Research Article

A. Kumar* and R.K. Aggarwal

Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modeling

https://doi.org/10.1515/jisys-2018-0417
Received Oct 17, 2018; accepted Dec 01, 2019

Abstract: This paper implements the continuous Hindi Automatic Speech Recognition (ASR) system using the proposed integrated features vector with Recurrent Neural Network (RNN) based Language Modeling (LM). The proposed system also implements the speaker adaptation using Maximum-Likelihood Linear Regression (MLLR) and Constrained Maximum likelihood Linear Regression (C-MLLR). This system is discriminatively trained by Maximum Mutual Information (MMI) and Minimum Phone Error (MPE) techniques with 256 Gaussian mixture per Hidden Markov Model (HMM) state. The training of the baseline system has been done using a phonetically rich Hindi dataset. The results show that discriminative training enhances the baseline system performance by up to 3%. Further improvement of ~7% has been recorded by applying RNN LM. The proposed Hindi ASR system shows significant performance improvement over other current state-of-the-art techniques.

Keywords: Automatic speech recognition, MFCC, GFCC, WERBC, PLP, discriminative training, MMI, MPE, RNN LM

1 Introduction

ASR is the process of taking speech utterance and converting it into text sequence as close as possible. There are many functional areas in ASR. Some are as follows: dictation, a program control application, dialog systems, audio indexing, speech-to-speech translation, and query-based information retrieval system, i.e., weather information system, or some travel information system. With the increase in need of end-user focused applications such as look for voice and voice communication with the cellular device and domicile amusement systems, the robust speech recognition that works in all the real-world noises and other acoustic distorting conditions is in demand [29]. To implement the ASR system, some obstacles may occur due to abnormality in speaking style and noises in the environment. The acoustic environment for ASR is much difficult or different than in the past [13]. Despite several technological advancements of the ASR system, there is a huge gap in terms of accuracy and speed in comparison to the human perspective [2]. The main objective behind developing the ASR system is to convert a speech utterance into text sequence, independent of a speaker and the surrounding environment.

***Corresponding Author: A. Kumar:** Research Scholar, Computer Engineering Department, National Institute of Technology, Kurukshetra, Haryana, India; Assistant Professor, Computer Science and Engineering Department, Galgotias University, Greater Noida, Uttar Pradesh, India; Email: anketvit@gmail.com
R.K. Aggarwal: Associate Professor, Computer Engineering Department, National Institute of Technology, Kurukshetra, Haryana, India; Email: rka15969@gmail.com

The feature extraction plays a vital role in the ASR as any loss of useful information cannot be retrieved in later processing stages. There are various number of techniques available to extract the speech features such as Mel Frequency Cepstral Coefficient (MFCC) [12], Perceptual Linear Predictive Analysis (PLP) [20], Gammatone Frequency Cepstral Coefficients (GFCC) [43, 44], Linear Prediction Cepstral Coefficients (LPCC) [49], and wavelet-based feature extraction techniques [45]. Among all these techniques, MFCC is more popular as it shows promising results in clean environment conditions, but the performance of MFCC deteriorates in noisy environmental conditions. Till now, there is no optimal feature extraction technique in the field of ASR. Each feature extraction technique has some advantages and disadvantages. The back-end processing of the ASR system includes the pattern matching of speech features that are stored in memory with the feature set extracted from the test speech signal. The acoustic and language modeling are the two major fields at the back-end processing. For more than four decades, HMM was the first choice for acoustic modeling [18]. The Expectation-Maximization (EM) algorithm is used to train the parameters in HMM Acoustic Modeling. HMM has some serious issues like an inability to train a large amount of training data with no intra-speaker variability [35]. As a result, various other acoustic modeling technique comes into existence such as Gaussian Mixture Model (GMM) [1], Subspace Gaussian Mixture Model (S-GMM) [34], and discriminative techniques. The discriminative techniques like MMI and MPE show significant improvement over Maximum-Likelihood Estimation (MLE). To overcome the problems of Baum Welch (BW) algorithm, Extended-BW (E-BW) algorithm is used in MMI and MPE techniques [36]. Speech recognition and machine translation are the areas in which language modeling plays a vital role [11]. Often, a better language model increases the performance of the underlying system, which makes LM valuable in ASR [22]. In the past few years, RNN and deep learning have fueled language modeling research [6–8]. The majority of work has been done on RNN models, which are capable of retaining long-term dependencies [22]. Speaker adaptation has also become more popular in the last few years [50]. While speaker-dependent (SD) speech recognition systems can show impressive performance, speaker-independent (SI) systems can provide an average Word Error Rate (WER), a factor of two to three lower than an SD system if both systems use the same amount of training data. In speaker adaptation, the small amount of training data from a new speaker is sufficient to adopt the characteristics of new speakers. Adaptation can significantly improve the WER for outlier speakers such as non-native or others who are not well represented in the SI training set.

In this research paper, we used a limited resource Hindi speech dataset (2.5 hours) [41]. For feature extraction, various heterogeneous feature combinations were done in this work, and relative information gain or losses were recorded. We have shown that heterogeneous features performed well and lead to the ASR system with better generalization capability. MPE and MMI discriminative techniques were used to train the acoustic model, which gave significant performance gain. For language modeling, the RNN model was used to train the text data. The implementation of LM was done by RNN LM training toolkit CUED-RNNLM developed at CUED [52]. In this work, MLLR and C-MLLR supervised speaker adaptation were also applied. This paper has three outcomes. First, Integrated acoustic features significantly improve the accuracy over traditional features. Second, it discriminatively trains the integrated feature vector using MMI and MPE discriminative techniques. Third, it applies RNN LM to improve the accuracy of the proposed system further.

The remaining part of the paper is organized as follows: Section 2 explains the concept of different feature extraction techniques, speaker adaptation, discriminative techniques, and RNN LM. Section 3 gives the details of the proposed architecture. Section 4 describes the Hindi speech corpus. Section 5 covers the experiment part of the paper, and section 6 is the conclusion of the proposed system.

