Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Baseline HMM-GMM Speech Recognition System for the Low-Resource Arebhashe Dialect Proverbs Using Kaldi

Venkatesh U C
Visvesvaraya Technological University
Kusumadhara S
echod@kvgce.ac.in

Visvesvaraya Technological University
Savitha C K
Visvesvaraya Technological University
Smitha M L
Visvesvaraya Technological University

Research Article

Keywords: Automatic Speech Recognition, Arebhashe, Kannada dialect, HMM-GMM, Low-resource speech recognition

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10445651/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Baseline HMM-GMM Speech Recognition System for the Low-Resource Arebhashe Dialect Proverbs Using Kaldi

***Abstract*— Automatic Speech Recognition (ASR) for low-resource dialects remains a significant challenge due to limited speech corpora, dialectal variability, and the scarcity of linguistic resources. This paper presents the development and evaluation of an Automatic Speech Recognition (ASR) system for Arebhashe, a low-resource dialect of Kannada spoken in the Sullia and Puttur taluks of South Canara, Kodagu district of Karnataka and the Kasaragod district of Kerala, India. The very basic step is to address the challenge of building ASR for dialectal variations through progressive acoustic modeling using the Kaldi toolkit. The Arebhashe proverbs dataset is recorded from male and female speakers of different age groups from Sullia and nearby places. A grapheme-based pronunciation lexicon and a 3-gram statistical language model were developed from a manually collected corpus of traditional Arebhashe proverbs. The acoustic modeling pipeline progressively employed Monophone, context-dependent Triphone (Tri1), LDA+MLLT (Tri2b), and Speaker Adaptive Training (Tri3b) models using 39-dimensional MFCC and 40-dimensional LDA/fMLLR features. Word Error Rate (WER), decoding statistics and training log-likelihood were used as performance metrics in the experimental evaluation. Among the baseline acoustic models, the Monophone model achieved the lowest WER of 85.90%, while the Speaker Adaptive Training (SAT) model significantly reduced decoding cost and improved acoustic log-likelihood. Furthermore, experiments conducted on a cleaned subset of the corpus reduced the WER of the SAT model to 89.47%, demonstrating that recording quality has a considerable impact on ASR performance in low-resource settings. Despite the challenges of limited training data, this work establishes a baseline for the Arebhashe Kannada dialect and provides a foundation for future research using larger datasets and modern neural ASR models.**

***Index Terms*— Automatic Speech Recognition, Arebhashe, Kannada dialect, HMM-GMM, Low-resource speech recognition.**

1. Introduction

Automatic Speech Recognition (ASR) has made great progress during the last two decades. Progress in statistical modelling, machine learning and deep neural networks. Experiments are carried out on several high-resource languages using large annotated speech corpora and powerful acoustic and language models. However, many low-resource languages and dialects do not have annotated speech data, pronunciation lexicons and linguistic resources. This constrains the development of dependable speech recognition systems for various under-recognized languages around the globe. (UNESCO, 2009) reports that in India only 197 languages at risk, with dialectal variations receiving minimal technological support. The development of ASR to address the need for language technologies will contribute to the preservation and accessibility to indigenous languages. There is a lot of research in Kannada in India as compared to other Dravidian

Languages. Most of the work has been directed towards isolated-word recognition and speaker-dependent systems using Hidden Markov Models (HMMs) and conventional feature extraction techniques. Recent work also contributes to dialect classification and multilingual ASR architectures for Indian languages. Most of the recent advances on ASR focused on Standard Kannada, while dialectal variations of Kannada remain largely unexplored.

Arebhashe (Gowda Kannada) is a dialect of Kannada spoken mainly in the Kodagu and Dakshina Kannada districts of Karnataka. It has unique phonological, lexical and morphological features as against Standard Kannada. Dialect-specific vocabulary, variations in pronunciation and grammatical structures are included (Ramaswamy, 2019). Such linguistic variations pose a challenge to ASR systems trained on Standard Kannada speech. Therefore, it is necessary to develop dialect-specific acoustic and language models for ASR. Also, there are very few publicly available speech corpora and computational resources for Arebhashe. This makes it a typical low-resource speech recognition problem. Developing robust ASR systems for such low-resource dialects is challenging due to factors such as scarcity of annotated speech data, speaker variability, recording quality, dialect-specific pronunciation patterns, and vocabulary sparsity. Hence, it is important to build baseline systems before building an end-to-end ASR or self-supervised neural architectures.

Motivated by these challenges, this work develops a baseline Kaldi-based automatic speech recognition system for the Arebhashe dialect using conventional HMM-GMM acoustic modeling. A grapheme-based pronunciation dictionary and a statistical 3-gram language model were constructed from a manually collected corpus of traditional Arebhashe proverbs. Four acoustic modeling stages—Monophone, context-dependent Triphone (Tri1), LDA+MLLT (Tri2b), and Speaker Adaptive Training (Tri3b)—were systematically developed and evaluated using identical training and testing data to analyze the influence of increasing model complexity on recognition performance.

2. Related Work

Automatic Speech Recognition (ASR) has evolved from statistical Hidden Markov Model (HMM)-based approaches to hybrid and end-to-end deep learning architectures (L. R. Rabiner, 1989). Early ASR systems predominantly employed Hidden Markov Models combined with Gaussian Mixture Models (HMM-GMMs), which provided a robust statistical framework for acoustic modelling under limited computational resources (Gales, 1998). Enhanced evaluation metrics for speech recognition and showed that Word Error Rate (WER) provides a reliable measure of recognition accuracy by considering substitution, deletion, and insertion errors (Morris et al., 2004). Kaldi, in particular, has become one of the most widely adopted open-source toolkits because of its flexibility in implementing context-dependent acoustic models, speaker adaptation, and weighted finite-state transducer (WFST)-based decoding (Povey, 2011). Hidden Markov models with Gaussian mixture models have formed the foundation of ASR systems for Indian languages. Research on Kannada ASR initially focused on isolated-word and speaker-dependent recognition systems. Support Vector Machine (SVM)-based isolated-word recognition is investigated by (S. Hegde et al., 2012), while (P. & G., 2014) developed a discrete HMM-based speaker-dependent recogniser. Subsequently, (Ananthakrishna T et al., 2015) demonstrated the feasibility of

continuous Kannada speech recognition using HTK-based HMM acoustic models under controlled recording conditions. These early studies established the feasibility of statistical speech recognition for Kannada but were primarily limited to standard Kannada vocabulary and relatively small speech corpora.

