







## Enhanced Dialectal Speech Recognition in Punjabi Using Pitch-Based Acoustic Modeling

Vivek Bhardwaj<sup>1</sup>, Deepak Thakur<sup>2\*</sup>, Tanya Gera<sup>2</sup>, Vikrant Sharma<sup>3</sup>

<sup>1</sup> School of Computer Science and Engineering, Manipal University Jaipur, Jaipur 303007, India

<sup>2</sup> Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab 140401, India

<sup>3</sup> Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun 248002, India

Corresponding Author Email: [deepak.thakur@chitkara.edu.in](mailto:deepak.thakur@chitkara.edu.in)

<https://doi.org/10.18280/isi.280612>

### ABSTRACT

**Received:** 19 June 2023

**Revised:** 11 October 2023

**Accepted:** 16 December 2023

**Available online:** 23 December 2023

#### Keywords:

*Automatic Speech Recognition (ASR), pitch, Punjabi language, speech signal processing, dialectal variations*

Automatic Speech Recognition (ASR) systems usually have difficulty accurately transcribing dialectal variations, resulting in subpar performance in areas where dialectal variants are common. The pitch-based Dialect ASR method we described in this paper aims to improve voice recognition for dialectal differences of Punjabi language. We use the pitch information that was taken out of the voice signal as a feature to enhance the dialectal nuance recognition. The suggested system includes a cutting-edge pitch-based feature extraction module that records minute differences in pitch patterns linked to various dialects. This module gives the ASR system the ability to distinguish between phonetic units more effectively and faithfully depict the distinguishing features of dialectal speech. To develop reliable representations from the pitch-based data, we also use deep learning approaches, speaker adaptive training, and vocal-tract length normalization (VTLN). The experimental results show the significant reduction in the WER of 6.63% and 4.98% for Malwa and Majha dialects. Language learning applications could benefit from the developed Punjabi dialectal speech recognition system by offering learners exposure to various dialects and accents. This can help learners develop a well-rounded understanding of the language and better adapt to different regional variations.

## 1. INTRODUCTION

Automatic Speech Recognition (ASR), also known as Speech-to-Text or Speech Recognition, is a technology that enables machines to convert spoken language into written text [1, 2]. The goal of this area of artificial intelligence and signal processing is to close the communication gap between humans and computers. Regardless of the speaker's accent, language, or background noise, ASR aims to create algorithms and systems that can accurately transcribe spoken words. It has several uses, including voice assistants, dictation software, call centre automation, language learning tools, and transcription services. Figure 1 illustrates the major steps that make up the ASR process:

**Acoustic Signal Capture:** A microphone or other audio input device is used to capture the acoustic signal containing the spoken voice in the beginning of an ASR system.

**Pre-processing:** Pre-processing is done on the collected signal to improve audio quality and get rid of undesired artefacts. This involves methods like noise reduction, filtering, and signal normalization.

**Feature Extraction:** The pre-processed audio signal is changed into a representation that the ASR system can process in this step. Mel-frequency cepstral coefficients (MFCCs) and filter banks are common methods for capturing the speech's spectral information.

**Acoustic Modeling:** The retrieved features are mapped to phonetic units or subword units using acoustic models. The link between the audio features and associated textual transcriptions is often learned utilising large datasets of speech transcriptions to train these models.

**Language Modeling:** To capture the linguistic context and increase recognition accuracy, language models are used. They aid in the resolution of ambiguities in the recognition process and evaluate the probability of word sequences. Statistical n-gram models or more sophisticated methods like recurrent neural networks (RNNs) or transformers can serve as the foundation for language models.

**Decoding:** The ASR system now looks for the transcription that is most likely given the auditory and linguistic models. The procedure entails decoding the audio information into a string of words or other language constructions that most closely approximate the spoken input.

**Post-processing:** The output of the identified text can be improved using post-processing techniques like language-specific rules, spell checking, or context-based correction.

