Machine Learning Paradigms for Speech Recognition of an Indian Dialect

N. D. Londhe, *IEEE Senior Member*, M. K. Ahirwal, P. Lodha

***Abstract*—Present era is full of speech recognition based services and products. The machine learning paradigms is at the centre stage of speech recognition methodology. Automatic speech recognition (ASR) technology has vastly evolved in recent years including emerging applications in mobile computing, natural user interface, and man-machine assistive technology. In this paper, it's the first time we are presenting ASR designs based on two important machine learning paradigms Artificial Neural Network (ANN) and Support Vector Machine (SVM) for an rare and geographically important Indian dialect *'Chhattisgarhi'*. The conventional feed-forward ANN and SVM have been applied on the dataset of maximum 50 isolated words of 15 speakers. The performance of these machine learning paradigms is compared with state of art Hidden Markov Model (HMM). The tendency of ASR to be speaker dependent and independent has been extensively investigated with speaker variation experiments. Furthermore the reliability and stability of ASR has been confirmed with numerical validation. The exhaustive review of ASR techniques from the literature along with the ASR systems designed on Indian languages is presented.**

***Index Terms*—Machine learning paradigm, MFCC, HMM, ANN, SVM.**

I. Introduction

AUTOMATIC speech recognition (ASR) is a feature based data driven technology, requires a relatively large amount of labelled data to train acoustic models with several machine learning paradigms. Basically feature extraction in front end and pattern recognition at back ends are used in ASR [1]. Researchers have implemented several feature extraction techniques like linear predictive coefficients (LPC), linear predictive cepstral coefficients (LPCC), Mel frequency cepstral coefficients (MFCC), perceptual linear predictive coefficients (PLP) and human factor cepstral coefficient (HFCC), Gammatone filter cepstral coefficient (GFCC) feature to improve system performance [2].Selection of these techniques depends on their capabilities to reduce redundancy towards irrelevant information of speech data [3]. Mel frequency cepstral coefficient (MFCC) is most suitable feature extraction technique for speech processing as it mimics the functioning and working of human auditory system very efficiently [4], Hidden Markov Model (HMM) [5], Artificial Neural Networks (ANN) [6],SVM [7] and Dynamic Bayesian Networks [8] are used in back end as pattern classifiers.

On the basis of data collection and data processing ASR may be implemented with any of isolated, connected or continuous mode of operation. In isolated speech recognition, information falling between two silence zones of a word is used to extract speech features. Connected word systems allow separate utterances of a word to be 'run-together' with a minimal pause between them. In Continuous speech recognition system it is difficult to find several word boundaries present in a sentence. Different utterances of same word may change in different sentences due to co-articulation effects and position of word relative to other words present in a sentence. In given work we have implemented three machine learning paradigms to recognize isolated Chhattisgarhi words under speaker dependent (SD) and speaker independent (SI) speech recognition environments. It is implemented on self recorded speech data in various geographical locations of Chhattisgarh state of India consisting of 7500 utterances of 50 common but acoustically challenging words for 15 speakers.

India has about 1650 dialects /native languages [9]. Presently Chhattisgarhi is considered to be one of the several dialects of Hindi like Braj, Bundeli, Kanauji, Bangru, Awadhi, Bhojpuri, Marwari, Bagheli etc. [10]. It is not only spoken widely in Chhattisgarh, but also in some parts of other states including M.P., Orissa etc. Recently Chhattisgarhi is declared as second official language of Chhattisgarh state [11] and spoken by approximately seventeen million people. Chhattisgarhi is most closely related to Bagheli and Awadhi and these languages are classified in the east central zone of the Indo-Aryan languages [12].

It is required a huge data set to implement ASR for any language. Some of researchers used broadcast news, public corpora and other speech related resources; some used their proprietary speech and text corpora to implement ASR with a wide variety of topics and speaking styles. Insufficiency of such resources is the major hurdle in research for Chhattisgarhi or any other Indian languages. Unfortunately any speech corpus or standard database in Chhattisgarhi

N. D. Londhe is with National Institute of Technology, Raipur, India. He is Asstt. Prof. in Department of Electrical Engineering.