With the availability of the Kaldi toolkit, Kannada ASR research progressed toward context-dependent acoustic modeling. (Yadava & Jayanna, 2017) developed continuous Kannada speech recognition systems using Kaldi and demonstrated that triphone acoustic models combined with Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) significantly improved recognition performance compared with conventional monophone systems. Similar observations were reported for domain-specific Kannada speech recognition, where context-dependent acoustic modelling reduced word error rates through improved modelling of phonetic context (Thimmaraja & Jayanna, 2017). (Mahadevaswamy & Ravi, 2021) further investigated continuous Kannada sentence recognition using Kaldi by comparing MFCC, PLP, wavelet features, and hybrid DNN-HMM architectures, demonstrating that advanced acoustic features and hybrid neural models can substantially improve recognition accuracy when sufficient training data are available (R & K Sreenivasa, 2016). (Anusuya & Katti, 2019) pioneered speaker-independent recognition using vector quantisation, achieving baseline accuracies for controlled vocabularies.

Several studies have extended Kaldi-based ASR to other Indian languages and low-resource scenarios. (Babu et al., 2018) developed a continuous Malayalam speech recognition system using Kaldi and reported consistent improvements as the acoustic models progressed from monophone to triphone and MLLT-based systems. Similarly, (Karan et al., 2015) developed an Odia speech corpus and demonstrated the importance of balanced corpus design and context-dependent acoustic modelling for low-resource ASR. (Talware et al., 2024) applied Kaldi to Hindi keyword spotting and observed systematic reductions in Word Error Rate (WER) from monophone to triphone and hybrid DNN-HMM models. (Kumar et al., 2022) also demonstrated the effectiveness of triphone modeling for Kannada by achieving improved recognition performance through progressive acoustic model refinement. (Mamyrbayev et al., 2019) developed a continuous speech recognition system to address the lack of digital resources for the Kazakh language using the Kaldi open-source toolkit, where DNN improved performance over monophone models.

Beyond individual languages, multilingual and cross-lingual approaches have received increasing attention for addressing data scarcity. (Manjunath et al., 2017) proposed multilingual phone recognition using shared phonetic units across Indian languages, while (Mirishkar et al., 2021) investigated multilingual hybrid architectures that combine acoustic transfer learning with recurrent neural network language models to improve recognition accuracy for low-resource Indian languages. These studies demonstrate that multilingual acoustic modelling and transfer learning provide practical solutions for languages with limited annotated speech resources.

Recent advances in ASR have shifted toward deep learning and self-supervised architectures. (Bhatt et al., 2020) presented a comprehensive survey of acoustic modelling techniques and highlighted the transition from HMM-GMM systems to Deep

Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Transformer-based architectures. Their review identified MFCC as the most widely adopted feature representation while emphasizing the continued importance of context-dependent triphone modelling in conventional ASR systems. More recently, (Ahlawat et al., 2025) reviewed modern ASR technologies, including recurrent neural networks, Transformers, Conformers, multilingual pre-training, and self-supervised learning, concluding that data scarcity, speaker variability, and noise robustness remain major challenges for low-resource speech recognition despite substantial advances in deep learning.

Research on dialectal and endangered language speech recognition remains comparatively limited. (Mothukuri et al., 2020) investigated automatic Kannada dialect classification using artificial neural networks across five regional dialects and demonstrated that combining MFCC and pitch features improved dialect identification. (Jones et al., 2024) Compared Kaldi-based Elpis and fine-tuned Whisper models for Čakavian Croatian, an endangered language variety, and reported that multilingual Transformer-based models achieved substantially lower Word Error Rates than conventional Kaldi pipelines.

All these literature reviews show that Kaldi remains a practical framework for developing a reproducible baseline system. From the literature review, to the best of our knowledge, no published study has developed a Kaldi-based baseline speech recognition system specifically for the Arebhashe dialect of Kannada. The baseline acoustic models Monophone, Triphone, LDA+MLLT, and Speaker Adaptive Training (SAT) were not compared under such low-resource conditions, as has not been reported. The present work addresses these research gaps by developing the first baseline ASR for Arebhashe using the Kaldi pipeline. The system development follows using a grapheme-based pronunciation lexicon, a statistical 3-gram language model, training and a comprehensive evaluation of conventional HMM-GMM acoustic modelling techniques.

3. Methodology

The complete experimental workflow for developing the baseline Automatic Speech Recognition system (ASR) for the Arebhashe dialect using the Kaldi speech recognition toolkit is illustrated in Fig. 1. The proposed framework consists of data preparation, feature extraction, language resource development, acoustic model training, decoding, and performance evaluation.