2 Preliminaries

2.1 Feature Extraction

2.1.1 Frequency Cepstrum Coefficient (MFCC)

Let $X(n)$ be the input speech signal and frames are blocked and smoothed by applying hamming window $W(n)$. Feature extraction through MFCC involves mainly the following five steps.

i) After performing pre-emphasis step over speech signal, short time Fourier analysis is done using hamming window [28]. To amplify the energy at a higher frequency, pre-emphasis is generally performed [10]. The power spectral estimation is done as:

$$\tilde{x}(k) = \sum_{n=0}^{N-1} x(n)W(n).e^{-J2\pi nk/N} 0 \le n < N \tag{1}$$

where N corresponds to hamming window.

ii) After that power spectrum is passes throgh the Mel-scale triangular filter bank, get the energies of each filter bank as:

$$E_m = \sum_{k=0}^{k-1} \phi_m(k)X_k; m = 1, 2, \ldots, M \tag{2}$$

where M denotes the number of triangular filters.

iii) Discrete Cosine Transform (DCT) is applied to the filterbank energies to get the MFCCs (c_i):

$$c_j = \sum_{m=1}^{M} log_{10}(E_m).cos(j(m + 0.5)\frac{\pi}{M}); j = 1, 2, \ldots, L \tag{3}$$

iv) Append normalized frame energy, producing a 13-dimensional standard feature vector.

v) More features can get by applying first and second derivatives as follows:

$$\frac{\partial c_i}{\partial \tau} = \frac{\sum_t \tau \left(c_i^t\right) - \left(c_i^{-t}\right)}{2.\sum_t t^2} \tag{4}$$

$$\frac{\partial^2 c_i}{\partial \tau^2} = \frac{\sum_t \tau \left(\frac{\partial c_i^t}{\partial \tau} - \frac{\partial c_i^{-t}}{\partial \tau}\right)}{2.\sum_t t^2} \tag{5}$$

where t is time, and $c_i^{(t)}$ and $c_i^{(-t)}$ represents the t^{th} following and previous cepstral coefficients in time frame, respectively.

2.1.2 Gammatone-Frequency Cepstral Coefficient (GFCC)

The MFCC [12] features give promising results in a clean environment, but in a noisy environment, the performance of MFCC decreases. The GFCC [43, 44] features work well in a noisy environment as their model is based on the human auditory system. The GFCC features are determined by Equivalent Rectangular Bandwidth (ERB) scale and set of gammatone filterbanks. We found GFCC features more robust in comparison to MFCC and PLP features [15]. The initial operations of GFCC and MFCC are similar. The output of the fourier transform is passed through gammatone filterbank where center frequency f can be defined as:

$$g(f, t) = a.t^{n-1}e^{-2\pi bt}cos(2\pi ft + \phi) \tag{6}$$

Where a denotes a constant, ϕ represent the phase of the filter, and n denotes the order of the filter. The filterbank bandwidth b factor is denoted as:

$$b = 25.17\left(\frac{4.37f}{1000} + 1\right) \tag{7}$$

After that DCT is performed to get the uncorrelated cepstral features.

2.1.3 Wavelet packet based ERB Cepstral features (WERBC)

The wavelet transforms effectively do the time-frequency analysis in the case of the non-stationary or quasi-stationary signal [4]. Wavelet packets (WP) [3, 26] shows their importance in signal representation schemes such as speech analysis [9]. WP's with the broad coverage of time-frequency characteristics outperform in comparison to standard MFCC features for the speech recognition task.

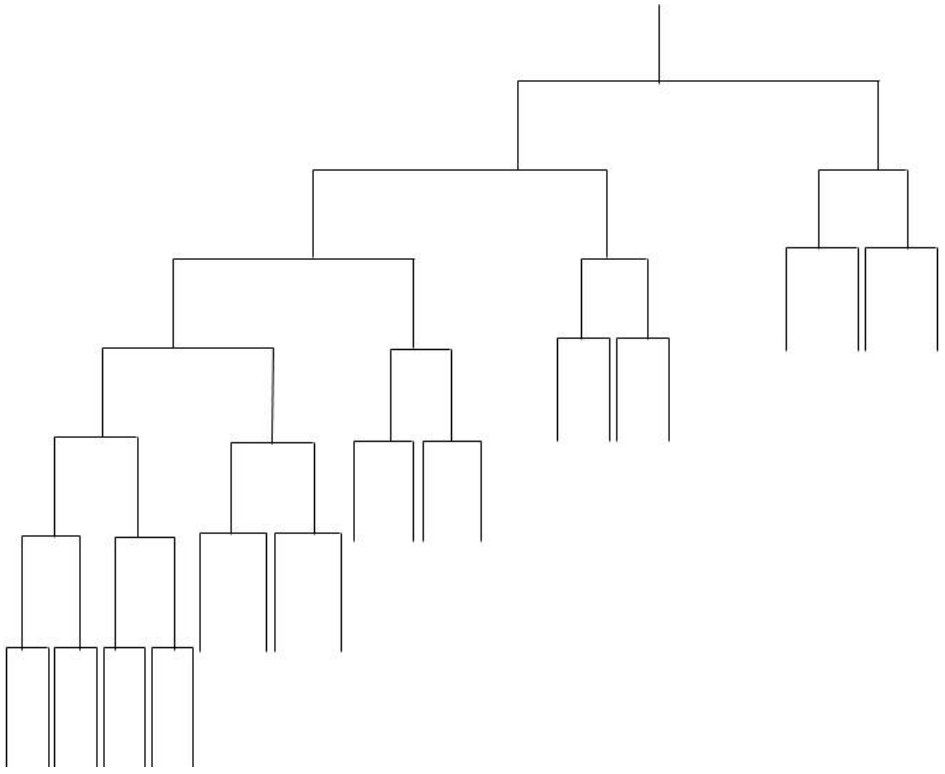

Figure 1: Distribution of center frequency (Hz) to 24 Wavelet Packet subbands

Some amount of research has been done by [9, 16, 21, 38, 39, 45] in Hindi speech recognition using wavelet transformation. The WERBC feature extraction technique is proposed in 2014 [4]. The process of converting the speech signal into WERBC features via admissible wavelet packet transform is shown in Figure 1. The frame size of 25 ms with 70% overlapping is used to extract WERBC features.

After applying the hamming window, the entire frequency band is decomposed with the help of 3-level WP decomposition. It will give 8 subbands of 1 kHz each. Again apply WP decomposition as shown in Figure 1 to get 24 frequency subband.

The first 0-500 Hz frequency band is divided into 8 subbands of 62.5 Hz each. This division finely emphasis a frequency band of 0-500 Hz, which contains a large amount of signal energy. Next, 500-1000 Hz is further decomposed by applying 2 level WP decomposition to get 4 subbands of 125 Hz each. Next, 4 sub-band of 250 Hz is found by applying 2 level WP decomposition on 1-2 Hz frequency band. Similarly, 4 subbands of 500 Hz and 4 subbands of 1 kHz are found subsequently. The equal loudness and log are performed to those 24 subbands to get 24 coefficients. Finally, DCT has been applied to get 12 cepstral features. To get more cepstral features, delta and acceleration coefficients are applied.

Table 1: 24 uniformly spaced wavelet packet sub-band Comparison with center frequencies (Hz) [25]

Filter	Wavelet Subband	Filter	Wavelet Subband	Filter	Wavelet Subband	Filter	Wavelet Subband
1	62.5	**7**	437.5	**13**	1250	**19**	3500
2	125	**8**	500	**14**	1500	**20**	4000
3	187.5	**9**	625	**15**	1750	**21**	5000
4	250	**10**	750	**16**	2000	**22**	6000
5	312.5	**11**	875	**17**	2500	**23**	7000
6	375	**12**	1000	**18**	3000	**24**	8000

2.2 Discriminative techniques

HMMs are the statistical model of speech production [36], whose parameters are optimized by the BW algorithm. Most popular ASR systems are based on statistical-based acoustic modeling. In the past few years, the discriminative technique gets more attention as it further optimizes the HMM parameters to achieve high accuracy [14, 37]. In conventional GMM-HMM based acoustic modeling, HMM parameters are estimated via MLE technique [18]. The MLE technique has the ability to produce an accurate system that is quickly trained using the BW algorithm [24]. The MLE is unbeatable if observations are from the known Gaussian family distribution, training data is unlimited, and the priorly known true language model is available. Unfortunately, these assumptions are not true in case of speech. The discriminative techniques try to optimize the model correctness in such a way to penalize parameters that are responsible to creating confusion between right and wrong predictions [48].