Prosodic features, which offer information beyond the auditory signal itself, are essential for speech identification. Prosody is the collective term for the rhythm, intonation, stress patterns, and additional suprasegmental features of speech that carry linguistic and emotional information. In speech, prosody carries expressive and emotional content. Prosodic features capture distinctive patterns and traits associated with various dialects, which are important for dialect speech recognition [3, 4]. Language dialects are regional or social variations that differ in prosody, vocabulary, grammar, and pronunciation.

The accuracy and performance of ASR systems are also heavily rely on the level of complexity of the models utilised, the size and quality of the training data, and the pipeline's different optimizations. Due to the availability of extensive speech datasets and improvements in deep learning techniques, ASR has significantly advanced recently. The handling of

various accents, robustness to background noise, and correct transcription of spontaneous speech are still issues. Overall, ASR is essential for enabling speech-based human-computer interaction and has a wide range of real-world uses that are always developing and getting better as technology progresses.

Thus, in this work we have used the pitch prosody which is an important acoustic feature in speech that refers to the perceived frequency of a sound and is related to the fundamental frequency of the vocal folds' vibration. Dialectal speech recognition involves training speech recognition systems to accurately transcribe and understand speech patterns that are specific to dialects or accents. So, for training the system, we have also developed a speech corpus with the help of different dialectal people. By introducing the pitch feature we found that there is significant decrease in the WER of the system under different dialects of Punjabi language.

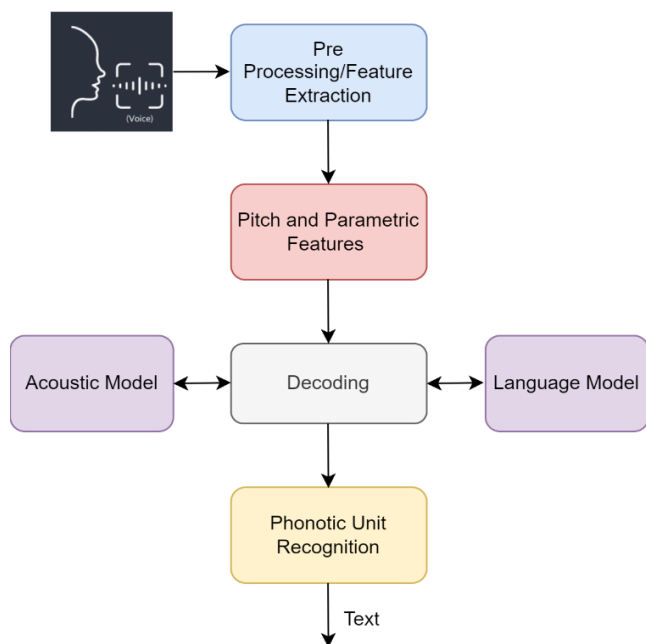


Figure 1. Basic ASR system

## 2. LITERATURE REVIEW

The book A review of the literature on speech recognition would include surveying and summarizing relevant academic papers, articles, and research studies on the topic. Here is a brief overview of some key areas and recent advancements in speech recognition research:

**Deep Learning Approaches:** Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers in particular have completely changed the field of speech recognition. To increase accuracy and robustness, researchers have looked into various designs and training approaches. For instance, end-to-end models that directly map audio features to text sequences, like Listen, Attend and Spell (LAS) and Connectionist Temporal Classification (CTC), have demonstrated encouraging results [5-7].

**Acoustic Modeling:** Speech recognition systems need to have acoustic modelling in place. Enhancing modelling methods to capture fine-grained acoustic information, accommodate many languages and accents, and overcome the difficulties brought on by noisy surroundings are the main goals of research. The use of deep neural networks (DNNs),

hybrid systems integrating DNNs and hidden Markov models (HMMs), and more recently end-to-end methodologies are examples of these breakthroughs [8, 9].