M. K. Ahirwal is with the National Institute of Technology, Raipur, India , CG-492101. He is Asstt. Prof. in Department of Computer Applications (e-mail: ahirwalmitul@gmail.com).

P.Lodha is M.Tech. Student in Department of Electrical Engineering, National Institute of Technology, Raipur, India.

978-1-5090-0396-9/16/$31.00 ©2016 IEEE

language is not available for public use. As per the authors' knowledge and observed from literature review, no ASR exists for Chhattisgarhi language. Therefore an attempt has been made to build an ASR specifically for Chhattisgarhi language. The organization of this paper is as follows. Section II explains the methodology. Section III includes brief description of related work on different Indian languages followed by Sections (IV-IV) showing results, discussion and conclusion.

II. Materials and Methods

ASR has been implemented in various application projects in different languages successfully. It follows a standard methodology consists data acquisition, speech segmentation, feature extraction and model formation stages to prepare efficient system. These models are used as benchmarks in testing phase. In recognition phase features extracted from testing data with some decision making algorithms are used to recognize uttered words. We have used MFCC as feature extraction technique. For pattern classification HMM, ANN and SVM are used here as machine learning paradigms. The overall methodology is discussed in more detail in following sub sections.

A) Data Acquisition

We used a basic noise cancelling handy recorder 'Zoom H4n' for data acquisition. We recorded speech samples at 16-bit/44.1 kHz in stereo mode. For recording purpose we selected native Chhattisgarhi speakers and asked them to utter each word ten times. We have used institute's lab and open environment (with moderate ambient noise) for recording. The words are selected from 'English to Chhattisgarhi dictionary' [13], Chhattisgarhi news paper articles and various Chhattisgarhi literatures. We recorded 50 words with multiple utterances per word from 15 individuals for training. To check the speaker dependent and speaker independent recognition rate speech signals are arranged in 3 groups and then trained with machine learning paradigms.

- Group A contains speech samples of speaker 1 to speaker 5.
- Group B contains speech samples of speaker 6 to speaker 10.
- Group C contains speech samples of speaker 11 to speaker 15.

In this way we have 150 utterances of each word and total 7500 speech samples.

B) Speech Segmentation

The separation of uttered speech samples of an individual speaker is perfumed using the speech segmentation. These speech samples required for machine learning are separated using bottom-up blind speech segmentation as it is language independent and doesn't require any prior information of speech samples [14].We can segment words of different lengths using this technique. With speech segmentation we removed silences present near utterances by using dynamic threshold based algorithms.The variations in amplitude of speech signal are tracked using short time energy by windowing the signal [15], spectral position or spectral centroid [16], spectral flux [17] and boundaries at the locations where amplitude or spectral changes exceed a minimum threshold level are calculated.

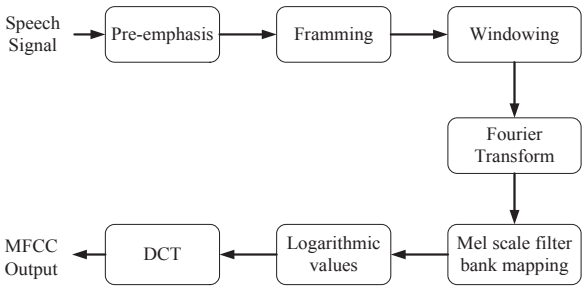

Fig. 1. Block diagram representation of feature extraction technique.

C) Feature Extraction: In feature extraction, we extract set of discriminative features from input speech samples by using MFCC technique. In MFCC, like human auditory system we arrange frequency bands on Mel scale [18]. The conversion from normal scale to Mel scale is done using equation 1. Feature extraction using MFCC technique is shown in Fig.1.

$$M(f) = 1125 * ln\left(\frac{1+f}{700}\right) \quad (1)$$

where f is normal frequency and $M(f)$ is the Mel scale frequency related to f.

Pre-emphasis is used to boost high frequency content of speech samples by increasing their energies. We used pre-emphasis coefficient alpha as 0.97. Speech signals are random in nature and length of each utterance depends on number of letters present in a word. For analysis of variable length utterances we use framing. Signal is converted in set of small frames having a length of 25 milliseconds with frames shifting of 10 milliseconds. Windowing technique has been applied before computing Fourier transform of the signals to reduce spectral leakages. We used Hamming windows as they are easy to implement and follow linear phase properties [19]. After windowing signals are passed to a filter bank of 20 filters and filter bank energies are computed. Discrete cosine transform is used to convert log compressed filter bank energies in time domain. These time domain coefficients are called Mel frequency cepstral coefficients.