In this work, the baseline acoustic modeling is based on HMM, where the states are represented by Gaussian mixtures, and the discriminative technique is applied over this to optimize the HMM parameters using the E-BW algorithm[19]. In this paper, lattice-based discriminative training is applied by using the HMMIRest tool of HTK3.5 toolkit, and these lattices are generated by a weak language model (e.g., bi-gram) to improve the generalization capability of the discriminative technique. To make the discriminative technique more efficient, phone-marked lattices are used by HMMIRest. For the discriminative technique, HTK uses more than one expected hypothesis for each speech utterance. In this paper, MMI and MPE discriminative techniques are applied to the integrated feature set supported by the HMMIRest tool. Speaker adaptation is also applied in acoustic modeling. Speaker adaptation is the process of modifying the acoustic model parameters by using a small amount of speech data of the specific user in such a way so that the resultant model able to recognize the speech of that speaker. Generally, these techniques are applied to the well trained SI model set to model the characteristics of a new speaker [47]. In many situations, if a large well-defined SI model is used, the baseline SI performance can be quite high. Hence, the error rate gain from speaker adaptation may be smaller than the simpler model. Speaker adaptation techniques can be grouped into two families: i) Linear transformation-based adaptation and ii) Maximum a posteriori (MAP) adaptation [52]. In this work, we only explored the linear transform based adaptation techniques. These techniques estimate the linear transformation from the adaptation data to modify HMM parameters.

2.2.1 Mutual Information (MMI)

The main motive behind the discriminative technique is to assess HMM parameters so as to boost the accuracy of the ASR system. The objective function of MMI to maximize the mutual information on the set of observation is defined as:

$$f_{MMI}(\lambda) = \frac{1}{R}\sum_{r=1}^{R} log\left(\frac{P\left(\frac{O^r}{H^r_{ref}}\right)P\left(H^r_{ref}\right)}{\sum_H P\left(\frac{O^r}{H}\right)P(H)}\right) \tag{8}$$

$P(H^r_{ref})$ denotes the word sequence probability given by LM and H^r_{ref} is the HMM corresponding to the word transcription. The denominator term sum over each possible word sequences. To boost the objective function, the numerator term should be high, and the denominator term should be low. The MMI tries to make the correct hypotheses more probable and incorrect hypotheses less probable at the same time [48].

2.2.2 Minimum Phone Error (MPE)

In MPE, we try to maximize the phone level accuracy rather than maximizing the word accuracy. The MPE training relies on minimum Baye's risk training. The objective function is defined as:

$$f_{MPE}(\lambda) = \sum_{r=1}^{R} \sum_{H} P\left(\frac{H}{O^r}\lambda\right) L\left(H, H^r_{ref}\right) \tag{9}$$

In HMMIRest MPE criterion is expressed as:

$$f_{MPE}(\lambda) = \sum_{r=1}^{R} \sum_{H} \left(\frac{P\left(\frac{O^r}{M_H}\right)}{P\left(\frac{O^r}{M_r^{den}}\right)} \right) L\left(H, H^r_{ref}\right) \tag{10}$$

M_H is a numerator acoustic model, and M_r^{den} is the denominator acoustic model for utterance r. The notation denotes the loss between hypotheses and reference. The main difference between MMI and MPE is lying in how the denominator and numerator lattices are computed. The parameter estimation is based on E-BW algorithm. The loss function is measured by Levenshtein-edit distance. This distance is measured between the phone sequences of the reference and the hypotheses. The MPE technique has been found to improve accuracy as it supports word transcription with corresponding phone accuracy [18].

2.3 Recurrent Neural Network (RNN) Language Modeling (LM)

Mostly, state-of-the-art LM used in LVCSR systems are based on RNN [30, 51]. In various work [31–33], the usefulness of RNN LM has been reported in LVCSR task. The traditional RNN is the three-layer architecture, as shown in figure 2. The first layer, known as the input layer, contains a full history vector by concatenating h_{i-2} and X_{i-1} as input to the hidden layer. For empty history, it is initialized to a vector of all ones. RNN LM encode full, non-truncated history $h^1_{i-1} = [X_{i-1}, \ldots, X_1]$ for current word X_i. The current word X_i is predicted by using 1-of-k encoding of the most recent preceding word X_{i-1} and history context h_{i-2}. Information receives at the hidden layer is further compressed using the sigmoid activation function. It also gives feedback to the input layer. An Out-of-Vocabulary (OOV) node is also added at the input layer to cover those words which are not present in the recognition dictionary. In the third layer, the softmax activation function is applied to produce normalized RNN LM probabilities [52]. The output of this layer is also feedback into the input layer as remaining history to compute LM probability for the next future word.

$$P_{RNN} = \left(X_{(i+1)}/X_i, h_{i-1}\right) \tag{11}$$

The training and decoding are computationally expensive in RNN LM, and major computation is done at the output layer. To reduce the computational cost at the output layer, one more node Out-of-Shortlisted (OOS) is used, which contains the most frequent words. The extension of the back-propagation algorithm, Back-propagation through Time (BPTT), is used to train the RNN LM [40]. In BPTT training, the output error is backpropagated in time for a specific number of time steps. This paper uses a 3-8 word length history. In our work, we use one million Hindi text corpus to train the RNN LM. The texts from the various sources were collected to train the language model. These sources were Emili corpus, magazines, newspapers, web text, and newsletters. One million text corpus is assumed as the medium-sized dataset for conventional n-gram LM, but it is reasonably large for RNN LM. In order to further reduce the computational cost at the output