**Language Modeling:** Language models are essential for voice recognition because they offer contextual data that helps with precise transcription. Large-scale language models based on transformers, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have been the subject of recent research to increase recognition accuracy. To develop contextual representations and improve language understanding, these models make use of enormous amounts of text data [6, 10].

**Multilingual and Code-Switching ASR:** In order to handle situations where speakers switch languages during a conversation and numerous languages, researchers have been attempting to develop ASR systems. This entails creating methods for modelling language boundaries, modifying acoustic and linguistic models for many languages, and dealing with issues unique to each language [11].

**Low-Resource ASR:** Creating ASR systems for languages or domains with little training data or poor resource availability is another topic of focus. Utilizing resources from high-resource languages and enhancing performance in low-resource environments are examined through the use of transfer learning, unsupervised learning, and semi-supervised learning techniques [12, 13].

**End-to-end speech recognition software** based on prosodic properties was developed by Liu et al. [14]. Prosodic parameters like pronunciation interval and energy were extracted using the coefficient distribution of attention mechanism. On spoken word recognition tasks lasting 1000 hours and 10,000 hours, experimental results demonstrate that the suggested method improves relative accuracy by 5.2% and 5.0%, respectively, over the baseline end-to-end method and increases the understandability of speech recognition outcomes.

Hema and Marquez [15] used the prosodic capabilities for the development of emotional speech Recognition using CNN and Deep learning techniques. The authors talked about how important emotions are in human communication and introduced the idea of speech emotion recognition, which uses auditory signals to identify and forecast the speaker's emotional state. It shows how spectral and prosodic features, with MFCC being one of the often-employed spectral features, can be utilised to determine the emotional state of the speaker. Results shows the accuracy rate of 78% and outperform existing systems.

We learned from the literature that the subject is developing quickly and that innovations are being developed frequently. In order to do a thorough literature review, you would need to look at a variety of articles and studies while taking into account the particular research questions or speech recognition-related topics you want to look into.

## 3. TRAINING AND TEST SPEECH CORPUS PREPARATION

This section the detailed description of the Punjabi speech dataset. We have used the following steps specific to preparing a Punjabi speech dataset:

**Define the objective:** Determine the specific tasks or applications for which you need the Punjabi speech dataset,

such as speech recognition, speaker identification, or sentiment analysis in Punjabi language.

### 3.1 Data collection

Gather Punjabi speech data from various sources. This may involve gathering publicly accessible Punjabi speech data from sources like podcasts, radio programmes, or online speech repositories or recording audio samples of native Punjabi speakers in controlled settings. Verify that you have the necessary authorizations and legal rights to use the data.

### 3.2 Annotation and labeling

Annotate the Punjabi audio data with relevant metadata and labels. The speech in Punjabi may need to be transcribed, timestamps assigned, speaker names determined, or any other relevant information. Praat tool was used for precise and dependable labelling, manual annotation by native Punjabi speakers or specialists is essential.

### 3.3 Data cleaning and preprocessing

Remove any noise, distortions, or background noise from the Punjabi audio data that can interfere with speech recognition. To enhance audio quality, use noise reduction algorithms, filters, or audio editing software. To guarantee uniform loudness levels across the dataset, normalise the audio levels.

Splitting the dataset: Make the necessary subsets of the Punjabi speech dataset for training, validation, and testing. This partitioning assists in correctly analysing model performance, fine-tuning parameters, and training machine learning models.

### 3.4 Data augmentation

Generate more training samples by modifying the Punjabi speech data in various ways. These modifications may keep the Punjabi language traits while changing the pitch, tempo, volume, or introducing background noise. The robustness and generalization of trained models are improved by data augmentation.

### 3.5 Data format and storage

Select a suitable format to store the Punjabi speech dataset, such as CSV, JSON, or TF Record for metadata and annotations, and WAV, FLAC, or MP3 for audio files. For easy access and management, arrange the data in a database or a hierarchical directory hierarchy. We have used the wav file format for storing the audio files.