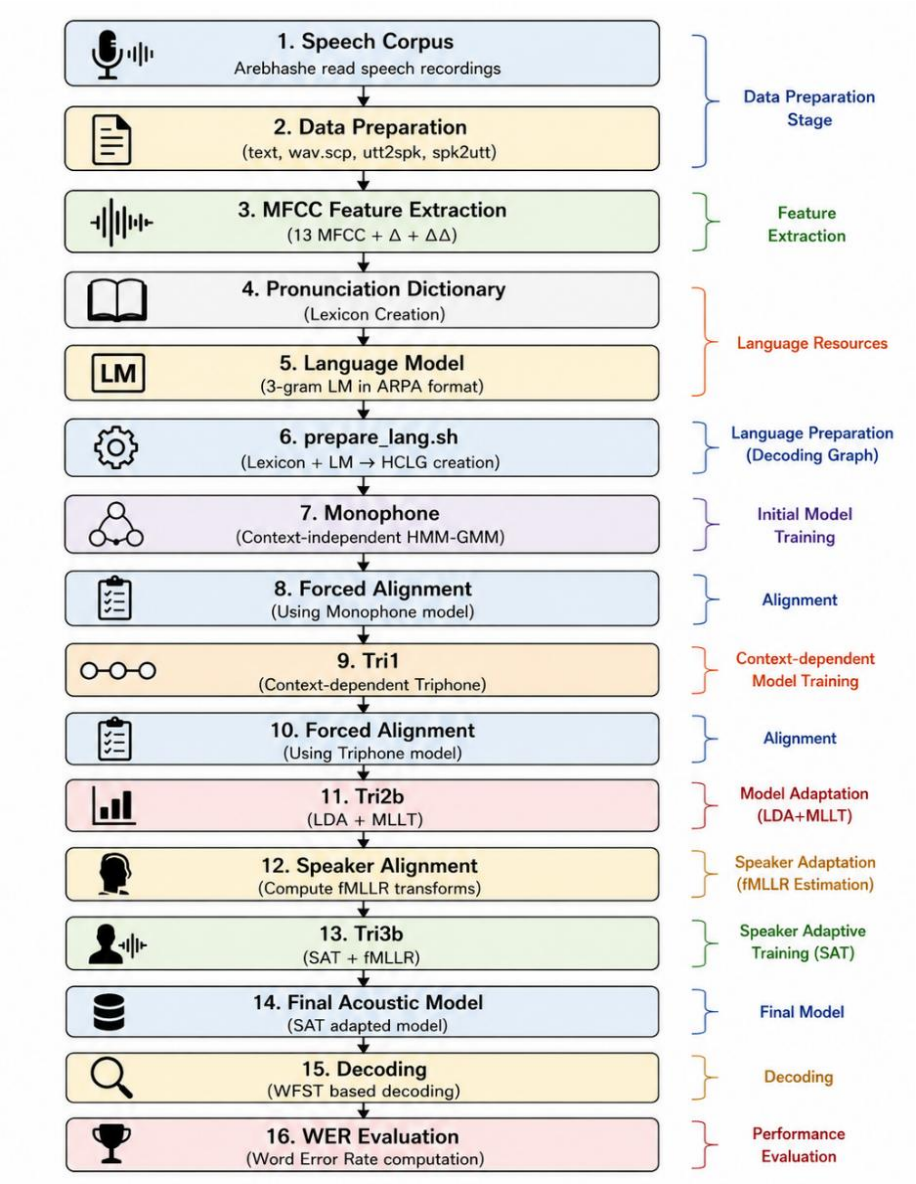


Fig.1: Kaldi Framework for Baseline ASR System Development on Arebhashe Speech

3.1. Speech Corpus and Data Preparation

Traditional Arebhashe proverbs were collected from 25 native speakers (14 male and 11 female) belonging to Sullia and nearby regions of Dakshina Kannada, Karnataka. The corpus contains 532 read utterances, each approximately 4–8 seconds long, totalling nearly one hour of speech. Recordings were captured using Audacity with a 16 kHz sampling rate, 16-bit PCM, and mono-channel configuration in a quiet indoor environment. The dataset is divided into Train and Test sets with a ratio of 80%(432) and 20%(100) wav files, respectively, in a speaker-dependent strategy. All the speech recordings were manually transcribed in Unicode Kannada (UTF-8 encoding). Each WAV file and its transcription are assigned identifiers like spk01_utt01. The Kaldi data directories were created using standard metadata files (text, wav.scp, utt2spk and spk2utt). The data verification and validation were performed by means of Kaldi's utils/validate_data_dir.sh utility. Table 1 shows transcriptions of a few sample utterances of Arebhashe proverbs with identifiers in the Test data. The last column of the table is the meaning of the Arebhashe Proverbs in English.

Table 1: Arebhashe Proverb Samples and English Translations

Unique ID	Transcriptions in Kannada	Synonyms in English
spk01_utt03	ಅಂಗ ಭಂಗ ಆದರೂ ಸಿಂಗಾರಕ್ಕೇನೂ ಕೊರತೆ ಇಲ್ಲೆ	Even after losing the property, people never stop showing their richness.
spk02_utt02	ಅಡಿಗೆ ಬಿದ್ದರೂ ಮೂಕು ಮೇಲೆ	Some people are not ready to accept their defeat or failure.
spk06_utt10	ಒಂದ್ ಕಣ್ಣಿಗೆ ಬೆಣ್ಣೆ ಒಂದು ಕಣ್ಣಿಗೆ ಸುಣ್ಣ	Showing partiality for selfishness
spk10_utt22	ನಯ ಇದ್ದರೆ ಭಯ ಇಲ್ಲೆ	Politeness avoids quarrels, and there is no fear of friction.
spk14_utt02	ಮಾತ್ ಸಕ್ಕರೆನಾಂಗೆ ಕೆಲ್ಸ್ ಕತ್ತೆನಾಂಗೆ	Polite words and hard work.

3.2. Feature Extraction

The speech features were extracted using Mel-Frequency Cepstral Coefficients (MFCCs) implemented in Kaldi . Each speech signal was divided into 25 ms frames with frame shift 10 ms, windowed with Hamming window and transformed into frequency domain by Fast Fourier Transform (FFT). The resulting spectrum was processed with 23 Mel-spaced triangular filters, logarithmic compression and Discrete Cosine Transform (DCT) to obtain 13 cepstral coefficients including the energy coefficient. To model the dynamic speech characteristics, first-order (Delta) and second-order (Delta Delta) derivatives were appended to generate 39-dimensional MFCC feature vectors. Then, Speaker-level Cepstral Mean and Variance Normalisation (CMVN) was applied to reduce variability due to channel and recording conditions. 40-dimensional discriminative feature vectors were derived by Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT). The final model, Speaker Adaptive Training (SAT), compensates for speaker differences by estimating speaker-dependent feature transformations using feature-space Maximum Likelihood Linear Regression (fMLLR).

