Journal of Engg. Research Vol. 7 No. (4) December 2019 pp. 96-114

Linear Collaborative Discriminant Regression and Cepstra Features for Hindi Speech Recognition

U. G. Patil*, S. D. Shirbahadurkar** and A. N. Paithane***

*JSPM Rajarshi Shahu College of Engineering, SPPU University, Tathawade, Pune, Maharashtra, India
**Dr. D. Y. Patil College of Engineering, SPPU University, Pimpri-Chinchwad, Maharashtra, Pune, India
***JSPM Rajarshi Shahu College of Engineering, SPPU University, Pune, India
*Corresponding Author: patil.ug47@gmail.com

ABSTRACT

Speech recognition system is one of the significant, yet challenging systems in computer-human interaction. Recognizing Indian languages, especially Hindi, faces many practical difficulties due to its wide grammatical and phonetic features from English. This paper focuses on Hindi speech recognition system for which Cepstra features and linear collaborative discriminant regression (LCDR) model are proposed for feature analysis and recognition. For definite audio signals, two models of test speech signals are synthesized and experimental investigations are carried out. The performance of the LCDR methods is analysed using Type I and II error functions and compared with the existing methods such as NN_2-cepstra and SVM_2-cepstra. Moreover, the best, worst, mean, median, and standard deviation are used for the statistical prediction and the proposed LCDR method is proved as the superior method for recognising Hindi speech.

Keywords: LCDR; cepstra; Hindi; speech; recognition.

INTRODUCTION

Automatic Speech Recognition (ASR) is a subsystem or a component of utmost importance in the human-computer interaction system, which accepts decoded text input to enable further operations (Zhang & Fung, 2014a). The two components that constitute the ASR primarily are (i) the language model and (ii) the acoustic model (Kumar *et al.,* 2004). The acoustic model deals with the modeling of the pronunciation that is related to an input word. On the contrary, the language model envisages the possibility of occurrence of the input word sequence in any language kind (Kumar *et al.,* 2004). In case of the acoustic model, the speech signal features, as well as an approach for pattern matching an input word or phone, serve as the major components comprising it (Jeong, 2012). ‘Phone’ indicates the fundamental unit of speech, whereas a word is formed of phones that can be a single phone or multiple phones (Kumar *et al.,* 2004).

Normally, the Perceptual Linear Predictive (PLP) Coefficients (Hermansky, 1990) and the Mel-Frequency Cepstral Coefficients (MFCC) (Davis & Mermelstein, 1980) are the features of frequent use in ASR, though domain transformation succeeds in other applications (Paithane & Bormane, 2014; Paithane & Bormane, 2015; Paithane *et al.,* 2014). Further, the regularly used approaches for pattern matching in ASR are the neural networks and the Hidden Markov Model (HMM) (Kumar *et al.,* 2004 ; Rabiner, 1989). The HMM undergoes a learning on the speech signal’s sequential nature to enable the modeling of the output probability distribution along with the state transition probability (Liu & Sim, 2014). At the time of speech recognition, a hypothesization process is done between several words and the signal being acquired. The matching procedure in HMM involves the computation of likelihood, which is related with a word being provided. For a word, the estimation of likelihood relies on the use of the joint likelihood of the entire number of phones that has relation to the word (Kumar *et al.,* 2004). In previous years, the HMM training was largely accomplished through the maximum likelihood estimation (ML) (Simand & Gales, 2006). Yet, nowadays, the ML training was found to be less hopeful than the discriminative training (Simand & Gales, 2006; Evermann *et al.,* 2004 ; Kim *et al.,* 2003; Donough & Waibel, 2003; Goel *et al.,* 2003).

The language model, in contrast, allows assessing the likelihood of occurrence of a given word sequence from the speech signal. The language model that is extremely used to envisage the probability of occurrence of a word from a sequence is the N-gram language model, and it makes utilization of the history of word sequences to do so. A massive text corpus, which is provided during the training phase, supports the computation of probability. In order to deal with the hypothetical word occurrences, the scores that result from the language, as well as the acoustic models, are exploited. A word is recognized to be an isolated word if its associated likelihood is the largest of all the joint likelihoods of the entire number of words (Kumar *et al.,* 2004). Yet, the performance that the traditional N-gram language models exhibits is found to lag behind the performance that is experienced from the Neural Network Language Models (NNLMs) (Bengio *et al.,* 2003; Schwenk & Gauvain, 2005; Schwenk, 2007; Le *et al.,* 2013; Mikolov *et al.,* 2010; Mikolov *et al.,* 2011). Improvisation in the ASR decoding process has been achieved using less number of hybrid models as well (Arisoy *et al.,* 2014; Arisoy *et al.,* 2013).

There are umpteen numbers of applications available, which requires support from the speech recognition system, and they include the field of commerce, automobiles, military healthcare, and many more areas. Hence, the robustness that the speech recognition system imparts to such kind of applications requires utmost attention. Speech recognition systems employing English have been under wide usage all around the globe. But, Hindi serves as the official language in India with the fact that it ranks second in the group of languages spreading far and wide, subsequent to English. The 2001 census report reveals the percentage of Hindi speaking Indians as 41%, which is fairly larger than the percentage of Indians speaking the rest of the languages. The language, which is largely spoken and ranks second in India, is Bengali. But, only 8.1% of Indians do speak it (Kalinli *et al.,* 2010). Therefore, all the applications concerned with government or commerce highly seek the evolution of speech recognition systems. Nevertheless, researches that are related to the recognition of Hindi speech are not exceedingly prevalent (Kumar *et al.,* 2004). We have thoroughly examined and discovered the truth that the literature lacks witness regarding the recognition of Hindi speech. The reason may be the complete variation of the syllables or the pronunciations from the English language. As an illustration, consider the acoustic classes that are unique to Hindi such as the stress plosives as well as the nasalized vowels. It can be understood that the design as well as the development of speech recognition system, which is solely attributed to Hindi, is of much importance.

This paper introduces a robust speech recognition system, which aids in Hindi speech recognition and word level recognizer is designed. The speeches are subjected to Cepstra analysis to distinguish them through its spectral features. Subsequently, a recently introduced regression classifier called as LCDR is exploited to recognize the speech based on its cepstra features. The paper is organized as follows: Section 2 gives a brief review about the works related to the speech recognition system and the drawbacks are noted, and Section 3 describes the feature analysis for Hindi speech. In Section 4, the speech recognition model is proposed and Section 5 analyzes the obtained results and Section 6 concludes the paper.

LITERATURE REVIEW

Related Works

Numerous speech recognition systems have been reported in the literature. However, a few relevant and significant works are reviewed here.

Khe Chai Sim and M.J.F. Gales (2006) have examined whether the precision matrix models could be possibly employed to perform the discriminative training of speech recognition systems. In addition, they have dealt with the issues concerned with the development of LVCSR systems and the utilization of minimum phone error condition. The approximation of the precision matrices was attained using a generic framework that incorporates a number of conventional models, which include the Subspace for Precision and Mean (SPAM) models, the Extended MLLT (EMLLT), and the Semi-Tied Covariance (STC). To carry out their testing, they have made use of English news broadcast as well as the non-stop telephone conversation with huge vocabulary.

Weibin Zhang and Pascale Fung 2013 (JM Baker, 2009) have trained the acoustic models involving inadequate training data through suggesting the sparse inverse covariance matrices. Here, the conventional objective function that is associated with ML estimation has undergone an enhancement with the inclusion of L1 regularization. The inverse covariance matrices have been sparsely met through the novel objective function. In addition, the parameters that are associated with HMM have been trained through the novel objective function exploiting the Expectation Maximization (EM) algorithm. The procedure for ML estimation as well as training looks identical. The testing was done with the help of a low resource language data, namely, the Cantonese dataset. The testing results have confirmed that the Sparse inverse covariance matrix could render improved performances, when compared against the performances that are yielded from the full covariance model or the diagonal covariance model.

Later, Weibin Zhang, and Pascale Fung 2014 (2014a) have employed the acoustic models involving sparse inverse covariance matrices in place of the diagonal covariance matrices or the conventional full covariance matrices. This replacement of the models has caused the discriminative training approaches to be enhanced. The inverse covariance matrices gain their sparsity through the inclusion of the lasso regularization term in the conventional objective function, while it extorts the maximum mutual information (MMI). Maximization of the enhanced novel objective function has resulted in the accomplishment of the training process. The ability of the sparse inverse covariance matrix in tackling the over-fitting issue, which frequently occurs during the discriminative training, has been proved with the aid of the Wall Street Journal as well as the Mandarin datasets.

Table 1. Merits and demerits of the existing methods.

