpus, and experiments were repeated for individual prosodic and integrated prosodic performance analysis. The scarcity of training data was removed by using an expanded corpus. This led to considerable performance improvement, as evidenced by experimental results. Furthermore, increasing the training complexity increased the training time while maintaining the original processing time.

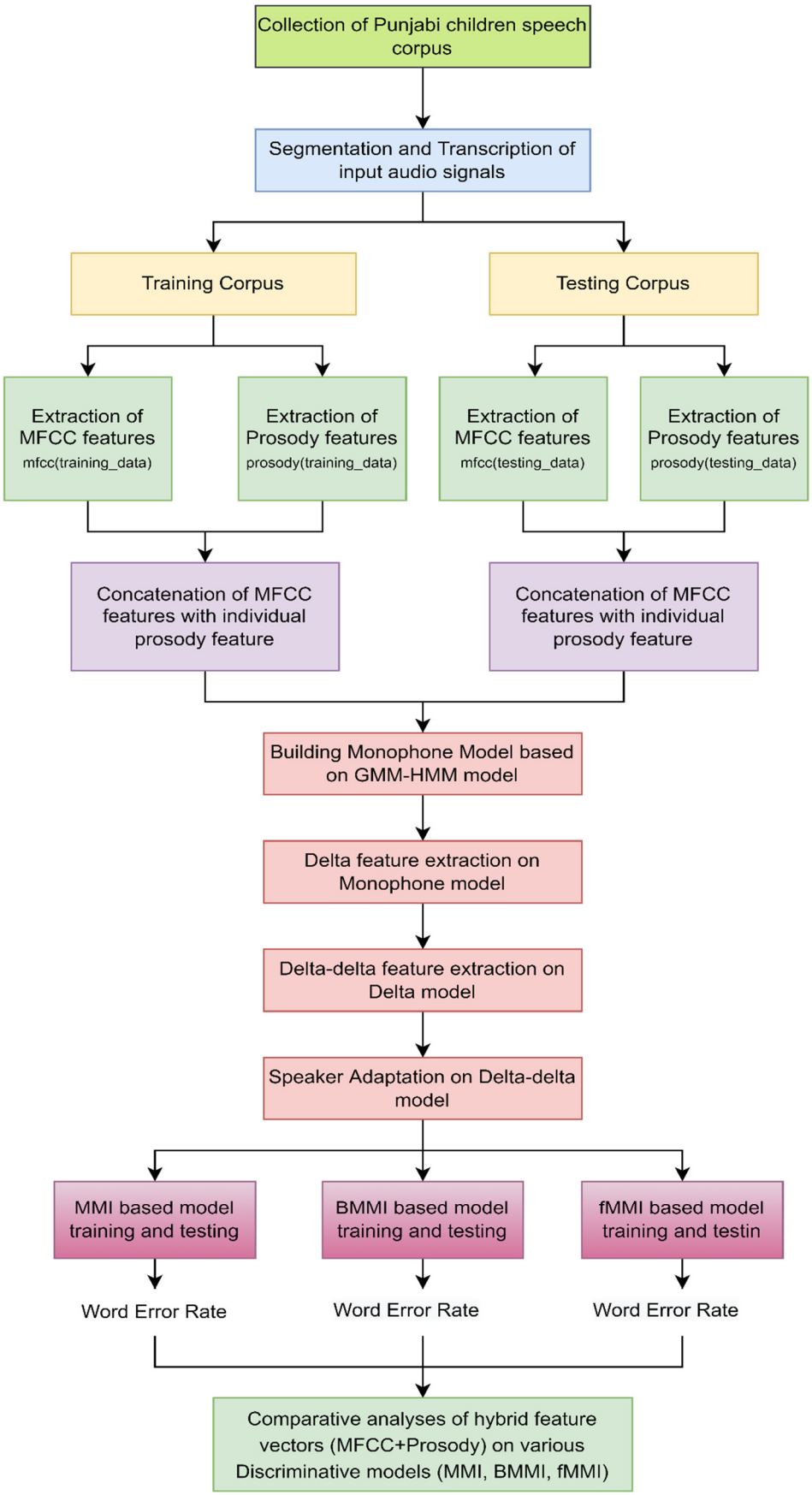

Figure 2. Methodology to implement prosodic features on discriminative techniques on Punjabi children's speech ASR.