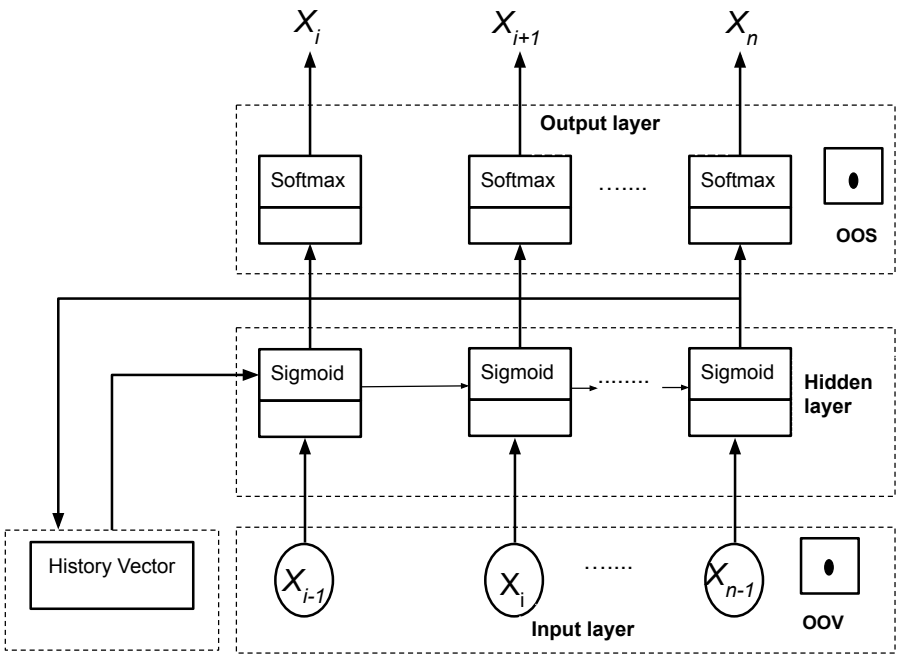

Figure 2: Architecture of Recurrent Neural Network Language Modeling

layer, this toolkit also supports certain other training criteria such as Variance Regularization (VR) and Noise Contrastive Estimation (NCE). Both methods use history independent normalization to increase the RNN LM evaluation speed.

3 Proposed Architecture

The proposed architecture is divided into two parts. The first part describes the process of feature extraction and the integration of different feature sets. In this part of the proposed architecture, feature vectors are generated using GFCC, MFCC, PLP, WERBC feature extraction techniques. The second part of the architecture covers the discriminative training of the feature vector proposed in the first part.

3.1 Proposed integrated feature set

The idea of an acoustic feature combination was initially proposed by Harmansky in 1994 [25]. In his work, he combined PLP features with RASTA features to improve the performance of the targeted ASR system. In the process of speech recognition, feature extraction plays a vital role in achieving high accuracy. Recently, Many papers come into existence to prove their superiority over the previous one by achieving high-performance gain [1, 15, 23, 42, 53, 54]. In the last few years, GFCC [5] and wavelet-based techniques [45] have become more popular as they work well in a noisy environment. This property attracts many researchers to combined these features with other features to improve the accuracy of the ASR system [1, 15, 23, 42, 54].

In this proposed work, the sequential combination of MFCC, GFCC, and WERBC features is done, as mention in Figure 3. The dimension of the feature vector is further reduced by applying Heteroscedastic Linear Discriminant Analysis (HLDA) [27] as HLDA reduces the feature dimension 25%, which helps to reduce the computational load of the ASR system. HLDA is the method of projecting high-dimensional acoustic representation into lower-dimensional spaces [46]. The finding of lower-dimensional representation reduces the

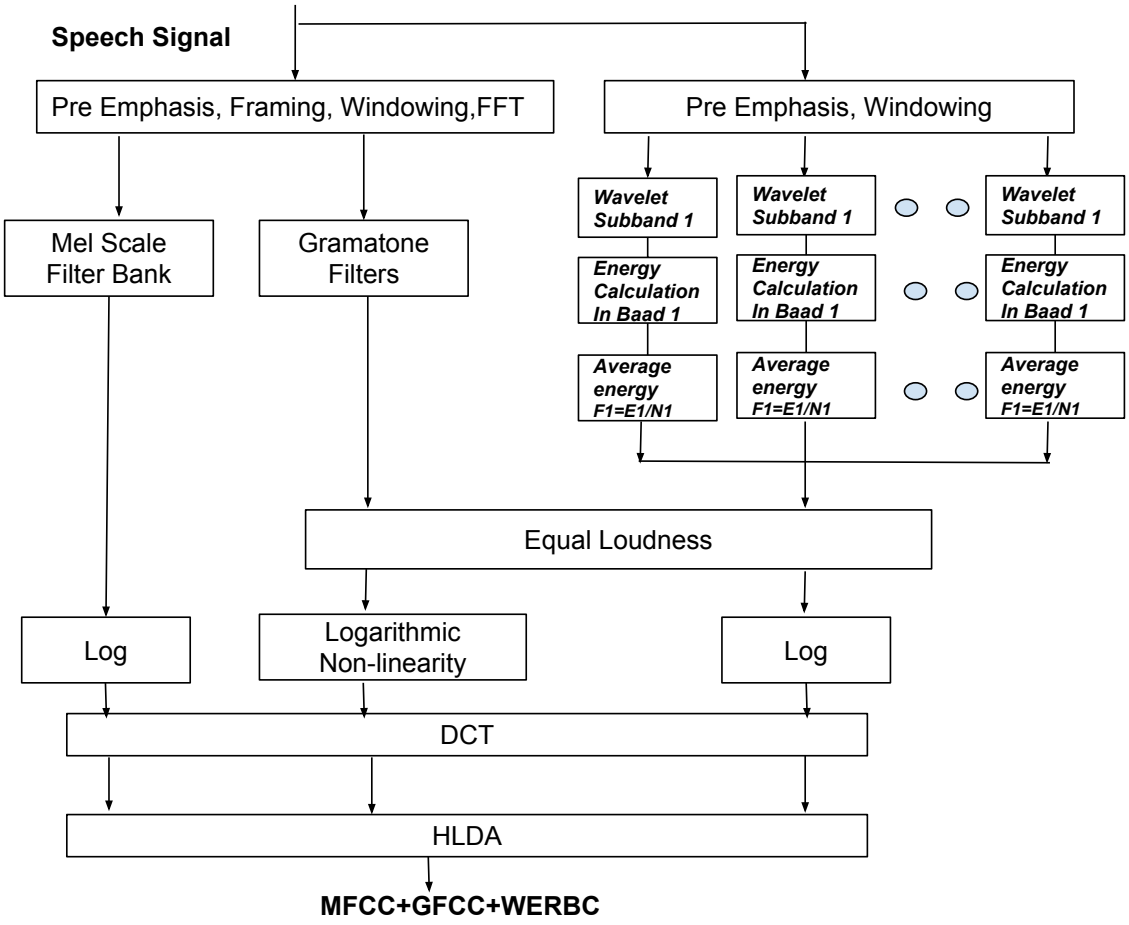

Figure 3: Procedure to obtain proposed integrated feature set (MFCC-GFCC-WERBC)

number of parameters that are used to train the acoustic model and thereby, a significant reduction in computational load.

3.2 Discriminative training

In the second part of the proposed architecture, an integrated feature set is trained using discriminative training approaches. The MMI and MPE discriminative techniques are used in the proposed work. The acoustic modeling is done by the HTK3.5 beta toolkit developed by Cambridge University, USA. To apply the discriminative technique, a cross-word triphone set of HMM's are initially trained using MLE. A weak language model (e.g., bi-gram LM) is the next requirement to apply the discriminative technique. The language model creates the lattice used by MMI and MPE in training.