Methodology	**Merits**	**Drawbacks**	**Year**
MPE Training of Precision Matrix Models (Simand & Gales, 2006)	Suitable for LVCSR environment	Suffers from over-fitting problem	2006
Dynamic Features in the Linear-Logarithmic Hybrid Domain (Kolossa et al., 2013)	Effective under reverberant environment	Noise impact affects performance	2010
Noise Adaptive Training (Zhang & Fung, 2014b)	Suitable for Noisy environment	Not asserted for LVCSR environment	2010
Sparse Inverse Covariance Matrices for Low Resource Speech Recognition (JM Baker, 2009)	Handles even low resource dataset and provides regularized training of speech	Investigated under no noise environment	2013
Noise-Adaptive Linear Discriminant Analysis (Zhang & Fung, 2013)	Suitable for Noisy environment	Not asserted for LVCSR environment	2013
Discriminatively Trained Sparse Inverse Covariance Matrices (Zhang, & Fung, 2014a)	Solves Over-fitting problem	Investigated under no noise environment	2014
Sparse Banded Acoustic Models (Tanja & Alex, 2001)	Good regularization and computational speed	Investigated under no noise environment	2014
Semi-Parametric Trajectory Model for HMM (Liu & Sim, 2014)	Proven on LVCSR under noisy environment	Computationally complex	2014
Derived Back-off Language Models from Neural Network Language Models (Arisoy et al., 2014)	Proven under LVCSR environment	Tested under no noise environment and suffers due to over-fitting	2014

Their approach has resulted in the regularization of the complexity associated with the model, in addition to enhancing the accuracy of recognition. The results of the traditional full covariance models as well as the diagonal covariance models have been found to lag behind the results of the acoustic models employing sparse inverse covariance matrices.

Weibin Zhang and Pascale Fung 2014 (2001) have successfully achieved the training of the sparse banded acoustic models through suggesting the weighted lasso regularization scheme. They have stated few features, which help in cutting down the bandwidth associated with the sparse-banded models, with the aim of improving the computational speed. The results of sparse banded models have orderly shown 15.1% increase and 9.5% increase, when compared orderly with the full covariance models and diagonal covariance models.

Osamu Ichikawa *et al.* 2010 (Kolossa *et al.,* 2013) have suggested the ways of representing the speech signals, which are subjected to reverberation, through the utilization of logarithmic delta. Approaches that compute delta along with the features that exist between the deltas were put forth in their work, in order to cancel out the effects of reverberation from the speech recognition systems. Further, their schemes were supported with newly suggested dynamic features. The testing platform has involved the Corpus as well as the Environments for Noisy Speech REcognition-4 (CENSREC-4) database, which are subjected to reverberation influences. A decrease in the dominant error was observed with the utilization of dynamic features in place of the conventional features.

Ozlem Kalinli *et al.* 2010 (Zhang & Fung, 2014b) have suggested a noise adaptive training (NAT) algorithm with the intention of dealing with the acoustic models, which are adversely affected by noises. In NAT, the 'pseudo-clean' model parameters were assessed in a straight forward manner, instead of taking the point estimates related to the clean speech features. Once the pseudo-clean model parameters were learned, they allow the noisy utterances that occur during testing to be decoded in combination with the vector Taylor series (VTS) model adaptation. When compared against the conventional VTS model adaptation, the NAT has caused improvements of about 32.03% and 18.83% in Aurora 3 and Aurora 2 tasks in a respective manner.

Dorothea Kolossa *et al.* 2013 (Zhang & Fung, 2013) have come up with a rule, which supports ASR that operates in a non-artificial or noisy environment. According to their strategy, "Reducing the dimensionality of the speech feature for optimal discriminance under observation uncertainty can yield significantly improved recognition performance". While finding ways to achieve their strategy, they have discovered the fact that Fisher's principle of discriminant analysis could hold good.

Shilin Liu and Khe Chai Sim 2014 (Liu & Sim, 2014) have analyzed the problems related to the Standard HMM. They have found that HMM results in poor trajectory model for speech due to the notion, "successive observations are independent to one another given the state sequence". Usually, the techniques offering semi-parametric trajectory modeling are capable of dealing with speech recognition tasks, which involve massive and non-stop vocabulary. Hence, Temporally Varying Weight Regression (TVWR) in combination with time-varying Gaussian weights has been intentionally employed for modeling the HMM trajectory in an implicit manner. An in-depth formulation of the Temporally Varying Weight Regression (TVWR) in accordance with the probabilistic modeling framework was portrayed here. The estimation of parameters has been attained in compliance with the phenomena of ML as well as the minimum phone error (MPE). The testing outcomes have outweighed the results of the standard HMM systems, when 20k open vocabulary recognition task (NIST Nov'92 WSJ0) and 5k closed vocabulary noisy speech recognition task were employed.

Ebru Arısoy *et al.* 2014 (2014) have employed non-complex language models to put forward an approximation approach that supports NNLMs. It was possible to transform a feed forward NNLM into a back-off n-gram language model with the suggested approximation approach. By doing so, usage of NNLMs in conventional LVCSR decoders can also be encouraged. Their testing of the back-off models on Broadcast News data has shown enhancements in both accuracy as well as gain, when compared against the traditional n- gram language models.

Problem Statement

Presently, more number of research works has been devoted to the recognition of spontaneous speech like, meetings, lectures, and telephone conversations (Akita & Kawahara, 2010). These works are nothing but an extension of the large-vocabulary continuous speech recognition (LVCSR). Despite the fact that the ASRs are rendered with sufficient enhancements, more numbers of massive challenges are still posed on them (Baker *et al.,* 2009a; Baker *et al.,* 2009b). As an illustration, consider the performance exhibited by the speech recognition systems of present day. It depends on the amount of training data that is being provided to the ASR. But gathering and transmitting massive data is impractical that it becomes even more complex and expensive in case of the training in acoustic modeling (JM Baker, 2009). The training in acoustic model and the speech database carry a trade-off all the time because the training of acoustic model necessitates the speech database to be arranged phonetically. Yet the arrangement of speech database in an automated way is possible only with the deployment of an acoustic model. If an enormous speech database is manually aligned, there will be wastage of time and heavy inaccuracy occurs (JM Baker, 2009). The most basic and chief need of ASR is to make it robust enough to noise (Zhang & Fung, 2013). This implies that the ASR should consider its signal of interest alone and eliminate all the other acoustic interferences (Kolossa *et al.,* 2013; Ichikawa *et al.,* 2010). All these disturbances would majorly result from the background noise only (Rennie *et al.,* 2010). The remaining challenges would have resulted from the training or testing of the pattern mismatch (Rose *et al.,* 1994; Hasan & Hansen, 2014), which are caused from factors like vocal effects (Hasan & Hansen, 2013; Fan & Hansen, 2011). Normally, in case of languages, excluding English, a mapping between the phone models as well as the English phone models is done in prior to recognition (Zhang & Hansen, 2011). But the ASR to recognize Hindi speech necessitates a group of certain utterances of isolated monosyllabic data (Kumar *et al.,* 2004).

The literature survey has revealed the attention of researchers towards developing an automatic speech recognition system, owing to the drawbacks existing in the conventional methods. The merits and demerits of the existing methods are tabulated in Table I. The objectives of them vary because of the wide research gaps that persist. For instance, few works have focused on considering LVCSR system, which is nearer to real environment. However, they fail to consider the impact of noise (Arisoy *et al.,* 2014). In contrast, works that have considered noisy environment did not consider LVCSR system (Zhang & Fung, 2014b; Zhang & Fung, 2013). Though the works have been focused on fast processing (Tanja & Alex, 2001) and solve over-fitting problem (Simand & Gales, 2006; Arisoy *et al.,* 2014), they are not robust against noise (Zhang & Fung, 2014a; Kolossa *et al.,* 2013; Zhang & Fung, 2014b; Tanja & Alex, 2001; Arisoy *et al.,* 2014). Few works have not focused well on improving the decoding speed, although they have worked on LVCSR (Liu & Sim, 2014). On the other hand, over-fitting problem occurs in a recognition system, which has considered LVCSR context (Simand & Gales, 2006). These are the significant research gaps that are extensively available in the existing speech recognition systems.

Instead of addressing all the aforesaid challenges, our ASR considers over-fitting problem of the supervised models. Hence, we introduce the LCDR for recognizing the speech signal. Though LCDR can be operated as a supervised model, our paper considers it as unsupervised to avoid over-fitting problem.