6. Experimental Results and Discussion

Experiments were carried out to see how prosodic features fared on discriminative approaches. The experiments were carried out in four phases. In the first phase, only a particular prosodic feature was concatenated with MFCC features of the Punjabi children's corpus. In the second phase, the combinations of prosodic features with concatenation of MFCC was given to the discriminative methods of ASR systems. In third phase, data were augmented with original speech using the out-domain technique, and features were extracted from augmented data. The process was repeated for individual prosodic feature combinations with MFCC; as fMMI was performing well at this phase, only the fMMI discriminative technique was used, and results were compared. In the last phase, all or some prosodic features computed from the data-augmented corpus were integrated with MFCC and fMMI.

6.1. Performance Analysis of Discriminative Techniques on the Individual Concatenation of Prosodic Feature with MFCC of Punjabi Children's ASR

The speech corpus was provided for MFCC feature extraction and then sent to the ASR system for discriminative approaches. Baseline results referred to the MFCC feature results, and performance analysis of prosodic features was conducted by comparing baseline results. Prosodic features were retrieved and merged with MFCC features before being sent into the ASR system for modeling, as shown in Figure 2. The results of MMI, BMMI, and fMMI techniques on a children's Punjabi corpus with MFCC features and prosodic features are shown in Table 3.

Table 3. Result of Mono, Tri 1, Tri 2, Tri 3, MMI, BMMI, and fMMI on Children's Punjabi speech corpus with MFCC and individual prosodic feature.

Feature Used	Mono	Tri 1	Tri 2	Tri 3	MMI	BMMI	fMMI
MFCC	20.58	18.62	17.9	15.02	17.94	17.45	14.53
MFCC + POV	20.16	18.45	16.88	14.74	14.98	16.53	14.25
MFCC + F0	20.37	18.92	18.33	15.34	18.4	17.69	14.53
MFCC + intensity	18.36	17.13	17.41	15.05	20.47	17.94	14.32
MFCC + loudness	19	17.13	17.41	15.05	15.05	17.69	14.69
MFCC + voice quality	19.42	17.06	17.24	15.57	19.63	20.8	14.64
MFCC + F0 raw	19.3	16.78	17.38	14.84	18.71	21.46	15.44
MFCC + F0 envelop	24.09	20.09	20	16.9	16.14	20.08	15.06

In Table 3, the first row represents the baseline results, while the subsequent rows reflect the outcomes of prosodic features. fMMI beats MMI and BMMI. Compared with MMI and BMMI results, fMMI results had a RI of 20% to 25%. Figure 3 illustrates a bar graph of particular prosodic features with regard to their WER when the MMI, BMMI, and fMMI methods were employed for speech recognition. The POV feature outperformed other features in all modeling techniques. It is recommended that the POV feature be included along with MFCC to improve the robustness of ASR systems since POV shows the voicing component of the spoken data, which increases the chances of recognition.

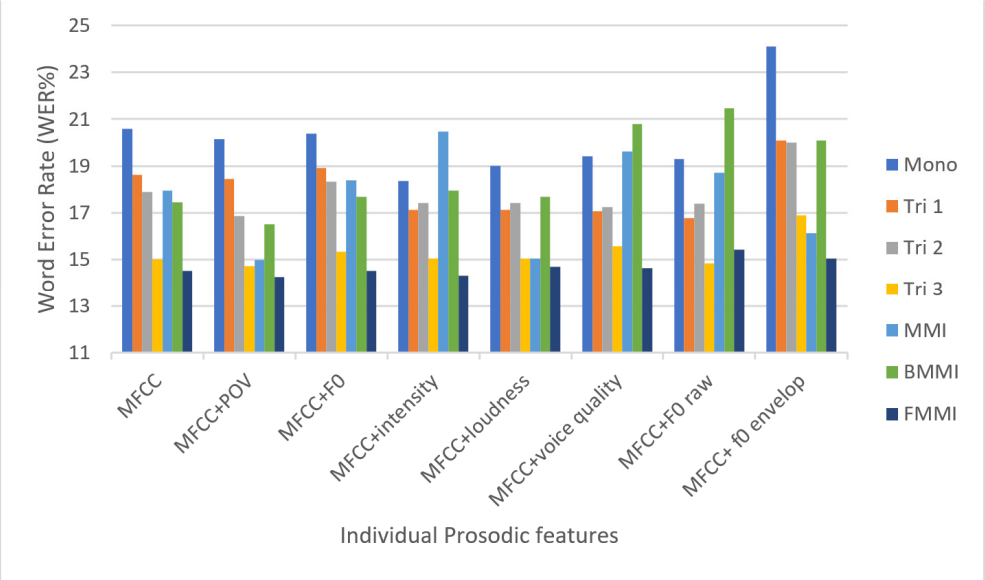

Figure 3. WER of an individual prosodic feature on MMI, BMMI, fMMI discriminative techniques.