Two sets of "phone-marked" lattices are required for the discriminative technique known as denominator lattice and numerator lattice. *HDEcode* is used to create the denominator lattice, and numerator lattice is created by the *HLRescore* tool. The numerator lattice includes language model log probabilities, and denominator lattice implements confusable hypotheses [36, 48]. The initial word lattices are further processed to create the phone marked lattice, and these phone marked lattices are used by *HMMIRest* tool to discriminatively trained the HMM. The E-BW algorithm does the parameter estimation. In this work, 4 iterations of each MMI and MPE is done to train the acoustic model.

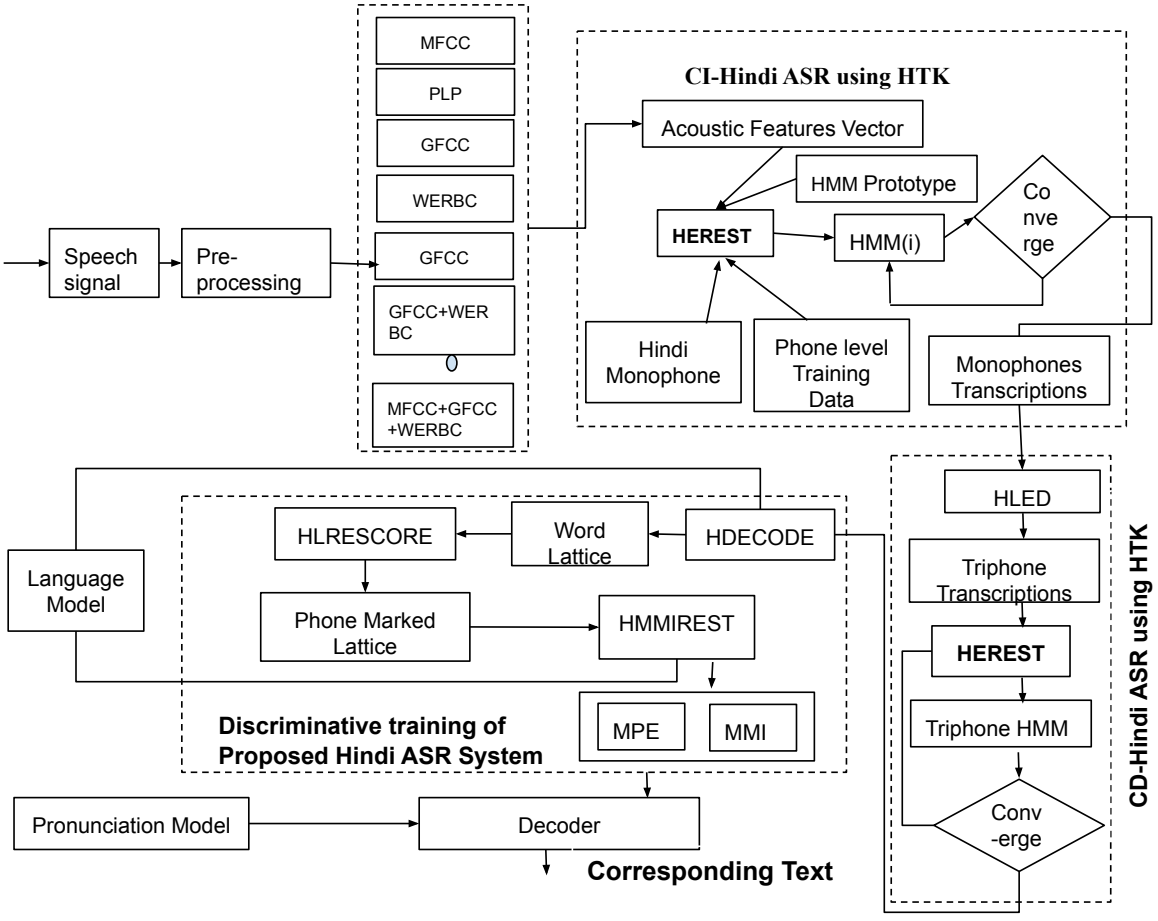

Figure 4: Training and testing procedure of an ASR system

4 Hindi Speech Corpus

Hindi is the fourth most natively spoken language in the world. According to Ethnologue, India have almost 260 million people who use this language. After Chinese and Spanish, English is in third place, with 335 million speakers. Except for Hindi, all these languages have well developed ASR system with their standard dataset. One major challenge for Hindi speech recognition is the deficiency in the Hindi speech dataset and text corpora. In this work, a well-annotated and phonetically rich Hindi dataset is used developed by TIFR, Mumbai [41]. This dataset contains 100 speakers, and each speaker utters 10 sentences out of which two sentences are common to everyone. These two common sentences cover all phonemes of the Hindi language. The next eight sentences also cover the maximum phones of the Hindi language. The recording was done by two microphones in a quiet room on 16 kHz sampling frequency. For training, 80 speakers out of 100 speakers are randomly selected, out of which 55 are male speakers, and 25 are female speakers. The remaining 20 speakers are used for testing purposes.

Table 2: Hindi Language Dataset

Dataset	No. of Speakers	Utterances	Total Words	Unique Words	Hours
Train	80	800	5420	2015	2.1
Test	20	200	1240	856	.20

5 Simulation details and experiment results

The acoustic modeling is done by the new version of the HTK toolkit 3.5. For front-end feature extraction and combination, MATLAB R2015a has been used. The training and evaluation of RNN LM have been carried out using the CUED-RNN LM toolkit [8]. The speech database is divided into two parts: training and evaluation. For training, 80 speakers out of 100 speakers randomly selected, and the remaining 20 speakers left for the testing purpose. The training data is further divided into three parts of 30 speakers each. Set 1 contains only 30 speakers out of 80 who speak Hindi frequently and belong to the northern part of India. Same as Set 2 contains 30 Hindi speakers out of the remaining 50 speakers who belong to the south region of India. Set 3 contains the remaining 20 speakers and the mixture of speakers from set 1 and set 2. Same as the training set, the testing set is also further divided into three parts. Set 1 includes only male speakers of count 12. Set 2 contains 8 female speakers and set 3 contain all 20 speakers for evaluation purposes.

5.1 Performance analysis of multiple feature combination

In this experiment, we continue with the work started in [15]. The baseline GMM-HMM system contains 256 Gaussian mixture per HMM state with tri-phone based acoustic modeling. The comparative analysis of various multiple feature combination techniques using the baseline system has been shown in Table 3. The standard feature set size without integration is 39 in this experiment. In the case of integrated acoustic features, the dimensionality of the feature vector is reduced by HLDA. To get the integrated feature set of MF-GFCC, the first four MFCC features were chosen to integrate with 13 GFCC features. In this way, MF-GFCC makes the set of 17 features. After taking the first and second derivatives, the feature vector size will become 51. These 51 MF-GFCC features are reduced to 39 features after applying the HLDA technique. The same procedure is applied to get the MFCC+GFCC+WERBC features, in which the first four features were taken from MFCC and combined with 13 GFCC and WERBC features (i.e., 30) and take the first and second derivative. In the same way, all other feature combinations have been taken place. The results clearly show that the combination of MFCC+GFCC+WERBC with HLDA transformation outperforms over all other feature combinations. The training set 3 with testing set 3 gives the best results where both male and female speakers come from the north and south region of India in comparison to other combinations. The proposed feature combination shows 9% relative improvement over MFCC based ASR system. The bi-gram LM was used in this experiment.