FEATURE ANALYSIS FOR HINDI SPEECH

Cepstra Analysis

Let us consider a speech sequence $x[n]: n = 1,2,\ldots,N$ of length *N*. The complex cepstra feature is widely used in the field of exponential sequences and its transform pair is given as

$$x[n] = \begin{cases} 0, & n < 0 \\ \beta^{n}, & n \ge 0 \end{cases} \tag{1}$$

$$\hat{x}[n] = \begin{cases} 0, & n \le 0 \\ \dfrac{\beta^n}{n}, & n \ge 1 \end{cases} \tag{2}$$

For $y[n]$, the complex cepstrum is given as

$$\hat{y}[n] = \begin{cases} 0, & n \le 0 \\ \omega \dfrac{\beta^n}{n}, & n \ge 1 \end{cases} \tag{3}$$

The above equation is the scaled version of eq. (2) and this equation can be subjected to z transformation to form

$$\hat{X}(z) = -\ln\left(1 - \beta z^{-1}\right) \tag{4}$$

The value of $y[n]$ is equal to the Sheffer polynomial set if it satisfies

$$\sum_{n=0}^{\infty} b_n(\omega) z^{-n} = \exp\left(-\omega \ln\left(1 - \beta z^{-1}\right)\right) = \frac{1}{\left(1 - \beta z^{-1}\right)^{\omega}} \tag{5}$$

Eq. (5) can be satisfied by polynomials given in (Rainville, 1960)

$$b_n(\omega) = \frac{\beta^n}{n!} (\omega)_n \tag{6}$$

where, $$(\omega)_n = \begin{cases} 1, & n = 0 \\ \omega(\omega+1)\ldots(\omega+n-1), & n > 0 \end{cases} \tag{7}$$

Eq. (5) and (6) combine to form a transform pair

$$y[n] = \begin{cases} 0, & n < 0 \\ \dfrac{(\omega)_n}{n} \beta^n, & n \ge 0 \end{cases} \tag{8}$$

$$\hat{y}[n] = \begin{cases} 0, & n \le 0 \\ \omega \dfrac{\beta^n}{n}, & n \ge 1 \end{cases} \tag{9}$$

If $\omega = 1$, the above equation changes to eq. (2). The z -transform of $y[n]$ is taken from eq. (5). If ω is a fraction, then $y[n]$ will be a fractional order signal. So, eq. (8) and (9) indicate the fractional order signals building blocks. Using those eqs.,

$$y[n] = \frac{\beta}{n} (\omega + n - 1) y[n-1], n \ge 1 \tag{10}$$

where $y[0] = 1$, the $y[n]$ in eq. (8) can be calculated. Let $\hat{y}_1[n] = -\omega\beta^n u[n-1]$ and $\hat{y}_2[n] = -4\omega n\beta^n u[n]$. With the Sheffer polynomial sets, the z transforms of $\hat{y}_2[n]$ and $\hat{y}_1[n]$ can be calculated as

$$\hat{Y}_1(z) = \frac{-\omega\beta z^{-1}}{1 - \beta z^{-1}} \quad \hat{Y}_2(z) = \frac{-4\omega\beta z^{-1}}{\left(1 - \beta z^{-1}\right)^2} \tag{11}$$

When $z = \beta$, the $\hat{Y}_2(z)$ and $\hat{Y}_1(z)$ had a second order pole and simple pole, respectively, but when $z = 0$, the function will attain zero. The Sheffer polynomial set is given as

$$\sum_{n=0}^{\infty} L_n(\omega) z^{-n} = \frac{1}{1-z^{-1}} \exp\left(\frac{-4\omega z^{-1}}{\left(1-z^{-1}\right)^2} \right) \tag{12}$$

$$\sum_{n=0}^{\infty} f_n(\omega) z^{-n} = \frac{1}{1-z^{-1}} \exp\left(\frac{-4\omega z^{-1}}{\left(1-z^{-1}\right)^2} \right) \tag{13}$$

where $f_n(\omega)$ and $L_n(\omega)$ represent the Fasenmyer and Laguerre polynomials (Rainville, 1960). Replace z by $\beta^{-1}z$ and multiply $1 - z^{-1}$ on both sides of eq. (12) and (13), so that we receive

$$\sum_{n=0}^{\infty} \beta^n \left[L_n(\omega) - L_{n-1}(\omega) \right] z^{-n} = \exp\left(\frac{-\omega\beta z^{-1}}{\left(1-\beta z^{-1}\right)} \right) \tag{14}$$

$$\sum_{n=0}^{\infty} \beta^n \left[f_n(\omega) - f_{n-1}(\omega) \right] z^{-n} = \exp\left(\frac{-4\omega\beta z^{-1}}{\left(1-\beta z^{-1}\right)^2} \right) \tag{15}$$

where $L_{-1}(\omega) = f_{-1}(\omega) = 0$. So, the sequences get changed to

$$y_1[n] = \beta^n \left[L_n(\omega) - L_{n-1}(\omega) \right] \tag{16}$$

$$y_2[n] = \beta^n \left[f_n(\omega) - f_{n-1}(\omega) \right] \tag{17}$$

Three-term recurrence relation and four-term recurrence correlations are used for computing the Lnaguerre and Fasenmyer polynomials. Therefore, $y_1[n]$ and $y_2[n]$ can be computed recursively with the provided β and ω.

Cepstra versus other features

The stationary Gaussian random process has been used to calculate the speech data one frame (~30 ms). Let $x = [x(0) \ldots x(n-1)]^T$ with discrete time and positive spectrum $s(f), 0 \le f < 1$ has been involved in real-valued Gaussian zero-mean stationary random process. If the time lag is more than n, then the value is set to be zero for the covariance function of the random process. The value of $c_M \in R^m$, MFCC of Gaussian process has to be calculated and so,

$$c_M \underset{=}{\Delta} \frac{1}{m} \Phi^H \log(Ms) \tag{18}$$

Φ, H, s and $M \in R^{m \times n}$ represents the m by m Fourier matrix with the element $(a,b): \Phi \underset{=}{\Delta} \left\{ e^{-i2\pi(a-1)(b-1)/m} \right\}_{ab}$, conjugate transpose, symmetrical spectrum vector, where $s = \left[s\left(0/n\right) \quad s\left(0/n\right) \ldots s\left(\left(n-1\right)/n\right) \right]^T$, and frequency warping filter bank, respectively. In the above equation, the log operates element-wise. The M attained the same symmetry of spectrum vectors and so, it is chosen as the filter bank. Because of this symmetry, eq. (1) seems to be real valued and computation can be done effectively with the discrete cosine transform (DCT). If the filter bank M is a n by n matrix, then the c_M will be minimized to ordinary cepstrum.

Let the MFCC estimator be $\hat{c}_M = \frac{1}{m}\Phi^H \log(M\hat{s})$ (19)

where $\hat{s}$ represents the multitaper spectrum estimator (Percival, & Walden, 1993; Thomson, 1982), and it can be represented as

$$\hat{s} = [\hat{s}(0) \quad \hat{s}(1) \; \ldots \; \hat{s}(n-1)]^T \tag{20}$$

and $\hat{s}(p) = \sum_{j=1}^{k} \lambda(j) \left| \sum_{t=0}^{n-1} w_j(t) x(t) e^{-i2\pi tp/n} \right|$ (21)

$$\hat{s}(p) = \sum_{j=1}^{k} \lambda(j) \left| w_j^T \Psi_p X \right|^2 , \; p = 0,....,n-1 \tag{22}$$

where k and Ψ_p represent the multitapers and n by n diagonal Fourier matrix, which can be given as

$$\Psi_p \underline{\underline{\Delta}} diag\left(\left[e^{-i2\pi p\frac{0}{n}} \quad e^{-i2\pi p\frac{0}{n}} \quad \ldots \; e^{-i2\pi p\frac{n-1}{n}} \right]^T \right) \tag{23}$$

Being a weighted average of k sub-spectra, $W_j^T \Psi_p X|^2$, $j = 1,....,k$, the multitaper estimate is determined and this type of estimation will minimize the variance, because each subspectrum is uncorrelated in a multitaper (Percival, & Walden, 1993 ; Thomson, 1982). Besides, it also minimizes the windowed periodogram, when $k = 1$, $\lambda = 1$ and $w_1(t) = 1/\sqrt{n}$. It also leads to the Welch and the Bartlett method, if accurate choice is selected for W_j and $\lambda(j)$, $j = 1,....,k$. If frequency warping matrix M is chosen, the multitaper estimator will change to ordinary non-warped cepstrum and MFCC.