6.2. Performance Analysis of Discriminative Techniques on the Integration of Prosodic Features with a Combination of MFCC of Punjabi Children's ASR

Following the particular prosodic feature, the performance was evaluated by integrating some or all of the features. At first, all features were integrated and then merged with MFCC features. These integrated features were supplied to the recognition system, assessing the outcome. After that, the best three features were integrated and concatenated with MFCC, and recognition was completed. The result of MMI, BMMI, and fMMI having integrated prosodic features merged with MFCC again is represented in Table 4.

Table 4. Result of Mono, Tri 1, Tri 2, Tri 3, MMI, BMMI, and fMMI having MFCC and a Combination of prosodic features.

Feature Used	Mono	Tri 1	Tri 2	Tri 3	MMI	BMMI	fMMI
MFCC + POV + F0 + voice quality + intensity + loudness + F0 raw + F0 envelop	19.66	18.12	16.99	15.16	19.21	19.24	14.6
MFCC + POV + F0 + voice quality	23.04	19.45	19.28	14.7	19.35	19.38	13.95

At fMMI, the combination of all prosody and MFCC had a WER of 14.6 percent, whereas the combination of three prosodic features and MFCC had a WER of 13.95 percent. The graph of integrated prosodic features coupled with MFCC is shown in Figure 4. Integration was the concatenation of all prosodic and MFCC features into a single matrix, and the WER was reduced. The performance improved when integration was carried out with the primary prosodic features of POV, F0, and voice quality. Combining POV, F0, and voice quality produced better results than combining all seven prosodic features, as seen in Figure 4. A 4–5% RI was obtained by combining POV, F0, and voice quality features with MFCC features rather than only MFCC features.

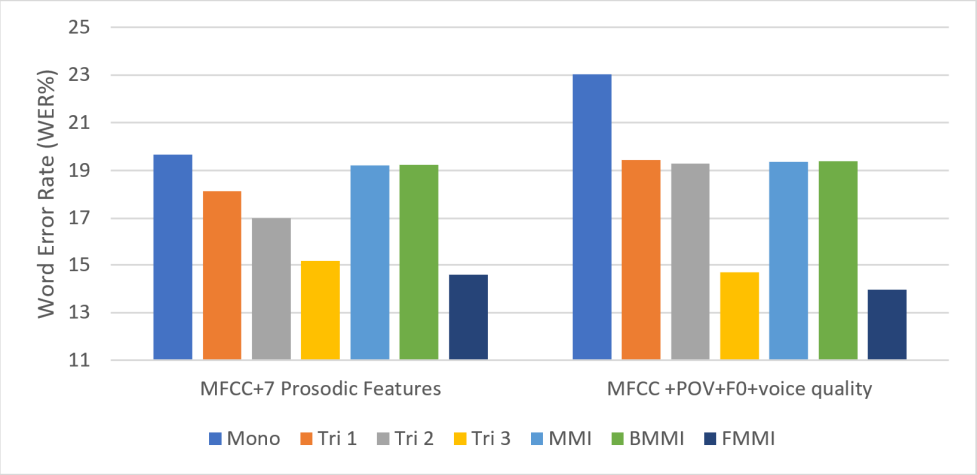

Figure 4. WER of integration of several prosodic features when MMI, BMMI, and fMMI discriminative technique was used.

6.3. Performance Analysis of Out-Domain Data-Augmented Punjabi Children's ASR Implementing Discriminative Techniques on Individual Prosodic Features Combined with MFCC

To overcome the issue of data scarcity, an artificial approach of data augmentation was used to enrich the training data while preserving the default test data. Synthesized speech was augmented with the actual speech corpus. After Tacotron synthesis, prosodic and MFCC features were retrieved, and a new matrix was formed by merging prosody and MFCC matrixes. The fMMI was then given the newly merged feature matrix. Only fMMI was used in this section since it performed well in Sections 6.1 and 6.2, and the results are presented in Table 5.