D) Machine learning paradigms: Efficient and satisfactory performance of ASR depends on successful training of system by using any one of machine learning paradigm. Using extracted features models and trained, networks are prepared which are able to classify testing data in correct class. 3 machine learning paradigms are applied in given work on features, extracted from isolated Chhattisgarhi words. Working of these machine learning paradigms are explained in following sub sections.

(i) *Hidden Markov Model:* HMM is a powerful statistical method, considered as a finite state machine. It characterizes observed data samples in discrete or continuously distributed time series. It utilizes a nondeterministic process to produce output observation symbols for given state [20]. HMM is a stochastic process generated by two interrelated procedures,

Markov chain and a set of random functions [21]. Number of states in Markov chain is fixed and the transitions between these states depend on calculation of transition probabilities. Second process is state output observations. First process is related to temporal structure of model and second process is related to spectral variability of speech. Usually the model outputs are linked with states rather than with the transition of the states [22]. According to Bakis topology reported in [23], three types of transitions are possible in HMM model; we implemented modified Bakis topology suggested in [24], only 2 types of transitions is allowed; the transition from a state to itself (loop) and the transition from a state to its immediate successor state (forward). Fig. 2 shows four states left right Markov model.

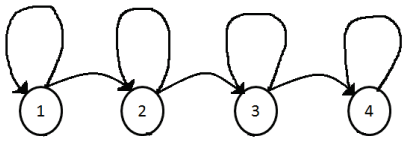

Fig. 2. A Markov model with 4 states.

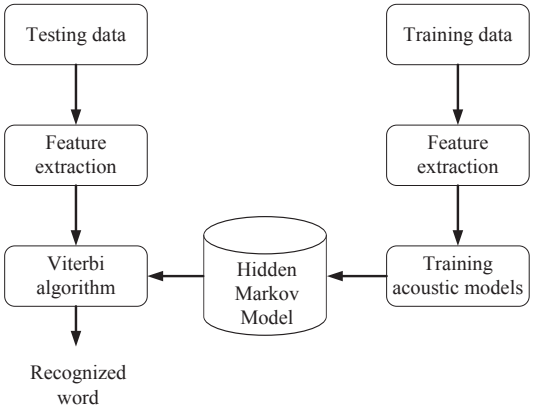

Fig. 3. Block diagram of ASR considering HMM as classifier.

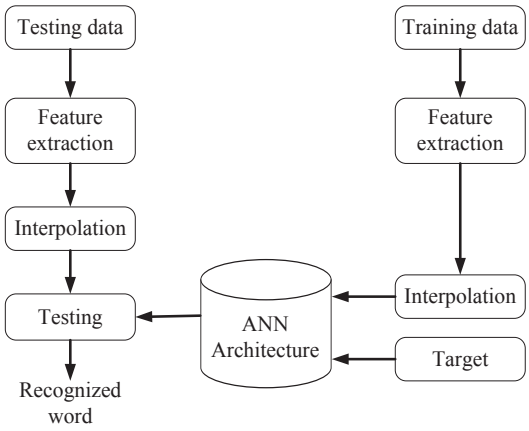

Fig. 4. Block diagram of ASR considering ANN as classifier.

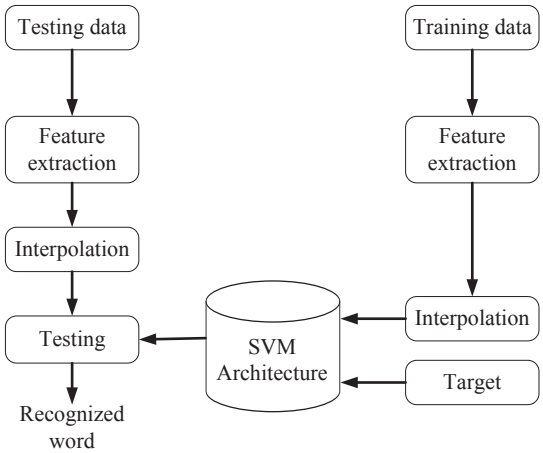

Fig. 5. Block diagram of ASR considering SVM as classifier

In training phase of HMM 3 parameters (π, A, B) are estimated using Baum-Welch [25] and Forward-Backward algorithms [22]. Where π represents initial state distribution, A is the state transition probability distribution and B is the output probability distribution. HMM model is prepared for each word and can be represented as:

$$\lambda = (A, B, \pi) \quad (2)$$

During recognition feature vectors from the speech signals are obtained using MFCC followed by measurement of Observation sequence $O = \{O_1, O_2, O_3, \ldots\ldots O_T\}$; then model likelihoods are calculated by implementing Viterbi algorithm [25] for all the models associated to each word present in vocabulary. Equation 3 is used to find out the word whose model likelihood is highest. This word is said to be recognized. The block diagram of system is shown in the Fig. 3.

$$v^* = [P(O|\lambda^v)],\ 1 \le v \le V \ \ (V \text{ is word length}) \quad (3)$$

(ii) *Artificial Neural Network*: ANNs are able to learn from complex input-output relationships and acts as pattern classifiers. Multilayer feed-forward neural networks are generally used [26], same has been implemented in this paper for isolated Chhattisgarhi words recognition. For training of neural network, it requires fixed number of input nodes in the network. As different utterances are varying in lengths, calculated MFCC features are also varying in lengths, Bicubic interpolation has been applied to fix the variable length MFCC features. MFCC output of a single utterance is arranged in a column matrix of size 585*1. To make system speaker independent concatenated utterances from different speakers were used to train it. Back propagation algorithm with 3 hidden layers, were used to train neural network. Number of neurons in each layer depends on word length, size of input and target matrices. Tangent sigmoid (tansig) layer activation function and scaled conjugate gradient (trainscg) network training function were used in training procedure.

As we have used 3 network layers in our neural network, the output equation can be written as:

$$a3 = f3(W3f2(W2f1(W1P + b1) + b2) + b3) \quad (4)$$

where $a3$ represents the output of neural network, P is input vector, $W1$, $W2$, $W3$ are weights, $b1$, $b2$, $b3$ are bias values, $f1$, $f2$, $f3$ are layer activation functions for layer 1, 2 and 3 respectively. Weights are updated with following relation:

$$W_{k+1} = W_k + \alpha_k P_k \quad (5)$$

where α_k is learning rate, P_k is search direction and these values are calculated according to principle of scale conjugate gradient algorithm [27].Fig. 4 shows block diagram representation of ASR system considering ANN as classifier.

(iii) *Support Vector Machine:* Extracted features from MFCC are processed same as in ANN to train SVM

classifier. SVMs are supervised learning based classifiers which generates a set of hyper-planes in a multidimensional space, which are used for classification [28]. Block diagram of SVM is shown in Fig. 5. Multiclass SVM with linear kernel has been utilized in this paper, A basic SVM classifies input vector $X \in Rn$, can be expressed as:

$$g(x) = w.\varphi(x) + b \tag{6}$$

Where w represents a normal vector of a separating hyper-plane in a feature space where the input vector X is mapped with a mapping function $\varphi(x)$ and b is a bias. SVM works on structural risk minimization principal, finds a compromise between the empirical error and a confidence measure corresponding to generalization. SVM learning as a constrained optimization problem [29] is defined as:

$$\min_{w,b,\xi} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \tag{7}$$

$$subjected\ to \quad y_i(w.\varphi(x_i) + b) \geq 1 - \xi_i \tag{8}$$

$$\xi_i \geq 0, for\ i = 1, \ldots, N$$

Where $x_i \in Rn$, $i=1,2,\ldots,N$ is a set of training vectors with corresponding targets (class labels) 21±0, $y_i \in \{-1, +1\}$, and the parameter C controls the trade-off between the two measures. Solving the optimization problem yields

TABLE I
DIFFERENT ASR SYSTEMS IMPLEMENTED IN INDIAN LANGUAGES.