SPEECH RECOGNITION MODEL

LDA Models

Let us consider the i^{th} word of the speech signal of the training data as $X_i \in R^{m \times n_i}$ and X_i column represents the m dimensional speech (word) of i class with n_i training words, such that, $i = 1,2,...c$, where c indicates the total number of words. Let the probe word be y and it is given as

$$y = X_i \alpha_i, \; i = 1,2,...,c \tag{24}$$

where $\alpha_i \in R^{n_i \times 1}$ refer to the regression parameter and it can also be evaluated with the least square method using the formula

$$\hat{\alpha}_i = \left(X_i^T X_i\right)^{-1} X_i^T y, \; i = 1,2,...,c \tag{25}$$

The reconstruction of probe word with each class can be estimated using

$$\hat{y}_i = X_i \hat{\alpha}_i X_i \left(X_i^T X_i\right)^{-1} X_i^T y = H_i y, \; i = 1,2,...,c \tag{26}$$

where H_i refers to the hat matrix, which maps y to $\hat{y}_i$ and the error in reconstruction can be obtained using

$$e_i = \|y - \hat{y}_i\|_2^2,\ i = 1,2,\dots,c \tag{27}$$

The LRC substitutes the value of y to the class with less error during reconstruction. Let $X = [x_1,\dots.x_i,\dots,x_n] \in R^{m\times n}$ be the training matrix generated from the training database, where m and n denote the dimensionality of the training word and number of training words, in order. The x_i class label is represented as $l(x_i) \in \{1,2,\dots c\}$ and the subspace projection matrix is given as $U \in R^{m\times d}$. The speech is intruded in the subspace as

$$y_i = U^T x_i \tag{28}$$

where $d < m$ and $y_i \in R^{d\times 1}$. Since the label of x_i is equal to y_i, $l(y_i) = l(x_i)$.

$$BCRE = \frac{1}{n(c-1)} \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq 1(x_i)}}^{c} \left\|y_i - \hat{y}_{ij}^{\text{inter}}\right\|_2^2 \tag{29}$$

and

$$WCRE = \frac{1}{n} \sum_{i=1}^{n} \left\|y_i - \hat{y}_i^{\text{intra}}\right\|_2^2 \tag{30}$$

where $\hat{y}_i^{\text{intra}}$ and $\hat{y}_{ij}^{\text{inter}}$ indicate the reconstruction of y_i by $l(y_i)$ class and jth class, respectively. Increasing the Between-Class Reconstruction Error (BCRE) and reducing Within-Class Reconstruction Error (WCRE) will give the subspace projection matrix.

LCDR

The error problem in reconstruction of class can be reduced using the proposed method. Let the training speech vector be $X = [X_1,.X_2,\dots,X_c] \in R^{m\times n}$ and $X_i = [X_{i1}, X_{i2},\dots,X_{in_{ui}}] \in R^{m\times n_i}$ where $n = \sum_{i=1}^{c} n_i$ and n_i, m represent the number of training words from ith class and dimension of the training word. Let $d < m$ and $U \in R^{m\times d}$ using $y_{ij} = U^T x_{ij}$, where $1 \leq j \leq n_i$, the x_{ij} can be mapped to the learned subspace. So, the complete training speech words can be mapped as $Y = U^T X \in R^{d\times n}$ and considering each class, it can be written as $Y_i = U^T X_i \in R^{d\times n_i}$.

$$CBCRE = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left\|y_{ij} - \hat{y}_{ij}^{\text{inter}}\right\|_2^2 \tag{31}$$

$$WCRE = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left\|y_{ij} - \hat{y}_{ij}^{\text{intra}}\right\|_2^2 \tag{32}$$

Here, $\hat{y}_{ij}^{\text{intra}} = Y_{ij}^{\text{intra}} \alpha_{ij}^{\text{intra}}$ and $\hat{y}_{ij}^{\text{inter}} = Y_{ij}^{\text{inter}} \alpha_{ij}^{\text{inter}}$ where Y_{ij}^{intra} and Y_{ij}^{inter} represent the Y_i with y_{ij} eliminated and Y with Y_i eliminated. The value of $\alpha_{ij}^{\text{intra}}$ and $\alpha_{ij}^{\text{inter}}$ is taken from eq. (25). $\hat{\alpha}$ is estimated in the original space and it is used in the learned subspace in the place of α.

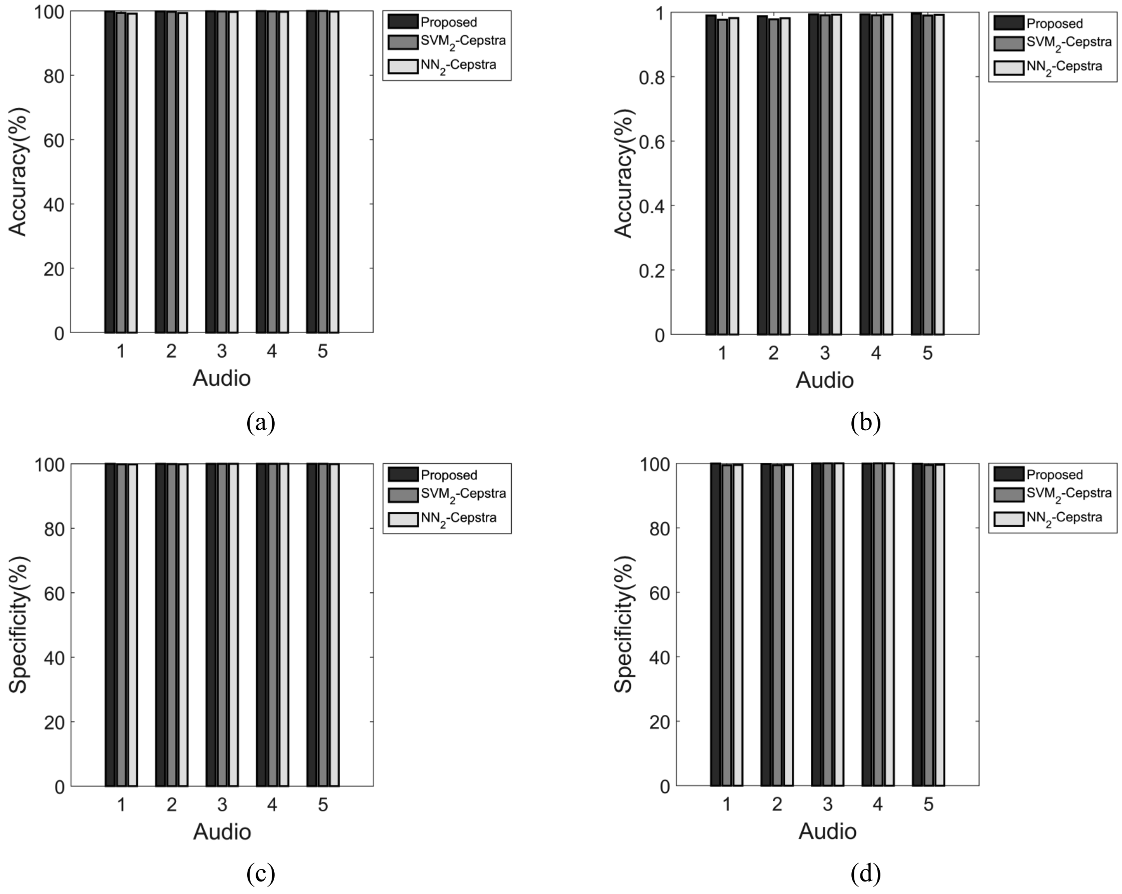


Fig. 1. Accuracy and Specificity of test cases A [(a) (c)] and B [(b) (d)].

It is also noted from eq. (32) that class-specific representation is used by BCRE and cross-class collaborative representation is used by Collaborative Between-Class Reconstruction Error (CBCRE). So, the WCRE and CBCRE can be also expressed as

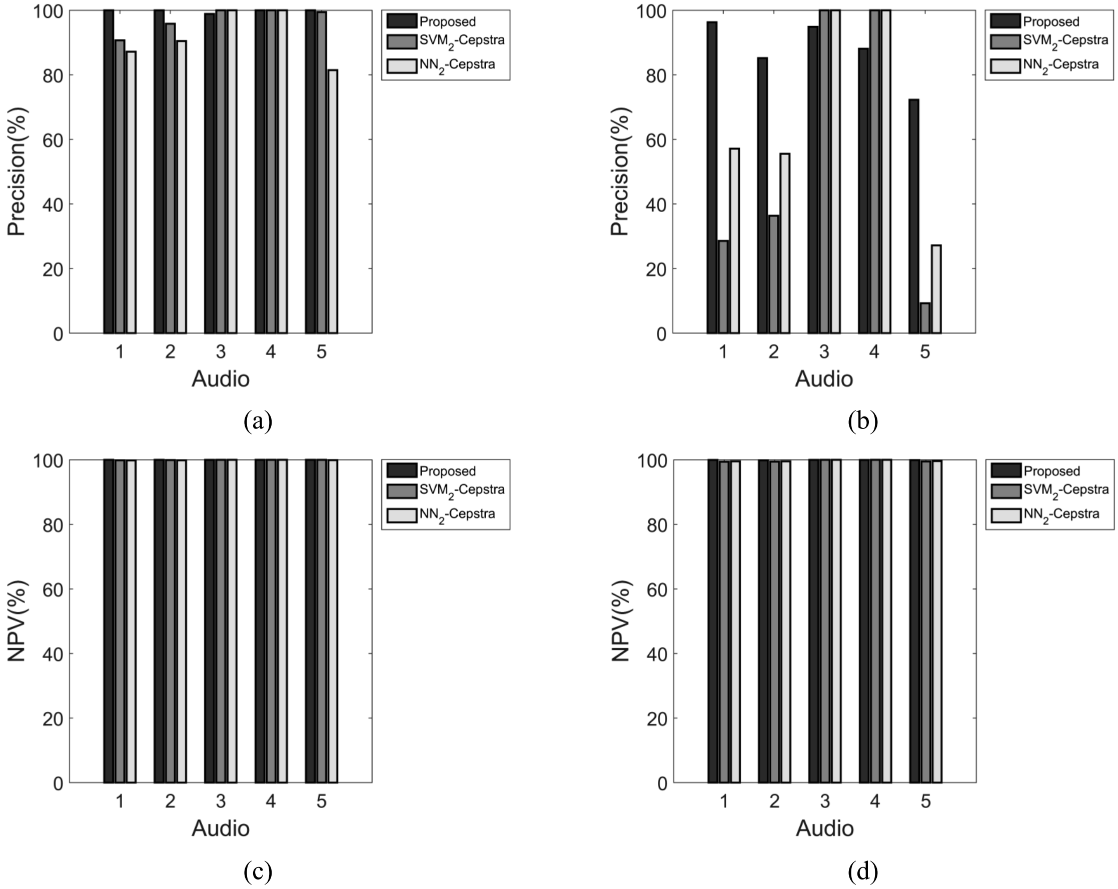
$$CBCRE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left\| U^T x_{ij} - U^T x_{ij}^{\text{int}\,er} \alpha_{ij}^{\text{int}\,er} \right\|_2^2 \tag{33}$$