Table 5. Data-Augmented Result of Mono, Tri 1, Tri 2, Tri 3, and fMMI having MFCC and Combination of prosodic features.

Feature Used	Mono	Tri 1	Tri 2	Tri 3	fMMI
MFCC	17.76	16.88	17.04	13.86	13.33
MFCC + POV	18.65	16.45	15.88	14.74	12.80
MFCC + F0	18.68	16.25	15.62	14.49	12.87
MFCC + intensity	18.36	16.55	16.05	14.75	13.25
MFCC + loudness	18.96	17.15	16.78	15.05	13.15
MFCC + voice quality	18.40	16.95	16.54	14.58	13.25
MFCC + F0 raw	17.99	16.95	16.56	15.86	15.17
MFCC + F0 envelop	18.15	17.89	17.59	16.15	15.04

The first row of Table 5 shows 13.33% WER after augmentation on solely MFCC features, with a RI of 8% compared with the original dataset, exhibiting good accuracy. Later, enhanced data were processed for prosodic features to improve performance, and the results were outperforming those using fMMI, as shown in Figure 5.

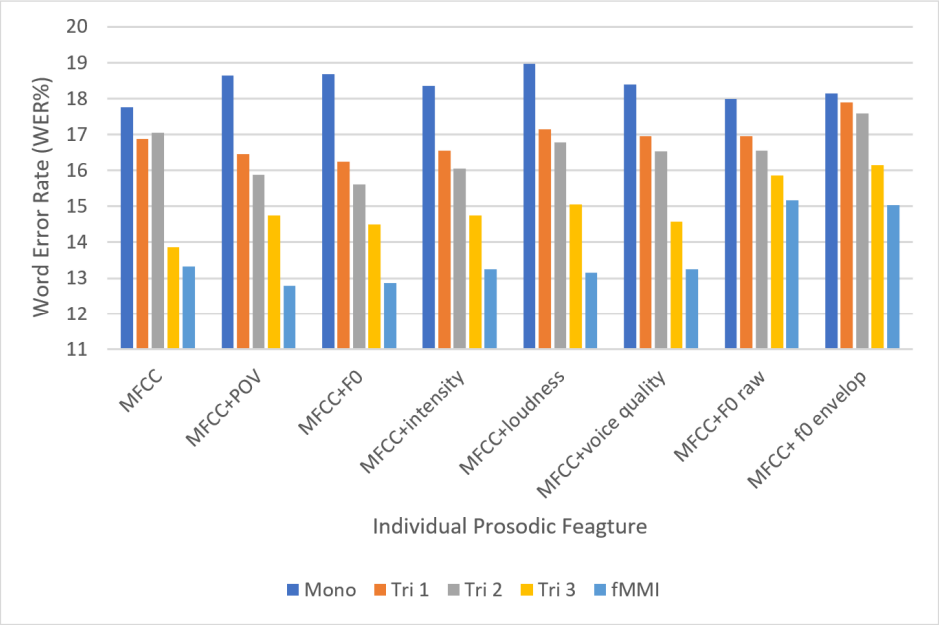

Figure 5. Data-augmented corpus WER of individual prosodic features using mono, tri 1, tri 2, and tri 3 modeling and fMMI discriminative techniques.

6.4. Performance Analysis of Out-Domain Data-Augmented Punjabi Children's ASR Implementing Discriminative Techniques on Integrated Prosodic Features Combined with MFCC

After implementing one prosodic combination, integrated prosodic features were implemented on fMMI, and all computations were performed on the augmented data corpus. The first row in Table 6 shows a WER of 13.85% for total prosodic feature integration. In comparison, the second row shows a WER of 12.61% for integrating significant features such as POV, F0, and voice quality, resulting in a RI of 13% more and outperforming other tests. The graphical representation of the result is shown in Figure 6, where the fMMI of MFCC + POV + F0 + voice quality shows the reduced WER.