S. No.	Authors	Language	Feature extraction	Classifier	Data set	SD/SI	Accuracy %
1	Ahad et. al. 2002 [45]	Urdu	MFCC	ANN	Isolated digits	SD	94.00
2	Hasnat et. al. 2007 [46]	Bengali	MFCC	HMM	Isolated Words	SD	90.00
						SI	70.00
					Continuous	SD	80.00
						SI	60.00
3	Krishnan et. al. 2009 [41]	Malyalam	Wavelet packet decomposition	SVM	Isolated Words (20 words)	SI	61.00
			Discrete wavelet Transform			SI	89.00
4	Pour et. al. 2009 [47]	Persian	MFCC and Discrete wavelet Transform	ANN	Isolated digits	SI	98.00
5	Venkateswarlu et. al. 2011 [48]	Indian	LPCC	Modified Self organizing map	Isolated words	SI	88.05
			MFCC			SI	89.27
6	Vimala.C et. al. 2011 [38]	Tamil	MFCC	HMM	Isolated words (50 words)	SI	88.00
7	Deka et. al 2011 [44]	BODO	LPC	ANN	Isolated vovels	SI	51.50
8	Patel et. al. 2011 [43]	Gujrati	Phonem Based feature extraction	English speech recognition Engine	Isolated words (30 words)	SD	88.71
9	Kumar et. al. 2011 [32]	Hindi	MFCC	HMM	Isolated words (30 words)	SI	94.63
10	Gawali et. al. 2011 [34]	Marathi	MFCC	Statistical method	Isolated words		94.65
			DTW				73.25
11	Dua et. al. 2012 [49]	Panjabi	MFCC	HMM	Isolated words (115 word for training and 30-35 words for testing)	SD	97.78 In room environment
						SD	95.49 In open environment
						SI	93.49 In room environment
						SI	92.68 In open environment
12 *****	Hegde et. al. 2012 [40]	Kannad		SVM	Isolated words		79.00

$$w = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i) \quad (9)$$

$$b = \sum_j \alpha_i \, y_i \varphi(x_i)^T . \varphi(x_j) + y_i \quad , \forall i \quad (10)$$

Where α_i, i=1,2,…,N are the Lagrange multipliers corresponding to each training vector x_i. Only the training vectors with non-zero coefficients α_i can contribute to determine the weight vector w and called as support vectors [30]. With kernel function defined as $K(x_i, x_j) = \varphi(x_i)^T . \varphi(x_j)$, equation 6 can be written as:

$$g(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x_j) + b \quad (11)$$

and Linear kernel function used in given work is defined as

$$K(x_i, x_j) = x_i^T x_j \quad (12)$$

III. Related Work

Automatic speech recognition is very first time implemented for Chhattisgarhi language. ASR has been implemented for Hindi [31-33], Marathi [34, 35], Bengali [36, 37], Tamil [38], Telgu [39], Kannad [40], Malayalam [41, 42], Gujrati [43] and Bodo [44] languages. Many researchers have worked in this area and used one of the machine learning paradigms to prepare ASR system. In TABLE I detailed analysis of ASR systems implemented for Indian languages is presented.

IV. Results and Discussion

Training and testing is done for 3 groups of data we have formed, with 3 classifiers HMM, ANN and SVM for speaker dependent and speaker independent speech recognition. We have also analyzed speaker variability with these classifiers to study inter speaker variations. We have done training and testing for bunch of 10 words, 20 words and 50 words with groups A, B and C. Following section shows results achieved from different experimental work.

Speaker Dependent Speech Recognition: In this category we trained our system with first 5 utterances of each word from all speakers present in a group and testing is done with last 5 utterances of each word for individual speaker present in same group. TABLE II shows percentage accuracies obtained with HMM, ANN and SVM classifiers.