$$WCRE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left\| U^T x_{ij} - U^T x_{ij}^{\text{int}\,ra} \alpha_{ij}^{\text{int}\,ra} \right\|_2^2 \tag{34}$$

The above two equations can also be written as

$$CBCRE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left(x_{ij} - X_{ij}^{\text{int}\,er} \alpha_{ij}^{\text{int}\,er} \right)^T UU^T \left(x_{ij} - X_{ij}^{\text{int}\,er} \alpha_{ij}^{\text{int}\,er} \right) \tag{35}$$

$$WCRE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left(x_{ij} - X_{ij}^{\text{int}\,ra} \alpha_{ij}^{\text{int}\,ra} \right)^T UU^T \left(x_{ij} - X_{ij}^{\text{int}\,ra} \alpha_{ij}^{\text{int}\,ra} \right) \tag{36}$$



(a) (b)

(c) (d)

Fig. 2. PPV and NPV of test cases A [(a) (c)] and B [(b) (d)].

The $\frac{1}{n}$ factor is eliminated from WCRE and CBCRE without changing the ratios between them. So,

$$CBCRE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} tr\left(U^T \left(x_{ij} - X_{ij}^{\text{inter}} \alpha_{ij}^{\text{inter}}\right)\left(x_{ij} - X_{ij}^{\text{inter}} \alpha_{ij}^{\text{inter}}\right)^T U\right) \quad (37)$$

$$WCRE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} tr\left(U^T \left(x_{ij} - X_{ij}^{\text{intra}} \alpha_{ij}^{\text{intra}}\right)\left(x_{ij} - X_{ij}^{\text{intra}} \alpha_{ij}^{\text{intra}}\right)^T U\right) \quad (38)$$

where $tr(\cdot)$ indicates the trace operator.

$$E_b = \frac{i}{n} \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left(x_{ij} - X_{ij}^{\text{inter}} \alpha_{ij}^{\text{inter}}\right)\left(x_{ij} - X_{ij}^{\text{inter}} \alpha_{ij}^{\text{inter}}\right)^T \quad (39)$$

$$E_w = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left(x_{ij} - X_{ij}^{\text{intra}} \alpha_{ij}^{\text{intra}}\right)\left(x_{ij} - X_{ij}^{\text{intra}} \alpha_{ij}^{\text{intra}}\right)^T \quad (40)$$

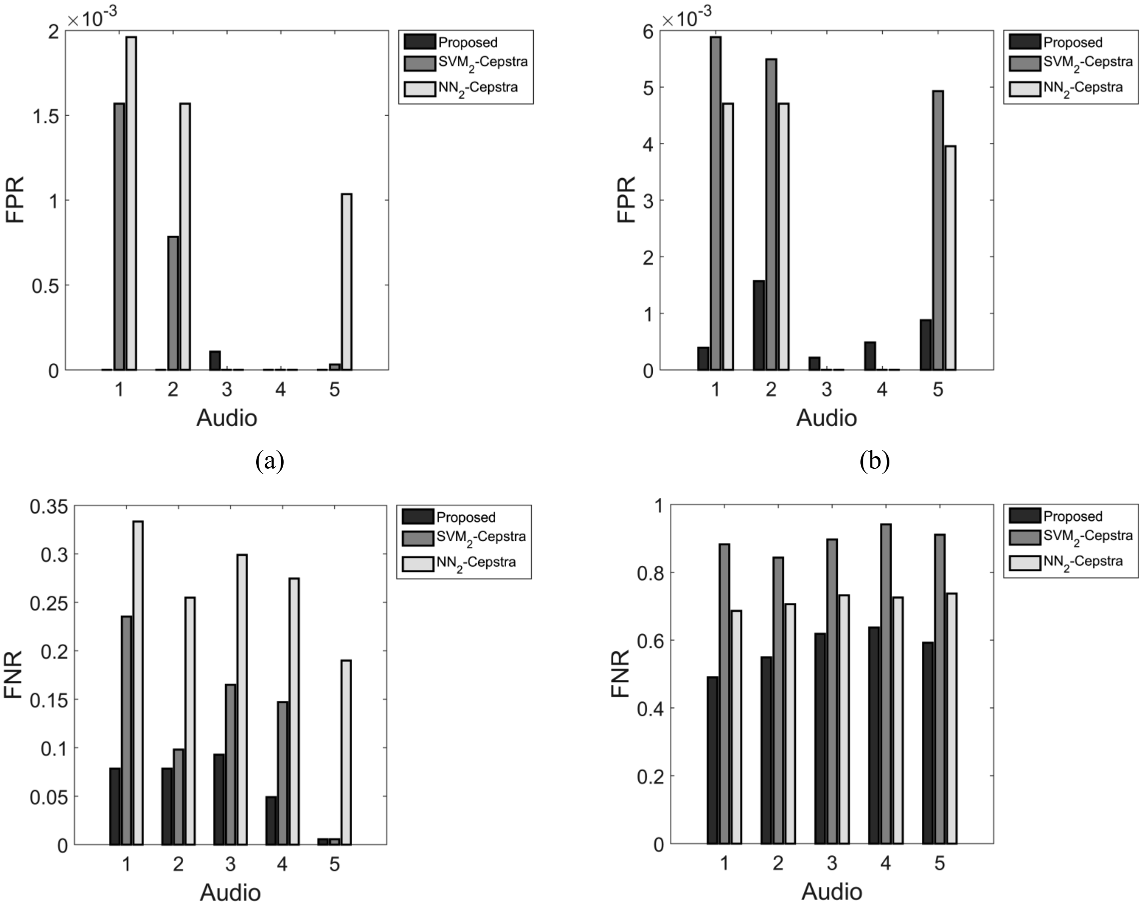
So, $CBCRE = tr\left(U^T E_b U\right)$ and $WCRE = tr\left(U^T E_w U\right)$

The maximum range criterion (MMC) (Li *et al.,* 2006) can be used to increase the CBCRE and WCRE. Hence, $\max_U J(U) = \max_U (CBCRE - WCRE) = \max_U \left(tr\left(U^T (E_b - E_w) U\right)\right)$

The above equation can be solved by identifying the biggest d Eigenvalue and the Eigenvectors are given as

$$(E_b - E_w)u_k = \lambda_k u_k, k = 1,2,...,d, \tag{41}$$

where $U = [u_1,...,u_k,...,u_d]$ and $\lambda_1 \geq ... \geq ...\lambda_k ... \geq \lambda_d$.



(a) (b)

(c) (d)

Fig. 3. FPR and FNR of test cases A [(a) (c)] and B [(b) (d)].