To ensure linguistic precision and cultural context, it's crucial to incorporate native Punjabi speakers or subject-matter experts in the data collecting and annotation processes. Review and improve the dataset iteratively to guarantee its quality and usability. Table 1 provides a description of the Punjabi dialect speech dataset.

**Table 1.** Training and testing speech corpus for Punjabi dialects

Dataset	# of Speakers		Dialect Spoken	Total Duration (hours)	# of Utterances
	Male	Female			
<b>Train_Punj</b>	35	24	Punjabi (All dialects)	10.5	1520
<b>Test_Malwa</b>	18	15	Malwa	7.6	1130
<b>Test_Majha</b>	17	12	Majhi	6.2	930

## 4. ACOUSTIC FEATURE EXTRACTION

The processing of speech and audio signals frequently makes use of acoustic features called mel-frequency cepstral coefficients (MFCCs). Pitch can be used as an extra acoustic feature in speech and audio processing tasks in addition to Mel-frequency cepstral coefficients (MFCCs) [16, 17]. An overview of how to use MFCCs with pitch is given below:

### 4.1 Preprocessing

Frame the audio signal: Divide the speech signal into short overlapping frames using a sliding window, typically around 20-40ms with a frame shift of 10-20ms.

Apply a window function: To minimise spectral leakage and soften the borders of each frame, multiply each one with windowing.

### 4.2 MFCC extraction

#### 4.2.1 Fourier transform

Compute the Discrete Fourier Transform (DFT) for each frame: Each frame should be transformed from the time domain to the frequency domain using the Fast Fourier Transform (FFT). Obtain the power spectrum: To obtain the power spectrum representation for each frame, compute the

squared magnitude of the complex valued FFT result.

#### 4.2.2 Mel-scale filterbank

Create a filterbank: Make a series of triangular filters that closely resemble how the human auditory system perceives frequency [18, 19]. On the Mel frequency scale, these filters are evenly spaced apart.

Apply the filterbank: Multiply the filterbank coefficients by the power spectrum of each frame. The output is an energy representation for each frequency band.

#### 4.2.3 Logarithm

Take the logarithm: Compute the logarithm of the filterbank energies to compress the dynamic range of the values and approximate the human perception of loudness.

#### 4.2.4 Discrete Cosine Transform (DCT)

Apply the DCT: To decorrelate the coefficients, apply the DCT to the energies of the logarithmic filterbank. Decide how many MFCC coefficients you want: Depending on the desired quantity of MFCCs to extract, keep a subset of the generated DCT coefficients. For speech applications, the first 12–13 coefficients are often kept.

#### 4.2.5 Delta and delta-delta coefficients

To capture temporal information, compute the delta

coefficients, or the rate of change of the MFCC coefficients over adjacent frames.

Calculate the coefficients for delta-delta: To record acceleration information, determine the delta coefficients' rate of change.

### 4.3 Pitch extraction

Use an algorithm for pitch estimation: Calculate the pitch period or fundamental frequency for each frame using a pitch estimation technique, such as autocorrelation, cepstral analysis, or the harmonic product spectrum [20, 21].

Pitch frequency to pitch period conversion: Take the reciprocal of the pitch period to determine the pitch frequency.

### 4.4 Combining MFCC and pitch

Concatenate MFCC coefficients and pitch values: Concatenate the MFCC coefficients (e.g., the first 12-13 coefficients) with the corresponding pitch values for each frame. Normalize the features: Perform mean normalization and variance normalization across the concatenated feature vector if desired. The MFCC coefficients and pitch values will be included in the feature vector that results for each frame. For a variety of speech and audio processing applications, such as speech recognition, speaker identification, emotion analysis, or prosody, this combined feature representation can be used.