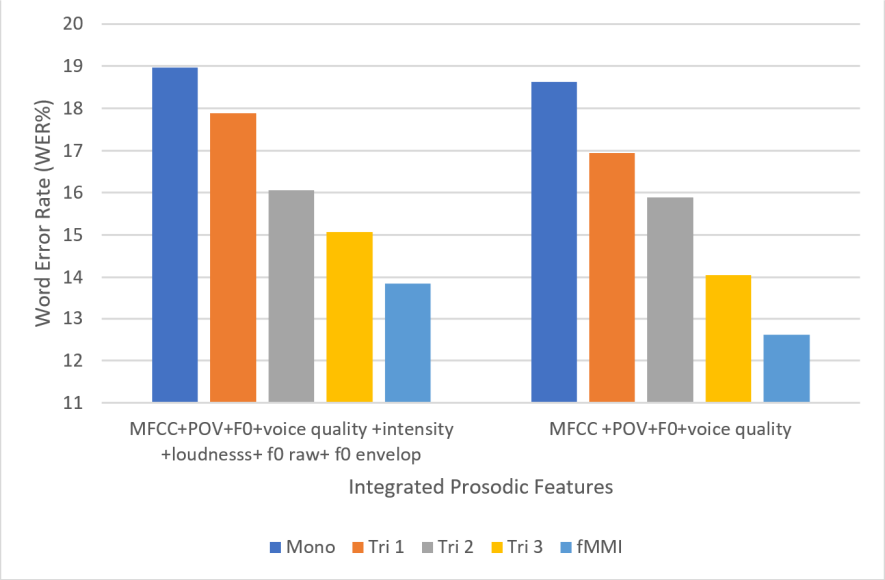

Figure 6. Data-augmented corpus WER of integrated prosodic features using mono, tri 1, tri 2, and tri 3 modeling and fMMI discriminative techniques.

Table 6. Data-augmented Result of Mono, Tri 1, Tri 2, Tri 3, and fMMI having MFCC and integrated prosodic features.

Feature Used	Mono	Tri 1	Tri 2	Tri 3	fMMI
MFCC + POV + F0 + voice quality + intensity + loudness + F0 raw + F0 envelope	18.97	17.88	16.06	15.07	13.85
MFCC + POV + F0 + voice quality	18.63	16.94	15.89	14.05	12.61

6.5. Comparative Analysis with Earlier Research Works on Punjabi Speech Recognition

The majority of ASR research has been focused on international languages. It is challenging to develop ASR systems for languages with limited resources. The Punjabi ASR system has received little attention, where the accumulation of training data has progressed from being isolated to continuous and then to spontaneous. Adult speech was used in the study of Punjabi speech, and much effort was undertaken to obtain high performance. Children's speech corpora are in their infancy, and we have gathered a children's Punjabi speech corpora for this research. Prosodic features were extracted from the data set, and discriminative methods were used. Table 7 shows the comparative result of the existing state-of-the-art on Punjabi speech ASR, which includes adult speech of isolated words and continuous sentences and very few works on children's Punjabi speech.

Table 7. Comparative analysis with existing state-of-the-art.

Sr. No.	Ref.	Data Set	Feature-Extraction Technique	Acoustic Modeling Technique	Performance
1.	(2012) [47]	2760 Distinct words (training dataset) (isolated data set)	MFCC	HMM (HTK toolkit was used)	Classroom environment (WER 4.37%, and accuracy was 95.63%), In open environment (5.92%, and accuracy was 94%)
2.	(2017) [48]	45,000 utterances (15 males, 10 males) (isolated data set)	MFCC, PLP, RASTA-PLP	HMM + GA (Genetic Algorithm), HMM + DE (Differential Evolution)	MFCC word accuracy was 67.38%. PLP word accuracy was 61.17%, and RATA-PLP word accuracy was 58.67%.
3.	(2017) [49]	58,700 utterances (training dataset) 6100 (test dataset) (isolated data set)	MFCC	HMM, HMM + GA (Genetic Algorithm), HMM + DE (Differential Evolution)	The system was tested on different real environment noises. RI of 3–4% (86% accuracy) using DE+HMM technique and 2–3% (83% accuracy) RI with GA + HMM as compared with HMM. (81% accuracy)
4.	(2018) [50]	3611 in training (6 male and 7 female), 422 sentences in test (phonetically rich sentences)	MFCC and GFCC	GMM + HMM and DNN + HMM	Using MFCC feature extraction, WER for DNN + HMM was 5.22%, and for GMM + HMM was 7.01%. Using GFCC feature extraction, WER for DNN + HMM was 24.67% and for GMM + HMM was 34.4%.
5.	(2020) [46]	1887 utterances in training, 485 in test (continuous children's speech)	MFCC	DNN + HMM	Hidden Layer 4 was performing well. Obtained WER was 14.46%.
6.	Proposed Work	1887 utterances in training, 485 in test (continuous children's speech)	MFCC + Prosody features	Discriminative Techniques (MMI, BMMI, and fMMI)	RI of 20–25% was observed using the fMMI discriminative technique. A combination of POV + F0 + voice quality prosody features with MFCC extracted on data-augmented corpus showed RI of 13% more than MFCC features alone.