5.2 System combination

For the detailed study of performance measurement of the proposed integrated feature set with other configuration settings at the back-end, several models have proposed in this experiment. The best three integrated feature sets are choosen from the previous experiment to apply discriminative training with speaker adaptation. Speaker adaptation can significantly improve the WER in the SI training set [52]. It has been observed in the previous work [1] that MLLR will help to achieve low WER as the size of vocabulary increases. In this experiment, we proposed a series of system configurations. The best possible system combination of front-end processing with back-end processing is chosen in this experiment. The proposed baseline system is tested with and without speaker adaptive training (SAT) to measure the performance gain.

The naming convention for the proposed series of the system is done by capital letters indicate the type of feature combination, speaker adaptation (MLLR,C-MLLR), and discriminative training criteria. For example, PG-M MMI indicates PLP+GFCC feature set with MLLR adaptation modeled by MMI discriminative technique. In this experiment, we choose 15 systems based on their feature extraction techniques, model type, discriminative training, no of iteration, and no of the transform for acoustic model adaptation. The acronyms used for the different number of systems helps for further discussion.

Table 3: Comparative analysis of multiple feature combinations

Training Set	Test Set	Features	HLDA	Accuracy%	Training Set	Test Set	Features	HLDA	Accuracy%	Training Set	Test Set	Features	HLDA	Accuracy%
Set-1	Set-1	MFCC	No	63.40	Set-2	Set-1	MFCC	No	67.20	Set-3	Set-1	MFCC	No	62.04
		PLP	No	62.20			PLP	No	64.02			PLP	No	61.86
		GFCC	No	71.05			GFCC	No	69.36			GFCC	No	70.40
		WERBC	No	72.56			WERBC	No	70.56			WERBC	No	71.36
		MFCC+PLP	Yes	65.02			MFCC+PLP	Yes	68.20			MFCC+PLP	Yes	63.40
		MFCC+GFCC	Yes	72.36			MFCC+GFCC	Yes	71.02			MFCC+GFCC	Yes	70.86
		MFCC+WERBC	Yes	73.20			MFCC+WERBC	Yes	71.36			MFCC+WERBC	Yes	7202
		PLP+GFCC	Yes	70.10			PLP+GFCC	Yes	69.02			PLP+GFCC	Yes	6820
		PLP+WERBC	Yes	72.86			PLP+WERBC	Yes	71.02			PLP+WERBC	Yes	71.40
		GFCC+WERBC	Yes	74.02			GFCC+WERBC	Yes	72.02			GFCC+WERBC	Yes	73.20
		MFCC+PLP +GFCC	Yes	74.56			MFCC+PLP +GFCC	Yes	72.86			MFCC+PLP +GFCC	Yes	74.10
		MFCC+GFCC +WERBC	Yes	75.20			MFCC+GFCC +WERBC	Yes	73.20			MFCC+GFCC +WERBC	Yes	75.36
	Set-2	MFCC	No	67.02		Set-2	MFCC	No	67.9		Set-2	MFCC	No	62.90
		PLP	No	65.36			PLP	No	64.86			PLP	No	62.40
		GFCC	No	70.02			GFCC	No	71.40			GFCC	No	71.86
		WERBC	No	71.86			WERBC	No	72.10			WERBC	No	72.20
		MFCC+PLP	Yes	68.02			MFCC+PLP	Yes	70.20			MFCC+PLP	Yes	65.02
		MFCC+GFCC	Yes	72.40			MFCC+GFCC	Yes	72.02			MFCC+GFCC	Yes	72.36
		MFCC+WERBC	Yes	72.56			MFCC+WERBC	Yes	72.96			MFCC+WERBC	Yes	72.86
		PLP+GFCC	Yes	73.36			PLP+GFCC	Yes	72.02			PLP+GFCC	Yes	72.02
		PLP+WERBC	Yes	72.02			PLP+WERBC	Yes	73.40			PLP+WERBC	Yes	72.56
		GFCC+WERBC	Yes	73.02			GFCC+WERBC	Yes	73.56			GFCC+WERBC	Yes	73.20
		MFCC+PLP +GFCC	Yes	73.86			MFCC+PLP +GFCC	Yes	73.96			MFCC+PLP +GFCC	Yes	75.02
		MFCC+GFCC +WERBC	Yes	74.56			MFCC+GFCC +WERBC	Yes	74.20			MFCC+GFCC +WERBC	Yes	76.10
	Set-3	MFCC	No	65.02		Set-3	MFCC	No	70.40		Set-3	MFCC	No	71.20
		PLP	No	64.56			PLP	No	68.20			PLP	No	69.40
		GFCC	No	71.56			GFCC	No	72.56			GFCC	No	73.56
		WERBC	No	72.02			WERBC	No	73.15			WERBC	No	74.04
		MFCC+PLP	Yes	67.56			MFCC+PLP	Yes	69.02			MFCC+PLP	Yes	70.56
		MFCC+GFCC	Yes	72.86			MFCC+GFCC	Yes	73.40			MFCC+GFCC	Yes	74.02
		MFCC+WERBC	Yes	73.20			MFCC+WERBC	Yes	73.86			MFCC+WERBC	Yes	74.86
		PLP+GFCC	Yes	72.02			PLP+GFCC	Yes	73.56			PLP+GFCC	Yes	73.40
		PLP+WERBC	Yes	73.56			PLP+WERBC	Yes	73.96			PLP+WERBC	Yes	74.36
		GFCC+WERBC	Yes	73.86			GFCC+WERBC	Yes	74.25			GFCC+WERBC	Yes	75.36
		MFCC+PLP +GFCC	Yes	74.36			MFCC+PLP +GFCC	Yes	75.10			MFCC+PLP +GFCC	Yes	76.56
		MFCC+GFCC +WERBC	Yes	75.02			MFCC+GFCC +WERBC	Yes	76.02			MFCC+GFCC +WERBC	Yes	77.86