3.3. Language Modeling

3.3.1 Pronunciation Lexicon

A pronunciation lexicon is built that maps vocabulary words to their acoustic units. Since Arebhashe is a low-resource dialect and there is no standard phonetic dictionary, we used a grapheme-based lexicon. Each lexical entry associates a word in Arebhashe with a sequence of Kannada unicode graphemes. This removes the requirement for hand-designed phoneme transcriptions. The lexicon also contains required tokens for silence (SIL), spoken-noise (SPN) and unknown-word (UNK) that are needed for Kaldi decoding. The grapheme approach eases the preparation of the lexicon and contains enough lexical material for low-resource ASR.

3.3.2 Statistical Language Model

The language model estimates the probability of a word sequence based on the two preceding words and provides linguistic constraints during decoding (Young & Young, 1993). A word-level trigram language model was developed using the SRILM

toolkit with modified Kneser–Ney smoothing and was used consistently for all acoustic models (Monophone, Tri1, Tri2b, and Tri3b). This provides reliable probability estimates for unseen word sequences by redistributing probability mass according to lower-order n-gram statistics.

$$P(w_i \mid w_{i-2}, w_{i-1})$$

The trained language model was generated in ARPA format and subsequently converted into a Finite State Transducer (FST) using Kaldi's language preparation utilities (Mohri et al., 2002). During graph construction, the pronunciation lexicon (L), context dependency (C), acoustic model (H), and language model (G) were composed to generate the decoding graph (HCLG), enabling efficient Viterbi beam-search decoding. The language model corpus was derived from the transcriptions of 431 Arebhashe utterances, comprising traditional proverbs and idioms. The text was normalized to maintain consistency in the Kannada Unicode representation and preprocessed to handle dialectal variations specific to the Arebhashe variety.

3.4. Acoustic Modeling

The acoustic model was developed using the Hidden Markov Model–Gaussian Mixture Model (HMM–GMM) framework implemented in the Kaldi speech recognition toolkit. The training procedure followed Kaldi's standard WSJ recipe, consisting of four successive acoustic models: Monophone, Context-dependent Triphone (Tri1), LDA+MLLT (Tri2b), and Speaker Adaptive Training (SAT, Tri3b). Each successive model utilized the alignments generated by the previous model to improve phonetic discrimination and recognition accuracy. Throughout the training process, Maximum Likelihood Estimation (MLE) was employed for parameter optimization using the Baum–Welch Expectation-Maximization algorithm.

3.4.1. Monophone HMM-GMM (Mono)

The initial acoustic model consisted of context-independent HMM-GMM units, where each phoneme was represented by a three-state left-to-right Hidden Markov Model (Rabiner, 1989). The observation probability for state jis given by

$$b_j(\mathbf{o}) = \sum_{m=1}^{M} c_{jm}\, \mathcal{N}(\mathbf{o}; \mu_{jm}, \Sigma_{jm})$$

where M denotes the number of Gaussian mixture components, c_{jm} represents the mixture weight, and $\mathcal{N}$ denotes the Gaussian probability density function.

3.4.2. Triphone HMM-GMM (Tri1)

The Triphone model extends the monophone system by incorporating left and right phonetic context to model co-articulation effects. Kaldi performs decision-tree state clustering to group acoustically similar HMM states, enabling reliable parameter estimation despite limited training data. The Triphone model was trained using the speaker-normalized 39-dimensional MFCC features and the alignments generated by the monophone model.

3.4.3. LDA+MLLT (Tri2b)

To improve class separability, Linear Discriminant Analysis (LDA) was applied to spliced contextual feature vectors. LDA projects the original feature space into a lower-dimensional space that maximizes discrimination between phonetic classes.

The optimal projection matrix is obtained by maximizing Fisher's discriminant criterion,

$$\boldsymbol{W = \arg\max_{W} \frac{| W^T S_B W |}{| W^T S_W W |}}$$

where $\boldsymbol{S_b}$and $\boldsymbol{S_w}$Denote the between-class and within-class scatter matrices, respectively (Fisher, 1936). The subsequent MLLT transformation maximizes the likelihood of the transformed features under diagonal-covariance Gaussian models, yielding more discriminative 40-dimensional feature representations.

3.4.4. Speaker Adaptive Training (Tri3b)

The final acoustic model employed Speaker Adaptive Training (SAT) using feature-space Maximum Likelihood Linear Regression (fMLLR) (Gales, 1998). Speaker-specific affine transformations were estimated as.

$$\mathbf{x}' = \boldsymbol{A_s}\mathbf{x} + \boldsymbol{b_s}$$

where $\boldsymbol{A}$ represents the transformation matrix and $\boldsymbol{b}$ Denotes the bias vector. The transformed features reduce speaker-dependent variability arising from differences in vocal tract characteristics, speaking style, and recording conditions, thereby improving model robustness.

3.5 Decoding

The trained acoustic models were evaluated using the Kaldi decoding framework, which performs recognition by searching for the most probable word sequence given the acoustic observations and the statistical language model. Decoding was carried out using the Weighted Finite State Transducer (WFST) framework, where the acoustic model, pronunciation lexicon, context-dependency model, and language model were combined into a single decoding graph (HCLG) (Mohri et al., 2002). The decoding graph is represented as

$$\mathbf{HCLG} = \boldsymbol{H \circ C \circ L \circ G}$$

The composed WFST enables efficient decoding by integrating acoustic, lexical, and linguistic constraints into a unified search graph. During recognition, Kaldi employs a beam-search Viterbi decoder to identify the most probable word sequence

$$\boldsymbol{\widehat{W} = \arg\max_{W} P(O \mid W)\, P(W),}$$

where $\boldsymbol{P(O \mid W)}$ **and** $\boldsymbol{P(W)}$ represents the acoustic likelihood estimated by the HMM-GMM model and the language model probability, respectively. Identical decoding parameters were used for all acoustic models to ensure a fair comparison. The decoding graph was generated using the trained trigram language model and grapheme-based pronunciation lexicon. Recognition experiments were performed on the independent Arebhashe test set using the same decoding parameters for all four acoustic models to ensure a fair comparison.