7. Conclusions

This paper presents a prosodic feature-based automatic Punjabi speech recognition system for children that was experimentally evaluated using discriminative approaches. In this context, an effort was made to overcome the zero resource issue for children's Punjabi ASR by developing a new speech corpus. With a RI of 20% to 25%, fMMI was a more promising method than MMI and BMMI. After a series of experiments, it was concluded that POV outperformed others when particular prosodic features were coupled with MFCC on fMMI modeling approaches. Integration of POV, F0, and voice quality prosodic features was conducted again to improve performance, and a RI of 4% was observed. The out-domain data augmentation technique was employed to enhance the training dataset to avoid the scarcity of data. The RI of the MFCC experiment using the data-augmented corpus was 8%, but when integrated prosodic features were included, the RI increased to 13%. It is proposed that extracting integrated prosodic features (POV + F0 + voice quality + MFCC) from the augmented corpus led to enhanced system performance, and the system's accuracy improved to 88% for the children's Punjabi ASR system. Further work can be extended by employing the spectrogram augmentation approach, which can help generate an artificial dataset wherein a few more essential factors responsible for building an efficient children's speech recognition system need to be identified.

Author Contributions: Conceptualization, T.H., V.K. and K.G.; methodology, T.H., V.K., K.G. and A.A.; software, T.H., V.K. and A.A.; validation, V.K., T.H., A.A. and H.A.; formal analysis, K.G., A.A., H.A. and N.G.; investigation, T.H., V.K., H.A. and N.G.; resources, V.K., A.A., H.A. and N.G.; data curation, T.H., V.K., K.G. and N.G.; writing—original draft preparation, T.H., V.K. and K.G.; writing—review and editing, K.G. and V.K.; visualization, K.G., A.A., H.A. and N.G.; supervision, K.G. and V.K.; project administration, V.K., K.G., A.A. and H.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Taif University Researchers Supporting Project number (TURSP-2020/231), Taif University, Taif, Saudi Arabia.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Not Applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yu, D.; Deng, L. *Automatic Speech Recognition*; Springer: London, UK, 2016.
2. Hoy, M.B. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Med. Ref. Serv. Q.* **2018**, *37*, 81–88. [CrossRef] [PubMed]
3. Benzeghiba, M.; De Mori, R.; Deroo, O.; Dupont, S.; Erbes, T.; Jouvet, D.; Fissore, L.; Laface, P.; Mertins, A.; Ris, C.; et al. Automatic speech recognition and speech variability: A review. *Speech Commun.* **2007**, *49*, 763–786. [CrossRef]
4. Radha, V.; Vimala, C. A review on speech recognition challenges and approaches. *World Comput. Sci. Inf. Technol. J. (WCSIT)* **2012**, *2*, 1–7.
5. Bosch, L.T. Emotions, speech and the ASR framework. *Speech Commun.* **2003**, *40*, 213–225.
6. Chou, W.; Juang, B.H. (Eds.) *Pattern Recognition in Speech and Language Processing*; CRC Press: Boca Raton, FL, USA, 2003.
7. Heigold, G.; Ney, H.; Schluter, R.; Wiesler, S. Discriminative Training for Automatic Speech Recognition: Modeling, Criteria, Optimization, Implementation, and Performance. *IEEE Signal Process. Mag.* **2012**, *29*, 58–69. [CrossRef]
8. Kesarkar, M.P.; Rao, P. Feature extraction for speech recognition. *Electron. Syst. EE. Dept. IIT Bombay* **2003**. Available online: https://www.ee.iitb.ac.in/~{}esgroup/es_mtech03_sem/sem03_paper_03307003.pdf (accessed on 10 March 2021).
9. Shrawankar, U.; Thakare, V.M. Techniques for feature extraction in speech recognition system: A comparative study. *arXiv* **2013**, arXiv:1305.1145.
10. Rafi, M.S. Semantic variations of Punjabi toneme. *Lang. India* **2010**. Available online: http://escholar.umt.edu.pk:8080/jspui/bitstream/123456789/543/1/Full%20View.pdf (accessed on 12 March 2021).
11. Hasija, T.; Kadyan, V.; Guleria, K. Recognition of Children Punjabi Speech using Tonal Non-Tonal Classifier. In Proceedings of the 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 5–7 March 2021; pp. 702–706.