Table 4: System Combination with various parameters

SN	System Name	Features	Model	SAT	Type	Transformation	Discriminative Training	Iterations
1	MPG-M	MFCC+PLP+GFCC	GMM	No	MLLR	1	No	1
2	MPG-M MMI	MFCC+PLP+GFCC	GMM	Yes	MLLR	1	MMI	4
3	MPG-M MPE	MFCC+PLP+GFCC	GMM	Yes	MLLR	1	MPE	4
4	MPG-C MMI	MFCC+PLP+GFCC	GMM	Yes	C-MLLR	2	MMI	4
5	MPG-C MMI	MFCC+PLP+GFCC	GMM	Yes	C-MLLR	1	MPE	4
6	GW-M	GFCC+WERBC	GMM	No	MLLR	1	No	1
7	GW-M MMI	GFCC+WERBC	GMM	Yes	MLLR	1	MMI	4
8	GW-M MPE	GFCC+WERBC	GMM	Yes	MLLR	1	MPE	4
9	GW-C MMI	GFCC+WERBC	GMM	Yes	C-MLLR	1	MMI	4
10	GW-C MPE	GFCC+WERBC	GMM	Yes	C-MLLR	1	MPE	4
11	MGW-M	MFCC+GFCC+WERBC	GMM	No	MLLR	1	No	1
12	MGW-M MMI	MFCC+GFCC+WERBC	GMM	Yes	MLLR	1	MMI	4
13	MGW-M MPE	MFCC+GFCC+WERBC	GMM	Yes	MLLR	1	MPE	4
14	MGW-C MMI	MFCC+GFCC+WERBC	GMM	Yes	C-MLLR	1	MMI	4
15	MGW-C MPE	MFCC+GFCC+WERBC	GMM	Yes	C-MLLR	1	MPE	4

5.3 Performance evaluation of different systems

The choice of the front-end feature combination has a tremendous impact on ASR performance. From the previous experiment, we choose three best feature combinations and evaluate them with a number of different parameters. In this section, the performance is evaluated of all the 15 proposed system described in the previous experiment. Here, again, the training set 3, which contains a mixture of north and south Indian dialects, gives maximum accuracy with the test set 3. Discriminative techniques help to optimize the HMM parameters, which leads to the performance gain. In all experiments, it was observed that discriminative techniques improve the generalization capability of the ASR system. The MGW-C MPE system gives the best performance results of 80.36%, which is ~3% more than the baseline configuration system. The MPE discriminative technique performs slightly better from MMI discriminative technique in all experiments. Speaker adaptation also helps to maintain low WER.

Table 5: Comparative analysis of various system performance

Training Set	Testing Set	System Accuracy %														
		MPG-M	MPG-M MMI	MPG-M MPE	MPG-C MMI	MPG-C MPE	GW-M	GW-M MMI	GW-M MPE	GW-C MMI	GW-C MPE	MGW-M	MGW-M MMI	MGW-M MPE	MGW-C MMI	MGW-C MPE
Set-1	set-1	74.86	75.56	76.02	75.86	76.56	74.20	75.02	76.20	75.56	76.86	75.4	76.86	77.56	77.20	78.02
	set-2	74.02	75.20	76.20	75.96	76.86	73.36	74.20	75.02	74.40	75.56	74.86	75.96	76.86	76.36	77.20
	set-3	74.56	75.86	76.56	76.40	77.02	74.20	75.10	75.96	75.56	76.40	75.20	76.86	77.20	77.02	77.56
Set-2	set-1	72.20	73.36	73.96	74.02	74.20	72.15	72.96	73.56	73.40	74.02	73.40	75.02	75.40	75.36	76.40
	set-2	74.36	75.56	76.02	76.20	76.40	73.86	74.56	75.20	74.96	75.56	74.56	76.02	76.56	77.02	77.20
	set-3	75.20	76.40	77.20	77.02	77.56	74.56	75.20	75.96	75.86	76.40	76.36	77.56	78.20	78.10	78.56
Set-3	set-1	74.36	75.20	75.86	75.86	76.36	73.40	73.86	74.65	74.40	75.02	75.86	77.20	77.86	78.02	78.86
	set-2	75.56	76.86	77.20	77.36	77.56	73.56	73.96	74.86	74.56	75.96	76.36	78.02	78.56	78.40	79.20
	set-3	76.86	77.56	77.96	77.96	77.02	75.86	76.40	77.36	76.86	77.86	78.02	79.20	79.96	79.56	80.36

5.4 Experiment with language modeling

Based on the performance evaluation in the previous section, the best four systems have been selected for this experiment. The performance of the proposed Hindi ASR system is further improved using RNN LM. RNN architecture is suitable to deal with variable-length inputs. RNN LM is well suited to model the sequential

data. By applying RNN LM, the computational load of the system increased but gave the significant performance gain. RNN LM can learn the long-term contextual information within the text. To implement RNN LM, CUED-RNN LM [8] toolkit has been used. The RNN LM experiment uses 500 hidden units and N-best lattice rescoring. The performance is further improved by up to 87.96% using RNN LM. One million text transcription is used from various sources to train the language model. One more observation has been recorded in this experiment that the performance of the ASR system is increased up tri-gram LM only.

Table 6: Performance comparison of n-gram LM with RNN LM

System Name	Bi-gram	Tri-gram	4-gram	RNN LM
MPG-C MPE	78.02%	80.56%	80.2%	84.96%
MGW-M MMI	79.2%	81.02%	80.96%	85.02%
MGW-C MMI	79.56%	81.4%	81.02%	85.96%
MGW-C MPE	80.36%	82.16%	81.56%	87.96%

6 Conclusion

A novel integrated features combination of MF-GFCC with WERBC features are discriminatively trained with RNN language modeling to improve the performance of the Hindi ASR system. For speaker adaptation, MLLR and C-MLLR techniques have been applied, and their corresponding improvements have been recorded. The performance of the proposed Hindi ASR has been evaluated on a different number of parameters. The results conclude that MFCC+GFCC+WERBC features are more robust and give maximum accuracy with MPE discriminative training. The MPE technique show 1% relative improvement over the MMI technique. This work can further be extended by applying these feature combinations to DNN based acoustic modeling, and various optimization techniques with the proposed feature set can also be tested.

References

[1] R. K. Aggarwal and M. Dave, Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system, *Telecommun. Syst.*, **52** (2013), 1-10.

[2] M. A. Anusuya and S. K. Katti, Front end analysis of speech recognition: a review, *International Journal of Speech Technology*, **14.2** (2011): 99-145.

[3] A. Biswas et al., Feature extraction technique using ERB like wavelet sub-band periodic and aperiodic decomposition for TIMIT phoneme recognition, *International Journal of Speech Technology*, **17.4** (2014): 389-399.

[4] A. Biswas et al., Hindi phoneme classification using Wiener filtered wavelet packet decomposed periodic and aperiodic acoustic feature, *Computers & Electrical Engineering*, **42** (2015): 12-22.

[5] W. Burgos, *Gammatone and MFCC Features in Speaker Recognition*, Dissertation, 2014.

[6] X. Chen et al., Improving the training and evaluation efficiency of recurrent neural network language models, in: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015.

[7] X. Chen et al., Efficient training and evaluation of recurrent neural network language models for automatic speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24.11** (2016): 2146-2157.

[8] X. Chen et al., CUED–RNNLM–An open-source toolkit for efficient training and evaluation of recurrent neural network language models, in:*2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016

[9] A. D. Cheveigné et al., Concurrent vowel identification. II. Effects of phase, harmonicity, and task, *The Journal of the Acoustical Society of America*, **101.5** (1997): 2848-2856.