3.6 Performance Evaluation

The performance of the proposed ASR system was evaluated using Word Error Rate (WER), the most widely adopted evaluation metric in automatic speech recognition. WER measures the difference between the reference transcription and the automatically recognized output by accounting for substitution, deletion, and insertion errors. The WER is defined as

$$\mathbf{WER} = \frac{\boldsymbol{S} + \boldsymbol{D} + \boldsymbol{I}}{\boldsymbol{N}} \times \mathbf{100}$$

where $\boldsymbol{S}$, $\boldsymbol{D}$, and $\boldsymbol{I}$ denote the numbers of substitution, deletion, and insertion errors, respectively, and $\boldsymbol{N}$ represents the total number of words in the reference transcription (Morris et al., 2004). In addition to the overall WER, individual insertion, deletion, and substitution errors, decoding statistics, acoustic model complexity, and training log-likelihood were analyzed to compare the performance of the four acoustic modeling strategies under identical experimental conditions.

4. Experimental Results

The proposed ASR system was implemented using the Kaldi Speech Recognition Toolkit on Ubuntu Linux within a VirtualBox environment running on Windows 11. The standard WSJ (s5) recipe was adapted for the Arebhashe corpus. The dataset comprised 532 utterances collected from native speakers, of which 432 were used for training and 100 for testing. Feature extraction, language resource preparation, acoustic model training, and decoding were performed using the standard Kaldi pipeline.

4.1 Acoustic Model Complexity

Table 2: Acoustic Model Complexity

Model	Feature Type	Feature Dim.	PDFs
Monophone	MFCC + Δ + ΔΔ	39	171
Triphone (Tri1)	MFCC + Δ + ΔΔ	39	480
LDA+MLLT (Tri2b)	LDA + MLLT	40	640
SAT (Tri3b)	fMLLR	40	736

Four conventional GMM-HMM acoustic models were developed sequentially using the Kaldi toolkit. Table 2 summarizes the complexity of each acoustic model. Context-independent HMM-GMM acoustic model trained using 39-dimensional MFCC features (including delta and delta-delta coefficients). As the training pipeline progressed from monophone to speaker-adaptive training (SAT), the number of probability density functions (PDFs), transition states, and Gaussian mixtures increased substantially, enabling more detailed acoustic modelling.

Table 3: Acoustic Model Performance Comparison

Model	Log-Likelihood	Relative Improvement	#Gaussians	#States
Mono	-92.95	Baseline	986	699
Tri1	-86.45	+6.99%	10,044	2175
Tri2b	-38.15	+58.95%	10,295	2913
Tri3b	-40.30	+56.65%	10,225	3412

The table 3 shows complexity of the acoustic models increased gradually from Monophone to SAT. Context-dependent modeling introduced a substantially larger number of Gaussian mixtures and hidden Markov model states, allowing more detailed representation of contextual speech variations. The training log-likelihood improved substantially after introducing LDA and MLLT feature transformations, demonstrating the effectiveness of discriminative feature representations. Although the SAT model slightly reduced the final likelihood compared with Tri2b, it provided improved speaker normalization while maintaining comparable acoustic modeling performance.

Table 4: Lexicon and Language Model Details

Lexicon Details		**Language Model Details**	
Total lexicon entries	1,879	Training utterances	432
Vocabulary entries	1,877	Total training word tokens	2816
Non-silence acoustic units	52	Unique training words	1575
Silence symbols	3 (SIL, SPN, <UNK>)	Vocabulary (words.txt)	1587
Optional silence	SIL	OOV token	<UNK>
Average graphemes per word	6.43	Grammar FST (G.fst)	196 KB
Total grapheme inventory	95	Final decoding graph (HCLG.fst)	5.0 MB

Table 4 shows the lexicon and language model statistics. A statistical trigram (3-gram) language model was developed using SRILM from the training corpus. The language model was compiled into a WFST grammar (G.fst, 196 KB) and integrated with the pronunciation lexicon and acoustic model to construct the final decoding graph (HCLG.fst, 5.0 MB).

4.2. Performance Metrics and Analysis

During decoding, lattices were successfully generated for all utterances using a four-job parallel decoding process. The average lattice depth was 2.84, indicating that the decoder maintained a limited number of competing hypotheses, which is expected for a small-vocabulary read-speech corpus.

Table 5: Kaldi Gmm-Hmm Results on Test Set

Model	Insertions	Deletions	Substitutions	Best WER (%)
Monophone	23	61	373	85.90
Triphone	70	18	393	90.41
LDA+MLLT	126	12	377	96.80
SAT	87	16	398	94.17

The table 5 summarizes the recognition errors produced by the four acoustic models. Although context-dependent acoustic models generally outperform Monophone systems in large-vocabulary speech recognition tasks, the present results exhibit the opposite trend. Under such low-resource conditions, the more complex Triphone, LDA+MLLT, and SAT models require substantially more training data to estimate context-dependent parameters reliably. Consequently, the advanced models exhibited higher insertion and substitution errors during decoding. Substitution errors dominate across all acoustic models, accounting for the largest proportion of recognition errors. This indicates that the acoustic models frequently confuse phonetically similar grapheme sequences, reflecting the rich phonetic variability and dialect-specific pronunciation characteristics of Arebhashe.