12. Shahnawazuddin, S.; Adiga, N.; Kathania, H.K.; Sai, B.T. Creating speaker independent asr system through prosody modification based data augmentation. *Pattern Recognit. Lett.* **2020**, *131*, 213–218. [CrossRef]
13. Shahnawazuddin, S.; Ahmad, W.; Adiga, N.; Kumar, A. In-Domain and Out-of-Domain Data Augmentation to Improve Children's Speaker Verification System in Limited Data Scenario. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7554–7558.
14. Hasija, T.; Kadyan, V.; Guleria, K. Out Domain Data Augmentation on Punjabi Children Speech Recognition using Tacotron. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2021; Volume 1950, p. 012044.
15. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S. Tacotron: Towards End-to-End Speech Synthesis. *arXiv* **2017**, arXiv:1703.10135.
16. Davis, K.H.; Biddulph, R.; Balashek, S. Automatic Recognition of Spoken Digits. *J. Acoust. Soc. Am.* **1952**, *24*, 637–642. [CrossRef]
17. Chengyou, W.; Diannong, L.; Tiesheng, K.; Huihuang, C.; Chaojing, T. Automatic Speech Recognition Technology Review. *Acoust. Electr. Eng.* **1996**, *49*, 15–21.
18. Wang, D.; Wang, X.; Lv, S. An Overview of End-to-End Automatic Speech Recognition. *Symmetry* **2019**, *11*, 1018. [CrossRef]
19. Juang, B.H.; Rabiner, L.R. Hidden Markov models for speech recognition. *Technometrics* **1991**, *33*, 251–272. [CrossRef]
20. Lee, K.-F.; Hon, H.-W.; Reddy, R. An overview of the SPHINX speech recognition system. *IEEE Trans. Acoust. Speech Signal Process.* **1990**, *38*, 35–45. [CrossRef]
21. Huang, X.; Alleva, F.; Hon, H.-W.; Hwang, M.-Y.; Lee, K.-F.; Rosenfeld, R. The SPHINX-II speech recognition system: An overview. *Comput. Speech Lang.* **1993**, *7*, 137–148. [CrossRef]
22. Milner, B.P. A comparison of front-end configurations for robust speech recognition. In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, Orlando, FL, USA, 13–17 May 2002; Volume 1, p. I-797.
23. Gupta, K.; Gupta, D. An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system. In Proceedings of the 2016 6th International Conference—Cloud System and Big Data Engineering (Confluence), Noida, India, 14–15 January 2016; pp. 493–497.
24. Wang, L.; Ambikairajah, E.; Choi, E. Automatic Language Recognition with Tonal And Non-Tonal Language Pre-Classification. In Proceedings of the 2007 15th European Signal Processing Conference, Poznan, Poland, 3–7 September 2007; pp. 2375–2379.
25. Wang, L.; Ambikairajah, E.; Choi, E.H. A Novel Method for Automatic Tonal and Non-Tonal Language Classification. In Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, Beijing, China, 2–5 July 2007; pp. 352–355.
26. Skerry-Ryan, R.J.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Saurous, R.A. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4693–4702.
27. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R. Natural Tts Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
28. Yasuda, Y.; Wang, X.; Takaki, S.; Yamagishi, J. Investigation of Enhanced Tacotron Text-to-speech Synthesis Systems with Self-attention for Pitch Accent Language. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6905–6909.
29. Povey, D.; Woodland, P. Improved discriminative training techniques for large vocabulary continuous speech recognition. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (Cat. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; Volume 1, pp. 45–48.
30. Povey, D.; Woodland, P.C. Minimum phone error and I-smoothing for improved discriminative training. In Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 13–17 May 2002; Volume 1, p. I-105.
31. Povey, D.; Kingsbury, B.; Mangu, L.; Saon, G.; Soltau, H.; Zweig, G. fMPE: Discriminatively Trained Features for Speech Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA, 23 March 2005; Volume 1, p. I-961.
32. Povey, D.; Kanevsky, D.; Kingsbury, B.; Ramabhadran, B.; Saon, G.; Visweswariah, K. Boosted MMI for model and feature-space discriminative training. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 4057–4060.
33. Veselý, K.; Ghoshal, A.; Burget, L.; Povey, D. Sequence-discriminative training of deep neural networks. *Interspeech* **2013**, *2013*, 2345–2349. Available online: http://www.fit.vutbr.cz/research/groups/speech/publi/2013/vesely_interspeech2013_IS131333.pdf (accessed on 20 March 2021).
34. McDermott, E.; Hazen, T.J.; Le Roux, J.; Nakamura, A.; Katagiri, S. Discriminative Training for Large-Vocabulary Speech Recognition Using Minimum Classification Error. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *15*, 203–223. [CrossRef]
35. Dua, M.; Aggarwal, R.K.; Biswas, M. Discriminative Training using Heterogeneous Feature Vector for Hindi Automatic Speech Recognition System. In Proceedings of the 2017 International Conference on Computer and Applications (ICCA), Dubai, United Arab Emirates, 6–7 September 2017; pp. 158–162.
36. Dua, M.; Aggarwal, R.K.; Biswas, M. GFCC based discriminatively trained noise robust continuous ASR system for Hindi language. *J. Ambient Intell. Humaniz. Comput.* **2019**, *10*, 2301–2314. [CrossRef]