[10] H. P. Combrinck and E. C. Botha, On the Mel-scaled cepstrum, in:*Department of Electrical and Electronic Engineering, University of Pretoria*, Pretoria, South Africa, 1996.

[11] A. Currey et al., Dynamic adjustment of language models for automatic speech recognition using word similarity, in: *Spoken Language Technology Workshop (SLT)*, IEEE, 2016.
[12] S. B. Davis, and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *Readings in speech recognition*, (1990): 65-74.
[13] L. Deng et al., Distributed speech processing in MiPadś multimodal user interface, *IEEE Transactions on Speech and Audio Processing*, **10.8** (2002): 605-619.
[14] M. Dua, R. K. Aggarwal, and M. Biswas, Discriminatively trained continuous Hindi speech recognition system using interpolated recurrent neural network language modeling, In:*Neural Computing and Applications*, (2018): 1-9.
[15] M. Dua, R. K. Aggarwal, and M. Biswas, Discriminative training using noise robust integrated features and refined HMM modeling, *Journal of Intelligent Systems*, (2018).
[16] O. Farooq O, S. Datta, M.C. Shrotriya, Wavelet sub-band based temporal features for robust Hindi phoneme recognition, *International Journal of Wavelets, Multiresolution and Information Processing*, **8.6** (2010):847-59.
[17] M. Ferras et al., Comparison of speaker adaptation methods as feature extraction for SVM-based speaker recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, **18.6** (2010): 1366-1378.
[18] D. Gillick, S. Wegmann, and L. Gillick, Discriminative training for speech recognition is compensating for statistical dependence in the HMM framework, in:*2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2012.
[19] G. Heigold, N. Hermann, S. Ralph, and W. Simon, Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance,In: *IEEE Signal Processing Magazine*, **29.6** (2012): 58-69.
[20] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, *the Journal of the Acoustical Society of America*, **87.4** (1990): 1738-1752.
[21] K. Ishizuka, and T. Nakatani, A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition, *Speech communication*,**48.11** (2006): 1447-1457.
[22] R. Jozefowicz et al., *Exploring the limits of language modeling*,preprint(2016), http://arxiv.org/abs/1602.02410.
[23] V. Kadyan, A. Mantri and R. K. Aggarwal, A heterogeneous speech feature vectors generation approach with hybrid hmm classifiers, *Int. J. Speech Technology*, **20.4** (2017): 761-769.
[24] S. Kapadia, *Discriminative training of hidden Markov models*, Doctoral dissertation, University of Cambridge, 1998.
[25] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch and G. Tong, Integrating RASTA-PLP into Speech Recognition, in: *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*,**1** Adelaide, SA, Australia, 1994.
[26] R. Kumar, A. Kumar, and R. K. Pandey, Beta wavelet based ECG signal compression using lossless encoding with modified thresholding, *Computers & Electrical Engineering*, **39.1** (2013): 130-140.
[27] N. Kumar and A. G. Andreou, Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition, *Speech Commun*, **26.4** (1998), 283-297.
[28] A. G. Kunkle, *Sequence scoring experiments using the TIMIT corpus and the HTK recognition framework*, Dissertation, Florida Institute of Technology, Florida, USA, 2010.
[29] J. Li et al., An overview of noise-robust automatic speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22.4** (2014): 745-777.
[30] K. Li et al., Recurrent neural network language model adaptation for conversational speech recognition, In: *INTERSPEECH*, Hyderabad (2018): 1-5.
[31] T. Mikolov et al., Recurrent neural network based language model, In:*Eleventh annual conference of the international speech communication association*,(2010).
[32] T. Mikolov et al., Extensions of recurrent neural network language model, In:*2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,IEEE, (2011):5528-5531.
[33] T. Mikolov et al., Context dependent recurrent neural network language model, In:*2012 IEEE Spoken Language Technology Workshop (SLT)*,IEEE, (2012):234-239.
[34] A. Mohan et al., Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain, *Speech Communication*, **56** (2014): 167-180.
[35] J. M. Naik, L. P. Netsch, and G. R. Doddington, Speaker verification over long distance telephone lines, in:*International Conference on Acoustics, Speech, and Signal Processing ICASSP-89*, IEEE, 1989.
[36] D. Povey, *Discriminative training for large vocabulary speech recognition*, PhD Diss. University of Cambridge, 2005.
[37] D. Povey et al.,Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI, In: *Interspeech*, (2016): 2751-2755.
[38] S. Ranjan. A discrete wavelet transform based approach to Hindi speech recognition, In:*2010 International Conference on Signal Acquisition and Processing*, Bangalore(2010):345-348.
[39] S. Ranjan, Exploring the discrete wavelet transform as a tool for Hindi speech recognition, *International Journal of Computer Theory and Engineering*, **2.4** (2010): 642.
[40] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors, *nature*, **323.6088** (1986): 533.
[41] K. Samudravijaya, P. V. S. Rao and S. S. Agrawal, Hindi speech database, in: *International Conference on spoken Language Processing*, Beijing, China, 2002, pp. 456–464.

[42] R. Schluter et al., Gammatone features and feature combination for large vocabulary speech recognition, in:*IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, **4** IEEE, 2007.
[43] Y. Shao, and W. DeLiang, Robust speaker identification using auditory features and computational auditory scene analysis, In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*,IEEE, (2008):1589-1592.
[44] Y. Shao et al., A computational auditory scene analysis system for speech segregation and robust speech recognition, *Computer Speech & Language*, **24.1** (2010): 77-93.
[45] A. Sharma et al., Hybrid wavelet based LPC features for Hindi speech recognition, *International Journal of Information and Communication Technology*, **1.3-4** (2008): 373-381.
[46] N. Singh-Miller, Natasha, M. Collins, and T. J. Hazen, Dimensionality reduction for speech recognition using neighborhood components analysis, in: *Eighth Annual Conference of the International Speech Communication Association*, 2007.
[47] A. Stolcke et al., MLLR transforms as features in speaker recognition, in:*Ninth European Conference on Speech Communication and Technology*, 2005.
[48] K. Vertanen, An overview of discriminative training for speech recognition, *University of Cambridge*, Cambridge, UK (2004).
[49] E. Wong, and S. Sridharan, Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification, in: *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing*, IEEE, 2001.
[50] Z. Wu et al., A study of speaker adaptation for DNN-based speech synthesis, in:*Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
[51] D. Yogatama et al., Memory architectures in recurrent neural network language models, in: *Seventh International Conference on Learning Representations*, (2018).
[52] S. Young, G. Evermann, M. Gales,T. Hain, D. Kershaw,X. Liu,V. Valtchev, *The HTK book*, Cambridge University Engineering Department, vol 3, pp 1–285,2002.
[53] X. Zhao and D. L. Wang, Analyzing noise robustness of MFCC and GFCC features in speaker identification, in: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013
[54] A. Zolnay, R. Schluter, and H. Ney., Acoustic feature combination for robust speech recognition, in: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*,**1** IEEE, 2005.