Speaker independent speech recognition: In this category we trained our system with first 5 utterances of each word

TABLE II
ACCURACY OF SPEAKER DEPENDENT SPEECH RECOGNITION

Speaker Dependent	10 words			20 words			50 words		
	HMM	ANN	SVM	HMM	ANN	SVM	HMM	ANN	SVM
Group A	99.20	100.00	100.00	98.60	98.80	98.32	98.64	94.88	98.32
Group B	98.40	96.80	97.20	98.80	97.40	97.60	98.32	93.36	97.20
Group C	98.80	97.60	97.20	98.60	98.00	97.80	97.68	94.48	97.68
Overall Accuracy	98.80	98.13	98.13	98.67	98.07	97.91	98.21	94.24	97.73

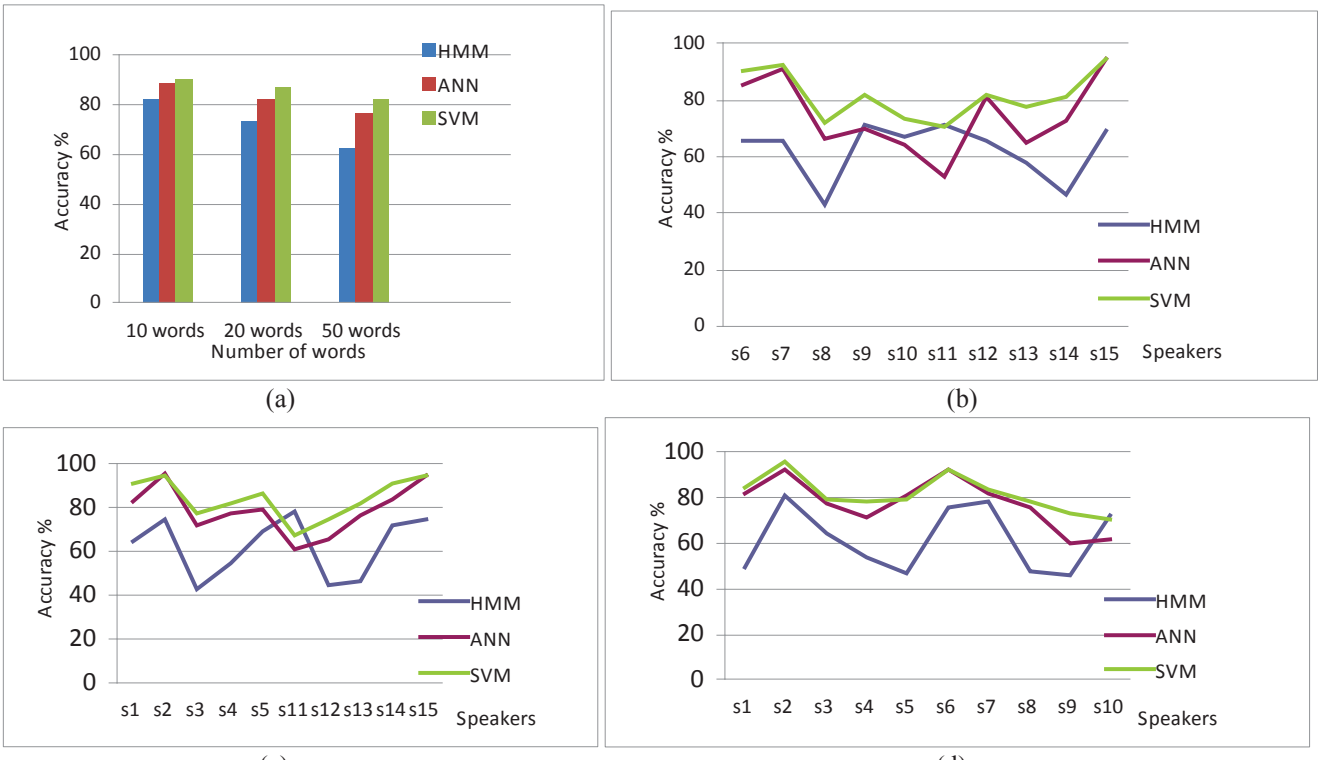

Fig. 6. (a) Comparison of different machine learning paradigm for speaker independent speech recognition, (b) Inter speaker variations for experiment 1, (c) Inter speaker variations for experiment 2 and (d) Inter speaker variations for experiment 3.

from all speakers present in a group and testing is done with speakers present in other 2 groups. 3 experiments have been performed to study speaker independent speech recognition. Data set used for training and testing in different experiments is shown in TABLE A1 (Provided as supplementary document, TABLE A1-A5) and percentage accuracies for 3 classifiers are shown in TABLE A2.