37. Kaur, H.; Kadyan, V. Feature Space Discriminatively Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC), New Delhi, India, 21–23 February 2020.
38. Mary, L.; Yegnanarayana, B. Extraction and representation of prosodic features for language and speaker recognition. *Speech Commun.* **2008**, *50*, 782–796. [CrossRef]
39. Teixeira, J.P.; Oliveira, C.; Lopes, C. Vocal Acoustic Analysis—Jitter, Shimmer and HNR Parameters. *Procedia Technol.* **2013**, *9*, 1112–1122. [CrossRef]
40. Czap, L.; Pintér, J.M. Intensity feature for speech stress detection. In Proceedings of the 2015 16th International Carpathian Control Conference (ICCC), Szilvasvarad, Hungary, 27–30 May 2015; pp. 91–94.
41. Kurata, G.; Itoh, N.; Nishimura, M. Acoustically discriminative training for language models. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 4717–4720.
42. Yu, D.; Deng, L. Deep Neural Network Sequence-Discriminative Training. In *Automatic Speech Recognition*; Springer: London, UK, 2015; pp. 137–153.
43. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 workshop on automatic speech recognition and understanding, Big Island, HI, USA, 11–15 December 2011.
44. Morris, A.C.; Maier, V.; Green, P. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In Proceedings of the Eighth International Conference on Spoken Language Processing, Jeju Island, Korea, 4–8 October 2004.
45. Somervuo, P.; Chen, B.; Zhu, Q. Feature transformations and combinations for improving ASR performance. *Interspeech* **2003**. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.408.9409&rep=rep1&type=pdf (accessed on 10 March 2021).
46. Bhardwaj, V.; Kadyan, V. Deep Neural Network Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 374–378.
47. Dua, M.; Aggarwal, R.K.; Kadyan, V.; Dua, S. Punjabi automatic speech recognition using HTK. *Int. J. Comput. Sci. Issues (IJCSI)* **2012**, *9*, 359.
48. Kadyan, V.; Mantri, A.; Aggarwal, R.K. A heterogeneous speech feature vectors generation approach with hybrid hmm classifiers. *Int. J. Speech Technol.* **2017**, *20*, 761–769. [CrossRef]
49. Kadyan, V.; Mantri, A.; Aggarwal, R.K. Refinement of HMM Model Parameters for Punjabi Automatic Speech Recognition (PASR) System. *IETE J. Res.* **2018**, *64*, 673–688. [CrossRef]
50. Kadyan, V. Acoustic Features Optimization for Punjabi Automatic Speech Recognition System. Ph.D. Dissertation, Chitkara University, Rajpura, India, 2018.