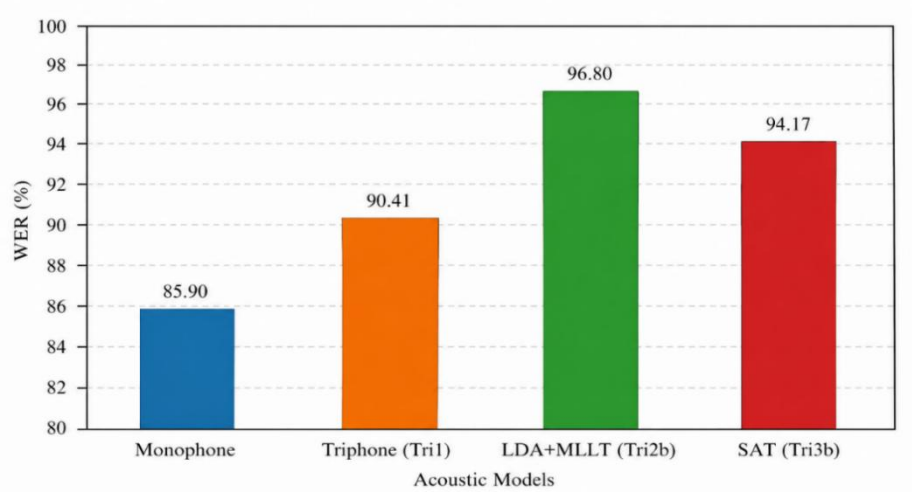


Fig.2: Comparison Of The Best Word Error Rate (WER) Obtained Using Four HMM-GMM Acoustic Models

Figure 2 shows Word Recognition Rate trend differs from conventional large-vocabulary ASR systems, where Triphone and speaker-adaptive models generally outperform Monophone systems. In the present work, the limited training corpus and recording-quality issues restricted the ability of the more complex acoustic models to generalize effectively. The comparatively poor performance of the Tri2b model indicates that LDA and MLLT transformations require substantially larger and more diverse training corpora to achieve reliable parameter estimation. Although the SAT model slightly reduced the WER relative to Tri2b, it was still unable to surpass the Monophone baseline because the speaker adaptation process was constrained by the limited number of speakers and the variability in recording conditions.

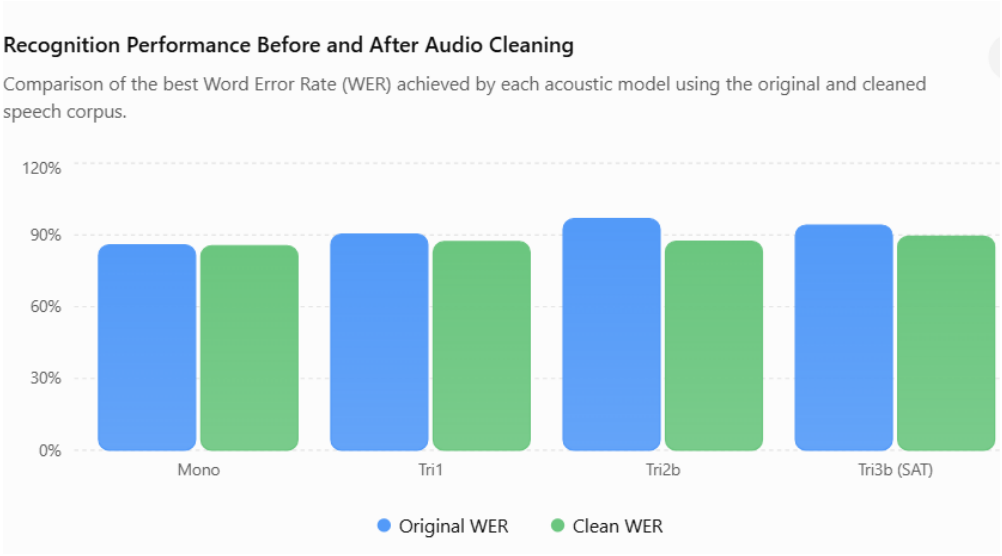


Fig .3: WER Comparison for Original Data and Clean Data of The Arebhashe Corpus

Figure 3 shows a comparison of word error rate (WER) obtained using the original and cleaned Arebhashe speech corpus. Audio cleaning is done to remove clipped and silence-heavy recordings. The same training and testing partition were maintained in the Kaldi pipeline. The resulting corpus consistently reduced the WER across all acoustic models. Confirming that recording quality is an important factor in low-resource ASR. A consistent reduction in WER was observed after removing clipped and silence-heavy recordings from the training corpus. The Monophone model showed only a marginal improvement (0.37 percentage points). The largest improvement was observed for the Tri2b model (96.80% to 87.41%),

indicating that LDA+MLLT-based acoustic modeling is particularly sensitive to clipped and silence-heavy recordings. Speaker-adaptive training (Tri3b) also benefited from audio cleaning, reducing the WER from 94.17% to 89.47%. The SAT-based Tri3b model also demonstrated improved recognition performance after cleaning, confirming that speaker adaptation benefits from higher-quality acoustic observations.

4.3 Comparative Analysis

A quantitative evaluation was carried out on four acoustic models using both the decoding lattices generated by each model and the actual recognition outputs obtained during decoding. The analysis demonstrates progressive improvements in recognition accuracy and computational efficiency from monophone to advanced triphone models with speaker adaptation (Tri3b). Table 6 summarizes the average total, acoustic, graph, and frame-normalized decoding costs. A consistent reduction in total decoding cost was observed as the acoustic model evolved from Monophone to Tri3b.

Table 6: Cost Analysis Using Lattice Best-Path Decoding

Model	Average Total Cost	Average Acoustic Cost	Average Graph Cost	Average Cost/Frame
Mono	41,011.9	40,937.2	74.68	93.81
Tri1	24,914.4	24,853.0	61.34	101.45
Tri2	13,925.3	13,859.2	66.14	56.26
Tri3	12,603.9	12,543.4	60.50	50.94

The graph in figure Fig.4 depicts the average total decoding cost. The percentage of cost reduction is relative to the Monophone model, where efficient frame-level likelihood estimation and improved decoding confidence suggests that the average cost per frame decreased substantially for the LDA+MLLT(Tri2b) and SAT(Tri3b) models.