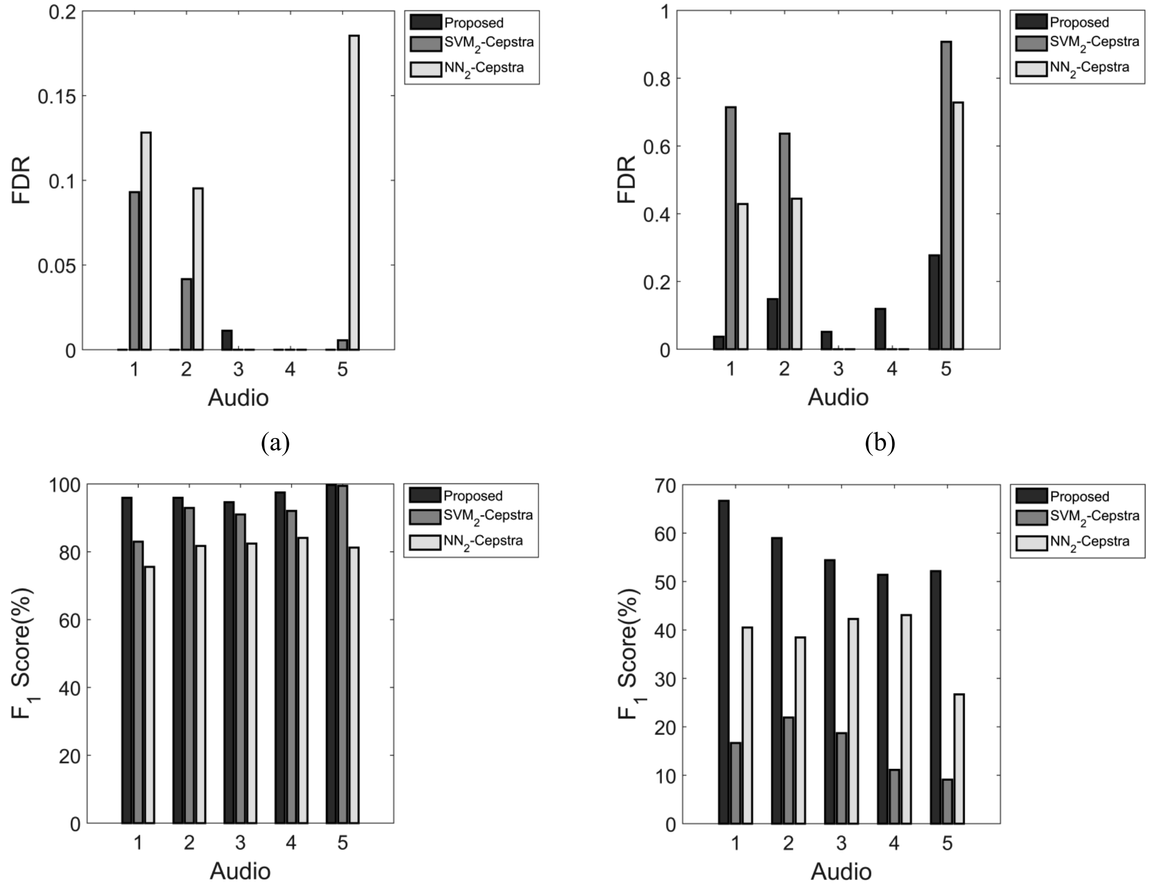
The small sample size problem (SSSP) with speech signal greater than the training speech number can be solved with MMC. The LCDRC algorithm is briefly described as follows:

1. To obtain the unit l_2-norm, the test speech signal and the training speech signal must be normalized.
2. The projection matrix U is estimated for all training speeches X and this X is substituted in the discriminant subspace and so, $Y = U^T X$ is generated.
3. For each class $i = 1,2,...,c$, the hat matrix H_i is estimated.
4. x is transformed to learned subspace for the test speech x using $y = U^T x$. The i^{th} class is used to reconstruct $\hat{y}_i = H_i y, i = 1,2,...,c$.
5. The error in reconstruction is calculated using $e_i = \|y - \hat{y}_i\|_2, i = 1,2,...,c.$

RESULTS AND DISCUSSION

Experimental Setup

Simulation experiments have been carried out in MATLAB with five standard audio sequences in .wav format that are downloaded from http://tdil-dc.in. The number of words in signal 1, 2, 3, 4, and 5 is about 51, 51, 97, 102, and 179, respectively. The performance testing is done by considering two-mode test case A and test case B. In test case A, the speech from speaker A is exploited for training and the testing speech is extracted from speaker B. The database has 12 audio signals of continuous speech of Hindi language from which seven audio signals were considered for training and the rest of the five signals were used for testing the system. The audio signals were segregated for training and testing in such a way that each training signal has used 1165 words, whereas the testing signal has 480 words. The speech signals were acquired from male Hindi speakers. In order to evaluate the insensitivity of the algorithm against the gender, female version of speech was synthesized for modifying the pitch of the test signal and so the experimentation was widened. The frequency spectrum of the complete database falls within the range of 16 KHz. The Hindi words are repetitive, but randomly throughout the database. The silence based preprocessing stage has been implemented to perform word based speech segmentation. Moreover, silence based preprocessing stage has been implemented to perform word based speech segmentation.



(a)

(b)

(c)

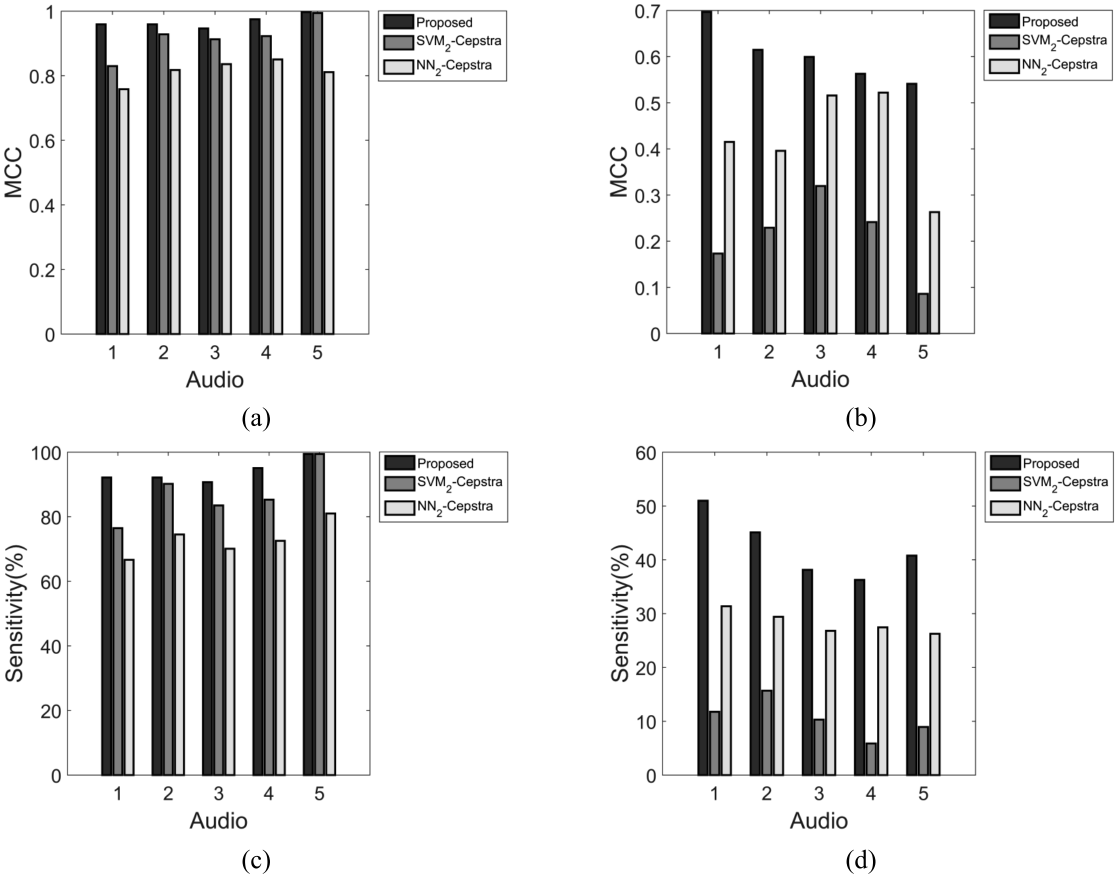
(d)

Fig. 4. FDR and F$_1$ Score of test cases A [(a) (c)] and B [(b) (d)].

Table 2. Statistical analysis on speech recognition performance for test case A.