## 5. EXPERIMENTAL SETUP AND RESULTS

We give a basic overview of the Kaldi toolkit. Two external libraries, OpenFst [22] for the finite-state framework and numerical algebra libraries, both of which are open-source, were required by the toolkit. We use the standard "Linear Algebra PACKage" (LAPACK)2 and "Basic Linear Algebra Subroutines" (BLAS) routines for the latter. The library modules can be split into two distinct groups, each of which is solely dependent on one of the external libraries. The C++ command-line utilities used to provide access to the library's features are used to create and execute speech recognizers. There are several executables that may be used to update a GMM-based acoustic model by adding statistics, adding accumulators, and using maximum likelihood estimation. Each tool only supports a small number of command line inputs and has a very small set of features. Every tool can read from and write to pipelines, making it easy to chain together various tools. To avoid "code rot," we have structured the toolkit so that adding new features frequently necessitates writing new code and command-line tools rather than altering old ones. This research project's front-end employs MFCC to extract acoustic information from an audio source. Each audio stream had a sample frequency of 16 kHz. The basic ASR acoustic model for Punjabi language is produced by combining the Kaldi toolbox and Karel's DNN recipe [23, 24]. The 13 mean and variance normalised cepstral coefficients are used to train the hidden Markov models (HMMs). MLLR, fMLLR for speaker independence, and SAT have all been employed as additional adaptation techniques [25-27], and [28]. Through VTLN [29], inter- and intra-speaker variability in young children is addressed. A hybrid DNN-HMM-based Punjabi-ASR system was developed in addition to the GMM-HMM system. The dialect-specific ASR system based on DNN-HMM replaces the posterior probabilities of the GMM-based system. The transcription of the training set (Text file)

was used to build an ARPA-based trigram model that decoded the speech.

Evaluating an Automatic Speech Recognition (ASR) system's accuracy in transcription is a key component of evaluation. Here, the Word Error Rate (WER) statistic is being utilised to assess the ASR systems.

$$WER = \frac{\text{Substitution Error} + \text{Insertion Error} + \text{Deletion Error}}{\text{Total Words}} \quad (1)$$

WER, which is represented in equation 1, is a commonly used metric to assess how accurately the ASR system transcribes words. In comparison to the reference transcript, it determines the percentage of substitution, deletion, and insertion errors. Better performance is indicated by lower WER values; a flawless transcription produces a WER of 0. Using a database of 10.5 hours of speech recorded from four different dialects of Punjabi, the system was trained.

### 5.1 Performance evaluation of Malwa Punjabi dialect

This section Provides a discussion of the ASR system's test results using the dataset for the Malwai Punjabi dialect. Speech files totaling 7.6 hours are in the dataset. Table 2 displays the results of the Malwai dialect recognition. The performance of the base ASR system and the experimental results of the developed ASR system with pitch features are also contrasted. Using the DNN-HMM, the baseline ASR system generates a WER of 27.7%. While the WER was decreased to 21.07 percent when the created ASR system was evaluated employing the DNN-HMM and pitch characteristics. According to the findings, the WER has decreased by 6.63 percent when compared to the baseline ASR system evaluated utilizing Malawian dialect speakers.

**Table 2.** Recognition results of the baseline ASR system and pitch based ASR system for Malwa Punjabi dialect

Acoustic Model	Pun_Malwa	Pun_Malwa (Pitch)
Monophone	55.44	48.44
Triphone	44.59	39.4
Triphone (LDA+MLLT)	42.19	36.52
Triphone (LDA+MLLT+SAT)	37.5	33.53
Triphone (LDA+MLLT+SAT+VTLN)	34.7	30.2
DNN-HMM	32.5	28.54
DNN-HMM (LDA+MLLT)	31.14	26.5
DNN-HMM (LDA+MLLT+SAT)	29.71	24.15
DNN-HMM (LDA+MLLT+SAT+VTLN)	27.7	21.07