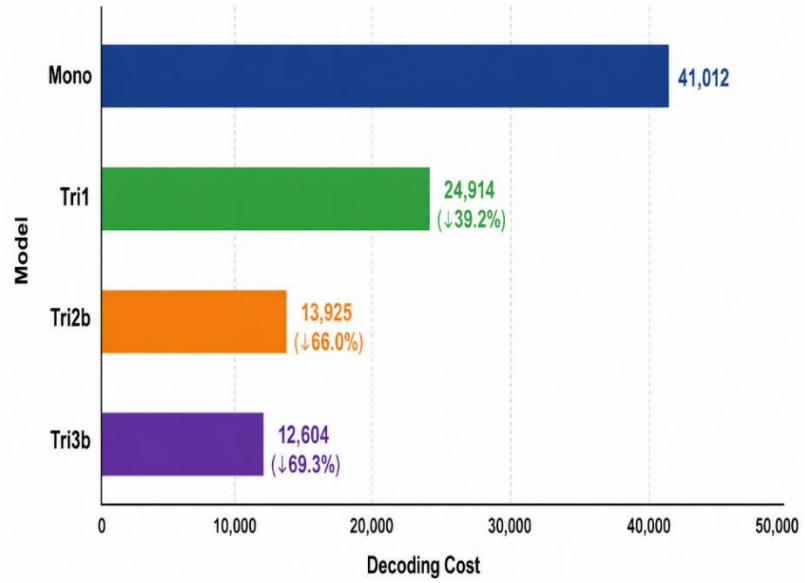


Fig.4: Graph Cost Reduction Model comparison

Table 7 below presents representative recognition outputs from the four acoustic models for five test utterances. Although

the SAT model did not achieve the lowest WER, its decoded outputs were generally more linguistically coherent than those of the preceding models, indicating improved contextual modeling despite limited training data.

Table 7: Word Recognition Comparison

Utterance ID	Test Data	Monophone	Triphone	LDA+MLLT	SAT
spk01_utt03	ಅಂಗ ಭಂಗ ಆದರೂ ಸಿಂಗಾರಕ್ಕೇನೂ ಕೊರತೆ ಇಲ್ಲ	ಅಂಗ ಬಂಗ ದೂರು ಸಿಂಗಾರ ಕೇಳುವ ಕದ್ದ ಇಲ್ಲೆ	ಅಂಗ ಬಂಗ ದೂರು ಸಿಂಗಾರ ಕೇಳೊವ್ ಕುರುಂಜಿ ಇಲ್ಲಿ	ಕೂಡಾ ಭಂಗ ದೂರು ಸಿಂಗಾರ ಕೇಳೊವ್ ಕೊಟ್ಟ ಅದ್ ಇಲ್ಲೆ	ಅಂಗ ಬಂಗಾರ ರೂಪ ಸಿಂಗಾರ ಕೇಳೊವ್ ಕೊಟ್ಟರ್ ಇಲ್ಲೆ
spk02_utt02	ಅಡಿಗೆ ಬಿದ್ದರೂ ಮೂಕು ಮೇಲೆ	ಅಡಿ ಲಿ ಗೆದ್ದವು ಮೂಕು ಮೇಲೆ	ಅಡಿ ಇದ್ ಇದ್ದರೂ ಮೂಕು ಮೇಲೆ	ಅಡಿ ಇದ್ ಇದ್ದರೂ ಮೂಕು ಮೇಲೆ	ರಿಗೆ ಇರ್ ಇದ್ದರೂ ಮೂಕು ಮೇಲೆ
spk06_utt10	ಒಂದ್ ಕಣ್ಣಿಗೆ ಬೆಣ್ಣೆ ಒಂದು ಕಣ್ಣಿಗೆ ಸುಣ್ಣ	ಒಂದು ಗಂಡಿಗೆ ಬೆಣ್ಣೆ ಒಂದು ಕಣ್ಣಗೆ ಸಣ್ಣ	ಒಂದು ಗಂಡಿಗೆ ಬೆಣ್ಣೆ ಒಂದು ಕಡ್ಡೆಗೆ ಸಣ್ಣ	ಒಂದು ಕಣ್ಣ್ ರಿಗೆ ಬೆಣ್ಣೆ ಲೆ ಒಂದು ಕಣ್ಣ್ ಎಲಿ ಸಣ್ಣ	ಒಂದು ಕಣ್ಣ್ಗೆ ಬೆಣ್ಣೆ ಒಂದು ಕಣ್ಣ್ಗೆ ಸಣ್ಣ
spk10_utt22	ನಯ ಇದ್ದರೆ ಭಯ ಇಲ್ಲೆ	ನಾಯಿ ದರೆ ಭಯ ಇಲ್ಲೆ	ನಾಯಿ ಇದ್ದರೆ ಒಬ್ಬ ಎಇಲ್ಲೆ	ನಾಯಿ ಇದ್ದರೆ ಭಯ ಇಲ್ಲೆ	ನಾಯಿ ಇದ್ದರೆ ಭಯ ಇಲ್ಲೆ
spk14_utt02	ಮಾತ್ ಸಕ್ಕರೆನಾಂಗೆ ಕೆಲ್ಸ ಕತ್ತೆನಾಂಗೆ	ಮಾತ್ ಸತ್ತರೆ ನಂಗೆ ಕೆಲ್ಸ ಕತ್ತೆ ನಿಂಗೆ	ಮಾತ್ ಸತ್ತರೆ ನಂಗೆ ಕೆಲ್ಸ ಕತ್ತೆ ನಂಗೆ	ಮಾತು ಸತ್ತರೆ ನಂಗೆ ಕೆಲ್ಸ ಕತ್ತೆ ನಂಗೆ	ಮಾತ್ ಸತ್ತರೆ ನಾ ಅಂಗಿ ಕೆಲ್ಸ ಕತ್ತೆ ನಂಗೆ

5. Discussion

Basic triphone performance shows an important challenge in a limited dataset of the Arebhashe dialect of the Kannada language for baseline ASR. Context-dependent models require sufficient examples of each context, and with only 432 training utterances, many triphone contexts are seen rarely or not at all, resulting in poor generalization. The quantitative analysis showed that the majority of recognition errors were caused by word substitution with acoustically similar words, not by omission of whole utterances or insertion of lengthy sequences. The recognition performance was slightly improved with the cleaned corpora. This shows that the quality of the recorded WAV files influences the performance.