	Audio signal 1			Audio signal 2			Audio signal 3			Audio signal 4			Audio signal 5		
	LCDR	SVM	NN	LCDR	SVM	NN	LCDR	SVM	NN	LCDR	SVM	NN	LCDR	SVM	NN
Best	99.49	98.18	98.67	99.26	98.3	98.64	99.82	99.55	99.74	99.82	99.57	99.77	100	99.5	99.64
Worst	98.53	97.20	97.69	98.27	97.31	97.65	98.84	98.59	98.75	98.84	98.59	98.79	99.09	98.5	98.69
Mean	99.06	97.69	98.16	98.77	97.78	98.11	99.32	99.08	99.25	99.33	99.06	99.25	99.59	98.97	99.16
Median	99.1	97.67	98.15	98.8	97.76	98.08	99.3	99.04	99.24	99.33	99.07	99.25	99.59	98.98	99.14
SD	0.27	0.29	0.28	0.29	0.27	0.28	0.29	0.30	0.28	0.27	0.29	0.29	0.28	0.29	0.28

In test case B, the test speech signals are synthesized to by changing the tempo of the training speech signals so that speaker independent environment is simulated. For the purpose of comparing the proposed recognition model's performance, renowned classifier models such as a neural network (NN) and support vector machine (SVM) are used. Henceforth, the NN, SVM, and proposed classifier models are referred to as NN_2-cepstra, SVM_2-cepstra, and LCDR, and they are analysed for both cases using the performance metrics such as accuracy, specificity (measures the proportion of negatives), sensitivity (measures the proportion of positives), PPV (Positive Predictive Value), NPV (Negative Predictive Value), FDR (False Discovery Rate), FPR (False Positive Rate), MCC (Matthews Correlation Coefficient), F_1 score, which is the harmonic mean of precision and sensitivity, and FNR (False Negative Rate). The performance of the three methods is statistically studied by determining the best, worst, mean, median, and standard deviation (SD) and the best method is sorted. Here, the precision percentage represents the ratio between the true positive and sum of true positive and false positive, where the true positive is the number of words that are recognized correctly, while the false positive is the number of words that are misclassified stating as positive words.



(a)

(b)

(c)

(d)

Fig. 5. MCC and Sensitivity of test cases A [(a) (c)] and B [(b) (d)].

Table 3. Statistical Analysis on Speech Recognition Performance for Test Case B.

	Audio signal 1			**Audio signal 2**			**Audio signal 3**			**Audio signal 4**			**Audio signal 5**		
	LCDR	**SVM**	**NN**	**LCDR**	**SVM**	**NN**	**LCDR**	**SVM**	**NN**	**LCDR**	**SVM**	**NN**	**LCDR**	**SVM**	**NN**
Best	99.04	97.74	98.24	98.81	97.85	98.2	99.39	99.12	99.29	99.37	99.12	99.33	99.63	99.05	99.24
Worst	98.95	97.64	98.14	98.72	97.76	98.1	99.29	99.02	99.19	99.28	99.02	99.24	99.53	98.95	99.14
Mean	98.99	97.69	98.19	98.76	97.81	98.15	99.34	99.07	99.25	99.32	99.07	99.28	99.58	99	99.19
Median	98.99	97.69	98.19	98.76	97.81	98.14	99.33	99.08	99.25	99.32	99.08	99.28	99.58	99	99.19
SD	0.029	0.03	0.027	0.028	0.027	0.026	0.028	0.03	0.028	0.031	0.029	0.029	0.029	0.03	0.026

Fig. 1, 2, 3, 4, and 5 represent the performance of NN_2-cepstra, SVM_2-cepstra, and LCDR methods for test cases A and B with considered parameters, accuracy and specificity, PPV and NPV, FPR and FNR, FDR and F_1 scores, MCC, and sensitivity, respectively. The statistical prediction of the performance of the audio signals for test cases A and B is tabulated in Tables 2 and 3, respectively. Tables 2 and 3 represent the accuracy as the recognition performance measure. The variation in recognition performance is observed because of varying pitch in the unknown audio signals.

Test Case A

Among various metrics considered, the proposed method shows 100% accuracy for signal 1 and 2, 100% F_1 score for signal 5, 100% PPV for signals 1, 2, 4, and 5, and 100% sensitivity for signal 5. For all signals except the fifth signal, the sensitivity of the proposed method is about 9% and 25% more than the SVM_2-cepstra and NN_2-cepstra methods, respectively.

In the case of FDR, the proposed method has performed well for signals 1, 2, 4, and 5, though just a performance lagging with the signal 3. For all signals with respect to FNR parameter, the proposed method shows nearly 0.22%, 0.02%, 0.07%, and 0.09% difference from SVM_2-cepstra and 0.26%, 0.18%, 0.21%, 0.23%, and 0.16 % difference from NN_2-cepstra method. Since the FNR needs to be minimized, the proposed method ensures its superior performance over the other two methods. In other words, the proposed method has lesser false negative outcomes, i.e., misclassification of desired words as undesired words, than the other two methods. The similar kind of better performance can be seen from FPR outcomes. For the signals 1, 2, 3, 4, and 5, the MCC and F_1 score are found to be higher for the proposed method than the SVM_2-cepstra and NN_2-cepstra method. Overall, the performance of the proposed method is higher with increased accuracy, F_1 score, PPV (except in signal 3), MCC, sensitivity and decreased FDR, FNR, FPR than the NN and SVM_2-cepstra methods. From Table II, it is noted that, for all audio signals, the LCDR method shows better mean, median, worst, best, and standard deviation. Among the audio signals, the best, worst, mean, median, and SD are better for audio signal 5 and they are 100%, 99.09%, 99.59%, 99.59%, and 0.28%, respectively.

Test Case B

In test case B, the NPV and specificity of all the three methods are found to be 100% for all signals. The accuracy of the proposed is found to be 99%, 98%, 99%, 99%, and 100%, which are higher than those of the other two methods. Similar to accuracy, the sensitivity is highly increased for the LCDR method, which is noted as 42%, 30%, 30%, 35%, and 32% more than SVM_2-cepstra and 22%, 23%, 12%, 8%, and 14% more than the NN_2-cepstra method with respect to signals 1, 2, 3, 4, and 5. Referring the PPV, the proposed method shows good performance for the signals 1, 2, and 5, but for signals 3 and 4, the SVM_2-cepstra and NN_2-cepstra possess 100% PPV and this is a very significant variation when compared to the proposed method. For all the signals except 3 and 4, the FDR of the proposed method is lesser than SVM_2-cepstra and NN_2-cepstra method. This in turn depicts that the proposed method has the characteristics of reduced misclassification. In the case of MCC, the proposed method shows better performance for signals 1, 2, 3, and 4 with 54%, 4%, 29%, and 33% more than SVM_2-cepstra and 28%, 39%, 5%, and 52% than NN_2-cepstra method. The signals in FNR and FPR graph (except 3 and 4) are found to be lesser than the SVM_2-cepstra and NN_2-cepstra method. From the results, the performance of the proposed method is higher than the NN_2-cepstra and SVM_2-cepstra methods. Table III interprets that the LCDR method performs better by considering the statistically predicting measures. The measures best, worst, mean, median, and SD are better for LCDR in audio signal 5 and they are noted as 99.63%, 99.53%, 99.58%, 99.58%, and 0.029%, respectively. Although the number of subjects used in the database is less, the recognition system attempted to recognize nearly 5000 words of Hindi. The number of records of the database, i.e., number of subjects, can be increased further since it is word based recognition system.

CONCLUSION

The present paper discussed the problems related to the speech recognition system for Hindi language. The challenges have been overcome in this paper by proposing a LCDR based recognition model with cepstra features. The performance of the LCDR method has been analysed by considering various parameters such as specificity, accuracy, PPV, NPV, FPR, FDR, F_1 score, sensitivity, MCC, and FNR, and the comparison has been done with SVM_2-cepstra and NN_2-cepstra. The mode of experiment has been done under two cases, A and B. The maximum possible performance has been accomplished by the proposed method for most of the audio signals in terms of accuracy, NPV, sensitivity, PPV, F_1 score, and specificity for test case A. In test case B, the maximum possible performance has been accomplished in terms of specificity and NPV. Since the adopted models are stochastic by nature, a statistical analysis has been carried out in which the best, worst, mean, median, and SD metrics are utilized for investigating the performance exhibited by the proposed model. The analysis has resulted in a finer performance of the proposed recognition model over the conventional models and hence ensured the consistency of the recognition accuracy. In the presence of noise, the robustness of the proposed scheme will be evaluated in future work.

REFERENCES

Akita, Y. & Kawahara, T. 2010. Statistical Transformation of Language and Pronunciation Models for Spontaneous Speech Recognition, *IEEE Transactions on Audio, Speech, and Language Processing,* **18**(6): 1539-1549.

Arisoy, E., Chen, S.F., Ramabhadran & B., Sethy, A. 2014. Converting Neural Network Language Models into Back-off Language Models for Efficient Decoding in Automatic Speech Recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **22**(1): 184-192.

Baker, J.M. 2009. Updated MINDS report on speech recognition and understanding, Part 2, *IEEE Signal Processing Mag.,* **26**(4): 78-85.

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.H., Morgan, N & Shaughnessy, D.O. 2009a. Research Developments and directions in speech recognition and understanding, Part 1, *IEEE Signal Processing Mag.,* **26**(3): 75-80.

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.H., Morgan, N. & Shaughnessy, D.O. 2009b. Updated MINDS report on speech recognition and understanding, Part 2 [DSP Education], *IEEE Signal Processing Magazine,* **26**(4): 78-85.