Speaker variability analysis: Inter speaker variations are important to study the performance of speaker independent speech recognition system. The speaker variability can be analyzed in two categories intra-speaker variability and inter-speaker variability. A speaker normally changes the quality of his voice, articulation pattern, speaking rate according to his physical and emotional situations, background noise and other environmental factors. Due to these reasons, there are a lot of variations in the speech signals even if a unique speaker produces the same linguistic message many times. These changes are readily observed in digital representations of speech signals. Physiological differences are an important source of variation between speakers. Each person has a different vocal tract, controlled by a unique brain. In speaker dependent cases, speech variations are typically less vast. When we generalize the ASR task to be speaker independent, as in most services for the general public, we face much larger range of variability that arises from different people, with their varied vocal tracts and diverse styles of speaking. It is challenging to handle all these variability for ASR. Intra-speaker variability is usually handled reasonably well via statistical models. Inter-speaker variability seems to be a greater problem. System can be made speaker independent by mixing training templates from a wide variety of speakers. Individual speaker's word recognition efficiency is calculated here with models prepared for all groups in 3 experimental categories. TABLES A3, A4 and A5 shows inter speaker variations for experiment 1, 2 and 3 respectively.

V. Discussion

With different training and testing experiments we come to know that our system fails to recognize words spoken by unknown speaker when trained with data set of single speaker. System can be made speaker independent if trained with mix utterances of different speakers. Efficiency of system increases as number of utterances increases to train the system. For automatic speech recognition of Chhattisgarhi we implemented 3 machine learning paradigms, achieved satisfactory results with them and observed that overall SVM performs well for speaker independent speech recognition as shown in Fig. 6(a).

Inter speaker variations for speaker independent speech recognition with 15 speakers have studied with 3 classifiers. A single speaker can provide variations in efficiency for different models generated by same machine learning tool. Inter speaker variations with 50 words for experiments 1, 2 and 3 are shown in Fig. 6(b) – 6(d) respectively. Form these Fig.we can conclude that svm provide consistent results as compared to HMM and ANN.

VI. Conclusion

Machine learning paradigm with traditional feed forward neural networks and SVM for maximum word length of 50 is studied and results are compared with state of art. For speaker dependent speech recognition all classifier perform well. For speaker independent speech recognition ANN gives better performance than HMM and SVM gives the best performance among the three classifiers. It gives accuracy of 90.60, 87.27 and 82.49 (in percent) for word length of 10, 20 and 50 respectively. Efficiency of the system reduces on increasing the dictionary size. With 3 experimental work detailed analysis of inter speaker variability for speaker independent speech recognition is analysed here. This shows that recognition accuracy varies on the basis of selection of speakers for training the system. If system learns with speakers having differences in there vocal tract characteristics, it performs well as in experiment 2. Recognition of larger dictionary size and continuous Chhattisgarhi speech is left as future work.

Acknowledgment

The work reported in this paper is funded by Chhattisgarh Council of Science and Technology (CCOST) Raipur, Chhattisgarh.

References

[1] D. Shaughnessy, "Interacting with computer by voice automatic speech recognition and synthesis*", In proc. Of IEEE,* Vol. 9, No. 91, 2003, pp.1272-1305.

[2] C. Vimala and V. Radha, "A Review on Speech Recognition Challenges and Approaches", World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, No. 1, 2012, pp. 1-7.

[3] M. Danubianu, P. Valentin and T. Iolanda, "Unsupervised Information-Based Feature Selection for Speech Therapy Optimization by Data Mining Techniques", In proc. Of 7th International Multi-Conference on Computing in the Global Information Technology, 2012, pp.206-211.

[4] Z. András, K. Daniil, S. Ralf and N. Hermann, "Using Multiple Acoustic Feature Sets for Speech Recognition", Speech Communication, Vol. 49, No. 6, 2007, pp.514-525.

[5] F. Jelinek, "Statistical methods for speech recognition", MIT Press, Cambridge, 2011.

[6] R. P. Lippmann, "Review of neural networks for speech recognition", Neural Computation, Vol. 1, No. 1, 1989, pp.1–38.

[7] A. Ganapathiraju, H. Jonathan and P. Joseph Picone, "Support vector machines for speech recognition", *In Proc. of IEEE Conf. on Spoken Language Processing,* 1988, pp. 2923–2926.