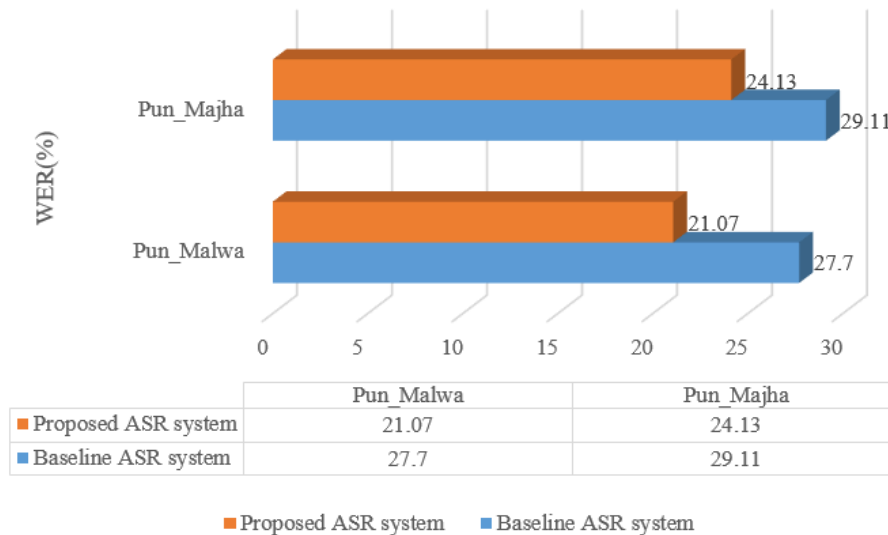
### 5.2 Performance evaluation of Majha Punjabi dialect

From the Table 3 it is clear that using the DNN-HMM, the baseline ASR system provides a WER of 30.45%. While the WER was decreased to 24.13 percent when the created ASR system was evaluated using the DNN-HMM and pitch characteristics. According to the findings, the WER has decreased by 6.32 percent when compared to the baseline ASR system evaluated utilising Malawian dialect speakers.

Figure 2 show the comparison of the best WER of the baseline ASR system and proposed ASR system. From the figure it is clear that there is huge reduction in the WER for both Malwa and Majha dialect of the Punjabi language. WER for the Majha and Malwa dialect is reduced to the 24.13 % and 21.07%.

**Table 3.** Recognition results of the baseline ASR system and pitch based ASR system for Majha Punjabi dialect

Acoustic Model	Pun_Majha	Pun_Majha (Pitch+POV)
Monophone	46.09	39.68
Triphone	42.93	34.93
Triphone (LDA + MLLT)	38.70	33.80
Triphone (LDA + MLLT+SAT)	36.75	31.30
Triphone (LDA+MLLT+SAT+VTLN)	33.19	29.11
DNN-HMM	35.74	32.26
DNN-HMM (LDA + MLLT)	33.95	31.10
DNN-HMM (LDA + MLLT + SAT)	32.93	28.12
DNN-HMM (LDA+MLLT+SAT+VTLN)	30.45	24.13

**Figure 2.** Comparison of the results of the baseline and proposed ASR system

Pitch is a fundamental aspect of prosody that relates to the perceived frequency of a person's voice, and it carries important information about intonation, emphasis, and emotional content in speech. From the results we found that proposed pitch-induced Automatic Speech Recognition (ASR) system performs better than the baseline system by leveraging the pitch information present in the speech signal to enhance its transcription accuracy.

## 6. CONCLUSIONS AND FUTURE WORK

In this research, we proposed a pitch-based Dialect ASR system that aims to improve speech recognition by accounting for dialectal variances. Our method made use of pitch information as a characteristic to capture the subtle subtleties of dialectal speech, allowing for a more precise identification of dialectal variants. We performed trials on a sizable dataset containing a variety of dialectal speech samples in order to assess the efficacy of our method. The outcomes show that our pitch-based dialect automatic speech recognition system outperforms baseline automatic voice recognition systems in dialectal recognition tasks, achieving notable increases in recognition accuracy and lowering the word error rate (WER) for dialectal speech to 21.07% and 24.13% for Punjabi Malwa and Majha dialects. Our investigation showed that pitch-based features were effective at identifying and separating the phonetic units connected to diverse dialects. For enhancing speech recognition in dialectal differences, the suggested Pitch-based Dialect ASR system offers a promising route. Our solution addresses the difficulties presented by dialectal

speech and lays the path for more precise and effective speech identification in dialect-rich regions by utilising pitch information and utilising cutting-edge deep learning techniques.