Bengio, Y., Ducharme, R., Vincent, P & Jauvin, C. 2003. A neural probabilistic language model, *J. Mach. Learn. Res.,* **3**: 1137-1155.

Davis, S.B & Mermelstein, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustic, Speech and Signal Processing,* **28**(4): 357-366.

Donough, J.M & Waibel, A. 2003. Maximum mutual information speaker adapted training with semi-tied covariance matrices, *in Proc.* ICASSP.

E. Arisoy, E., Chen, S.F., Ramabhadran, B & Sethy, A. 2013. Converting neural network language models into back-off language models for efficient decoding in automatic speech recognition, *in Proc. ICASSP,* 8242-8246.

Evermann, G., Chan, H.Y., Gales, M.J.F., Hain, T., Liu, X., Mrva, D., Wang, L & Woodland, P.C. 2004. Development of the 2003 CU-HTK conversational telephone speech transcription system, *in Proc.* ICASSP.

Fan, X & Hansen, J.H.L. 2011. Speaker identification within whispered speech audio streams, *IEEE Trans. Speech Audio Process.,* **19**(5): 1408-1421.

Goel, V., Axelrod, S., Gopinath, R., Olsen, P & Visweswariah, K. 2003. Discriminative estimation of Subspace Precision and Mean (SPAM) models, *in EUROSPEECH,* 2003.

Hasan, T., Hansen, J.H.L. (2013). Acoustic Factor Analysis for Robust Speaker Verification, *IEEE Transactions on Audio, Speech, and Language Processing,* **21**(4): 842-853.

Hasan, T. & Hansen, J.H.L. 2014. Maximum Likelihood Acoustic Factor Analysis Models for Robust Speaker Verification in Noise, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**(2): 381-391.

Hermansky, H. 1990. Perceptual Linear Predictive (PLP) analysis of speech, *Journal of the Acoustic Society of America,* **87**(4): 1738-1752.

Ichikawa, O., Fukuda, T. & Nishimura, M. 2010. Dynamic Features in the Linear-Logarithmic Hybrid Domain for Automatic Speech Recognition in a Reverberant Environment, *IEEE Journal of Selected Topics in Signal Processing,* **4**(5): 816-823.

Jeong, Y. 2012. Adaptation of Hidden Markov Models Using Model-as-Matrix Representation, *IEEE Transactions on Audio, Speech, and Language Processing,* **20**(8): 2352-2364.

Kalinli, O., Seltzer, M.L., Droppo, J. & Acero, A. 2010. Noise Adaptive Training for Robust Automatic Speech Recognition, *IEEE Transactions on Audio, Speech, and Language Processing,* **18**(8): 1889-1901.

Kim, D.Y., Evermann, G., Hain, T., Mrva, D., Tranter, S.E., Wang, L & Woodland, P.C. 2003. Recent advances in broadcast news transcription, *in Proc. ASRU.* 105-110.

Kolossa, D., Zeiler, S., Saeidi, R. & Astudillo, F. 2013. Noise-Adaptive LDA: A New Approach for Speech Recognition Under Observation Uncertainty, *IEEE Signal Processing Letters, Volume* **20**(11): 1018-1021.

Kumar, M., Rajput, N. & Verma, A. 2004. A large-vocabulary continuous speech recognition system for Hindi, *IBM Journal of Research and Development,* **48**, (5.6): 703-715.

Le, H.S., Oparin, I., Allauzen, A., Gauvain, J. & Yvon, F. 2013. Structured Output Layer Neural Network Language Models for Speech Recognition, *IEEE Transactions on Audio, Speech, and Language Processing,* **21**(1): 197-206.

Li, X., Jiang, T. & Zhang, K. 2006. Efficient and robust feature extraction by maximum margin criterion, *IEEE Trans. Neural Netw.* **17**: 157-165.

Liu, S. & Sim, K.C. 2014. Temporally Varying Weight Regression: A Semi-Parametric Trajectory Model for Automatic Speech Recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **22**(1): 151-160.

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J & Khudanpur, S. 2010. Recurrent neural network based language model, *in Proc. Inter-speech,* 1045-1048.

Mikolov, T., Kombrink, S., Burget, L., Cernocky, J & Khudanpur, S. 2011. Extensions of recurrent neural network language model, *in Proc. ICASSP.* 5528-5531.

Paithane, A.N., Bormane, D.S & Dinde. S. 2014. Human Emotion Recognition using Electrocardiogram Signals. *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC),* **2**: (2).

Paithane, A.N. & Bormane, D.S. 2014. Analysis of nonlinear and non-stationary signal to extract the features using Hilbert Huang transform, *in Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on,* **1-4**: 18-20.

Paithane, A.N. & Bormane, D.S. 2015. Electrocardiogram signal analysis using empirical mode decomposition and Hilbert spectrum, *in Pervasive Computing (ICPC), 2015 International Conference on,* **1-4**: 8-10.

Percival, D.B & Walden, A.T. 1993. Spectral Analysis for Physical Applications. Cambridge, U.K.: Cambridge Univ. Press.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE,* **77**(2): 257-286.

Rainville, E.D. 1960. Special Functions. New York: Chelsea.

Rennie, S.J., Hershey, J.R. & Olsen, P.A. 2010. Single-Channel Multitalker Speech Recognition, *IEEE Signal Processing Magazine,* **27**(6): 66-80.

Rose, R., Hofstetter, E & Reynolds, D. 1994. Integrated models of signal and background with application to speaker identification in noise, *IEEE Trans. Speech Audio Process.,* **2**(2): 245-257.

Schwenk, H & Gauvain, J.L. 2005. Training neural network language models on very large corpora, *in Proc.* HLT-EMNLP, 201-208.

Schwenk, H. 2007. Continuous space language models, *Comput. Speech Lang.,* **21**(3): 492-518.

Simand, K.C. & Gales, M.J. 2006. Minimum phone error training of precision matrix models, *IEEE Transactions on Audio, Speech, and Language Processing,* **14**(3): pp. 882-889.

Tanja, S & Alex, W. 2001. Language-independent and language-adaptive acoustic modeling for speech recognition, *Speech Communication,* **35**(1-2): 31-51.

Thomson, D.J. 1982. Spectrum estimation and harmonic analysis, *Proc. of the IEEE,* **70**(9): 1055-1096.

Zhang, C & Hansen, J.H.L. 2011. Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing, *IEEE Trans. Speech Audio Process*., **19**(4): 883-894.

Zhang, W. & Fung, P. 2013. Sparse Inverse Covariance Matrices for Low Resource Speech Recognition, *IEEE Transactions on Audio, Speech, and Language Processing,* **21**(3): 659-668.

Zhang, W. & Fung, P. 2014a. Discriminatively Trained Sparse Inverse Covariance Matrices for Speech Recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **22**(5): 873-882.

Zhang, W. & Fung, P. 2014b. Efficient Sparse Banded Acoustic Models for Speech Recognition, *IEEE Signal Processing Letters,* **21**(3): 280-283.

Submitted: 04/10/2016
Revised: 03/01/2017
Accepted: 01/02/2017

الانحدار التمييزي التعاوني الخطي وميزات Cepstra للتعرف على الكلام باللغة الهندية

*يو جي باتيل، **س. دي. شيرباهادوركار و*آي. إن بايثان
*كليةJSPM's Rajarshi Shahu للهندسة، جامعة SPPU، تاثاوايد، بون، ماهارشترا، الهند
**كلية D. Y. Patil للهندسة، جامعة SPPU، بيمبري-شينشواد، ماهارشترا، الهند

الخلاصة

نظام التعرف على الكلام هو أحد الأنظمة الهامة ولكن الصعبة في التفاعل بين الإنسان والحاسوب. يجد التعرف على اللغات الهندية العديد من الصعوبات العملية بسبب خصائصها النحوية والصوتية الأكثر توسعاً من اللغة الإنجليزية. يركز هذا البحث على نظام التعرف على الكلام باللغة الهندية والذي يقترح ميزات Cepstra ونموذج الانحدار التمييزي التعاوني الخطي (LCDR) لتحليل الكلام والتعرف عليه. بالنسبة لإشارات صوتية محددة، تم تجميع نموذجين من إشارات كلامية موضوع الاختبار وتم فحصها تجريبياً. ثم تم تحليل أداء طريقة LCDR باستخدام دوال الخطأ من النوع الأول والثاني ومقارنتها بالطرق الحالية مثل NN2-cepstra و SVM2-cepstra. علاوة على ذلك، تم استخدام الوسط الحسابي والوسيط والانحراف القياسي الأفضل والأسوأ للتنبؤ الإحصائي، وتم إثبات طريقة LCDR المُقترحة كأفضل طريقة للتعرف على الكلام باللغة الهندية.