The quantitative WER analysis also shows that the largest error component is substitutions. The acoustic cost also followed a similar trend and indicated a gradual improvement in the acoustic modelling. The graph cost is almost the same for all the models, probably all the models used the same pronunciation model. Proverbs of two or three short phrases tend to preserve more lexical content, while recognition errors tend to build up with longer sentences. The decoding cost is much lower for the SAT model than the Monophone model, but the recognised text is not necessarily more accurate. This confirms that decoding confidence alone is not enough to ensure better recognition performance in a low-resource setting. Overall, WER remained high, indicating that the dominating challenges are the limited corpus size and dialectal variability.

6. Conclusions and Future Work

The proposed ASR system for the Arebhashe dialect of Kannada shows that with limited data, progressive acoustic modelling can yield significant improvements. The best result, with the lowest WER of 85.90%, was achieved by the Monophone model. The WER of the Triphone model was 90.41%, the LDA+MLLT model 96.80%, and the SAT model 94.17%. Results show a consistent reduction in WER when we exclude clipped and silence-heavy recordings from the training corpus. The improvement is observed for the LDA+MLLT-based acoustic model (Tri2b), i.e is 96.80% to 87.41%, and speaker-adaptive training (Tri3b), with a reduction in WER from 94.17% to 89.47%. This study lays the groundwork for automatic speech recognition (ASR) in the Arebhashe dialect of Kannada. The proposed baseline system and benchmark results establish a strong foundation for future research, including the development of larger, more balanced datasets and more robust neural acoustic models.

REFERENCES

1. Ahlawat, H., Aggarwal, N., & Gupta, D. (2025). Automatic Speech Recognition: A survey of deep learning techniques and approaches. International Journal of Cognitive Computing in Engineering, 6, 201–237. tps://doi.org/10.1016/j.ijcce.2024.12.007
2. Ananthakrishna, T., Maithri, M., & Shama, K. (2015). Kannada word recognition system using HTK. 2015 Annual IEEE India Conference (INDICON)
3. Anusuya, M. A., & Katti, S. K. (2019). Speaker independent Kannada speech recognition using vector quantization. In Proc Int Conf Intell Comput Control Syst (ICICCS) (pp. 1210–1215).
4. Babu, L. B., George, A., Sreelakshmi, K. R., & Mary, L. (2018). Continuous speech recognition system for the Malayalam language using Kaldi. 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR).
5. 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). https://doi.org/10.14569/ijacsa.2020.0110455.
6. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.
7. Gales, M. J. F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech & Language, 12(2), 75–98. https://doi.org/10.1006/csla.1998.0043
8. Hegde, S., Achary, & Shetty, S. (2012). Isolated word recognition for Kannada language using support vector machine. In Wireless Networks and Computational Intelligence (pp. 262–269). Springer Berlin Heidelberg.
9. Jones, A., Zhang, S., Hale, J., Renwick, M. E. L., Vrzić, Z., & Langston, K. (2024). Comparing Kaldi-based pipeline Elpis and Whisper for Čakavian transcription. In Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics

(Field Matters 2024 (pp. 61–68). Association for Computational Linguistics.

10. Karan, B., Sahoo, J., & Sahu, P. K. (2015). Automatic speech recognition-based Odia system. In Proc Int Conf Microwave Optical Commun Eng (ICMOCE).
11. Kumar, T. N. M., Jayan, A., Bhat, S., Anvith, M., & Narasimhadhan, A. V. (2022). Monophone and triphone acoustic phonetic model for Kannada speech recognition system. 2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET).
12. Mahadevaswamy, & Ravi, D. J. (2021). Robust perceptual wavelet packet features for recognition of continuous Kannada speech. Wireless Personal Communications, 121(3), 1781–1804. https://doi.org/10.1007/s11277-021-08736-1.
13. Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Mukhsina, K., Keylan, A., BabaAli, B., Nabieva, G., Duisenbayeva, A., & Akhmetov, B. (2019). Continuous speech recognition of Kazakh language. ITM Web of Conferences, 24, 01012. https://doi.org/10.1051/itmconf/20192401012.
14. Manjunath, K. E., Rao, K. S., & Jayagopi, D. B. (2017). Development of multilingual phone recognition system for Indian languages. 2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES).
15. Mirishkar, G. S., Yadavalli, A., & Vuppala, A. K. (2021). An investigation of hybrid architectures for low-resource multilingual speech recognition system in Indian context. In Proceedings of the 18th International Conference on Natural Language Processing (pp. 205–212).
16. Mohri, M., Pereira, F., & Riley, M. (2002). Weighted finite-state transducers in speech recognition. Computer Speech & Language, 16(1), 69–88. https://doi.org/10.1006/csla.2001.0184.
17. Morris, A. C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. Interspeech 2004.
18. Mothukuri, S. K. P., Hegde, P., Chittaragi, N. B., & Koolagudi, S. G. (2020). Kannada dialect classification using artificial neural networks. 2020 International Conference on Artificial Intelligence and Signal Processing. AISP.
19. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*.
20. P., P., & G., H. (2014). Speaker dependent continuous Kannada speech recognition using HMM. 2014 International Conference on Intelligent Computing Applications.
21. R, P., & K Sreenivasa, R. (2016). Deep neural networks for Kannada phoneme recognition. 2016 Ninth International Conference on Contemporary Computing (IC3).

22. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. Institute of Electrical and Electronics Engineers, 77(2), 257–286. https://doi.org/10.1109/5.18626

23. Ramaswamy, C. (2019). *A grammar of Arebhashe*. Central University of Karnataka Press.

24. Talware, S. M., Singh, V., Sharma, C., & Tawania, H. (2024). Speech Keyword Spotting (KWS) using KALDI toolkit. International Journal of Research Publication and Reviews, 5(7), 3012–3016.

25. Thimmaraja, Y. G., & Jayanna, H. S. (2017). Creating language and acoustic models using Kaldi to build an automatic speech recognition system for Kannada language. 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT).

26. UNESCO. (2009). *Atlas of the world's languages in danger*. UNESCO Publishing.

27. Yadava, G. T., & Jayanna, H. S. (2017). Development and comparison of ASR models using kaldi for noisy and enhanced kannada speech data. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).

28. Young, S. J., & Young, S. (1993). The HTK hidden Markov model toolkit: Design and philosophy.