Pitch is also closely related to emotional content in speech. A pitch-induced ASR system could better identify emotional nuances in speech, which can be valuable in applications like sentiment analysis or emotion recognition. In future we work on recognizing the emotional speech of the Punjabi Language.

## REFERENCES

- [1] Nasr, S., Duwairi, R., Quwaider, M. (2023). End-to-end speech recognition for arabic dialects. *Arabian Journal for Science and Engineering*, 48: 10617-10633. <https://doi.org/10.1007/s13369-023-07670-7>
- [2] Bhatt, S., Jain, A., Dev, A. (2020). Acoustic modeling in speech recognition: A systematic review. *International Journal of Advanced Computer Science and Applications*, 11(4): 55. <https://doi.org/10.14569/IJACSA.2020.0110455>
- [3] 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), Greater Noida, India, pp. 374-378. <https://doi.org/10.1109/ICCCA49541.2020.9250780>
- [4] Murty, K.S.R., Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8): 1602-

1613. <https://doi.org/10.1109/TASL.2008.2004526>
- [5] Serizel, R., Giuliani, D. (2014). Deep neural network adaptation for children's and adults' speech recognition. In proceedings of the first Italian conference on computational linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014, Pisa, pp. 344-348.
- [6] Kaur, N., Singh, P. (2023). Modelling of speech parameters of Punjabi by pre-trained deep neural network using stacked denoising autoencoders. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3): 1-17. <https://doi.org/10.1145/3568308>
- [7] Imaizumi, R., Masumura, R., Shiota, S., Kiya, H. (2022). End-to-end japanese multi-dialect speech recognition and dialect identification with multi-task learning. *APSIPA Transactions on Signal and Information Processing*, 11(1): e4. <http://doi.org/10.1561/116.00000045>
- [8] Giuliani, D., BabaAli, B. (2015). Large vocabulary children's speech recognition with DNN-HMM and SGMM acoustic modeling. In Sixteenth Annual Conference of the International Speech Communication Association. <http://kaldi.sourceforge.net>
- [9] Qian, M., McLoughlin, I., Quo, W., Dai, L. (2016). Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM. In 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, pp. 1-5. <https://doi.org/10.1109/ISCSLP.2016.7918386>
- [10] Bhardwaj, V., Kukreja, V. (2021). Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions. *Applied Acoustics*, 177: 107918. <https://doi.org/10.1016/j.apacoust.2021.107918>
- [11] Hamed, I., Denisov, P., Li, C.Y., Elmahdy, M., Abdennadher, S., Vu, N. T. (2022). Investigations on speech recognition systems for low-resource dialectal Arabic-English code-switching speech. *Computer Speech & Language*, 72: 101278. <https://doi.org/10.1016/j.csl.2021.101278>
- [12] Bhogale, K., Raman, A., Javed, T., Doddapaneni, S., Kunchukuttan, A., Kumar, P., Khapra, M.M. (2023). Effectiveness of mining audio and text pairs from public data for improving ASR systems for low-resource languages. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5. <https://doi.org/10.1109/ICASSP49357.2023.10096933>
- [13] van der Westhuizen, E., Kamper, H., Menon, R., Quinn, J., Niesler, T. (2022). Feature learning for efficient ASR-free keyword spotting in low-resource languages. *Computer Speech & Language*, 71: 101275. <https://doi.org/10.1016/j.csl.2021.101275>
- [14] Liu, C., Wan, G., Gao, J., Fu, Z. (2023). End-to-end speech recognition method based on prosodic features. *Journal of Computer Applications*, 43(2): 380-384. <https://doi.org/10.11772/j.issn.1001-9081.2022010009>
- [15] Hema, C., Marquez, F.P.G. (2023). Emotional speech recognition using cnn and deep learning techniques. *Applied Acoustics*, 211: 109492. <https://doi.org/10.1016/j.apacoust.2023.109492>
- [16] Yadav, I.C., Pradhan, G. (2019). Significance of pitch-based spectral normalization for children's speech recognition. *IEEE Signal Processing Letters*, 26(12): 1822-1826. <https://doi.org/10.1109/LSP.2019.2950763>
- [17] Shahnawazuddin, S., Dey, A., Sinha, R. (2016). Pitch-adaptive front-end features for robust children's ASR. In *Interspeech*, pp. 3459-3463. <https://doi.org/10.21437/Interspeech.2016-1020>
- [18] Naing, H.M.S., Miyanaga, Y., Hidayat, R., Winduratna, B. (2019). Filterbank analysis of MFCC feature extraction in robust children speech recognition. In 2019 International Symposium on Multimedia and Communication Technology (ISMAC), Quezon City, Philippines, pp. 1-6. <https://doi.org/10.1109/ISMAC.2019.8836181>
- [19] Ghai, S., Sinha, R. (2011). A study on the effect of pitch on LPCC and PLPC features for children's ASR in comparison to MFCC. In Twelfth Annual Conference of the International Speech Communication Association, 2589-2592. <https://doi.org/10.21437/Interspeech.2011-662>
- [20] Cabral, J.P., Oliveira, L.C. (2005). Pitch-synchronous time-scaling for prosodic and voice quality transformations. In Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, pp. 1137-1140. <https://doi.org/10.21437/Interspeech.2005-209>
- [21] Shahnawazuddin, S., Sinha, R., Pradhan, G. (2017). Pitch-normalized acoustic features for robust children's speech recognition. *IEEE Signal Processing Letters*, 24(8): 1128-1132. <https://doi.org/10.1109/LSP.2017.2705085>
- [22] Bhardwaj, V., Bala, S., Kadyan, V., Kukreja, V. (2020). Development of robust automatic speech recognition system for children's using kaldi toolkit. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, pp. 10-13. <https://doi.org/10.1109/ICIRCA48905.2020.9182941>
- [23] Serizel, R., Giuliani, D. (2014). Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, pp. 135-140. <https://doi.org/10.1109/SLT.2014.7078563>
- [24] Qian, Y., Wang, X.H., Evanini, K., Suendermann-Oeft, D. (2016). Improving DNN-based automatic recognition of non-native children's speech with adult speech. In 5th Workshop on Child Computer Interaction, USA, pp. 40-44. <https://doi.org/10.21437/WOCCI.2016-7>
- [25] Klejch, O., Fainberg, J., Bell, P., Renals, S. (2019). Speaker adaptive training using model agnostic meta-learning. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, pp. 881-888. <https://doi.org/10.1109/ASRU46091.2019.9003751>
- [26] Takaki, S., Kim, S., Yamagishi, J. (2016). Speaker adaptation of various components in deep neural network based speech synthesis. In 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, pp. 153-159. <https://doi.org/10.21437/SSW.2016-25>
- [27] Tsao, Y., Lai, Y.H. (2016). Generalized maximum a posteriori spectral amplitude estimation for speech enhancement. *Speech Communication*, 76: 112-126. <https://doi.org/10.1016/j.specom.2015.10.003>
- [28] Guglani, J., Mishra, A.N. (2020). Automatic speech

recognition system with pitch dependent features for Punjabi language on KALDI toolkit. *Applied Acoustics*, 167: 107386.  
<https://doi.org/10.1016/j.apacoust.2020.107386>

[29] Madhavi, M.C., Patil, H.A. (2019). Vocal tract length

normalization using a Gaussian mixture model framework for query-by-example spoken term detection. *Computer Speech & Language*, 58: 175-202.  
<https://doi.org/10.1016/j.csl.2019.03.005>