[8] J.A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition", *IEEE Signal Processing Magazine,* Vol. 22, No. 5, 2005, pp.89–100, 2005.

[9] S. Hussain, N. Durrani and S.Gul, "Survey of Language Computing in Asia," Center for Research in Urdu Language Processing, NUCES, 2005.

[10] P. M. Colin, "The Indo-Aryan Languages",CAMBRIDGE UNIVERSITY PRESS, Melbourne, 1991,

[11] R. Pathak and S. Dewangan, "Natural Language Chhattisgarhi: A Literature Survey", International Journal of Engineering Trends and Technology (IJETT), Vol. 12, No. 2, 2014, pp. 113-117.

[12] http://archive.ethnologue.com/16/show_language.asp?code=hne

[13] English to Chhattisgarhi, Administrative Dictionary published by Chhattisgarhi National Language Committee, Raipur.

[14] Y. P. Estevan, V. Wan and O. Scharenborg, "Finding Maximum Margin Segments in Speech*", In Proc. Of IEEE Conf. on Acoustics, Speech, and Signal Processing (ICASSP '07),* Honolulu, Hawaii, USA, 2007, pp. 937-940.

[15] T. Zhang and J. C. CKuo, "Hierarchical classification of audio data for acrhiving and retrieving", *In Proc. Of IEEE Conf. on Acoustics,Speech and Signal Processing,* 1999, pp. 3001–3004
[16] T Giannakopoulos, "Study and application of acoustic information for the detection of harmful content and fusion with visual information" Ph.D. dissertation, Dept. of Informatics and Telecommunications, University of Athens, Greece, 2009.
[17] R. K. Aggarwal and M. Dave, "Acoustic modeling problem for automatic speech recognition system: conventional methods (Part I)", International Journal of Speech Technology, Vol. 14, No. 4, 2011, pp.297–308.
[18] D. O'Shaugnesy, "Speech Comunication: Human and Machine, second edition", University press, India 2004.
[19] J.G. Proakis and D.G. Manolakis, "Digital Signal Processing: Principles, Algorithms and Applications", 3rd edition, Prentice Hall, New York, 2000.
[20] H. Xuedong, A. Alex and W. Hsiao, "Spoken Language Processing A Guide to Theory", Algorithm and System Development, Prentice-hall Inc. 2001.
[21] R.J. Elliott, L. Aggoun and J.B. Moore, "Hidden Markov Models: Estimation and Control", Springer Verlag, 1995.
[22] R.R. Lawrence, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *In proc. of IEEE*, Vol. 77, No. 2, 1989, pp. 257-286.
[23] R. Bakis, "Continuous speech word recognition via centisecond acoustic states", In Proc. Of ASA meeting, Washington, DC, USA 1976.
[24] I. Bhardwaj and N. D. Londhe, "Hidden Markov Model Based Isolated Hindi Word Recognition", *In proc. Of IEEE conf. on Power, Control and Embedded Systems,* 2012, pp.1-6.
[25] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Trans. on Information Theory*, Vol. 13, No. 2, pp. 260-269, 1967.
[26] A. K. Jain, "Artificial Neural Network: A tutorial", Computers , Vol. 3, 1996, pp.31-33.
[27] M.F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning", Neural networks, Vol. 6, No. 4, 1993, pp.525-533.
[28] V.N. Vapnik and V. Vlamimir, "Statistical learning theory", New York: Wiley, 1998.
[29] C.W. Hsu, C. Chih-Chung and L. Chih-Jen, "A practical guide to support vector classification", Department of Computer Science, National Tiawan University, 2003.
[30] H. Shimodaira, N. Ken-ichi, N. Mitsuru and S. Shigeki, "Support vector machine with dynamic time-alignment kernel for speech recognition", Interspeech, 2001, pp. 1841-1844.
[31] N. Chalapathy, N. Rajput and A. Verma, "A large-vocabulary continuous speech recognition system for Hindi", IBM Journal for Research and Development, 2002, pp. 1-5.
[32] K. Kumar and R.K. Aggarwal, "Hindi Speech Recognition System Using HTK", International Journal of Computing and Business Research, Vol. 2, No. 2, 2011, pp. 2229-6166.
[33] R.K. Aggarwal and M. Dave, "Using Gaussian Mixtures for Hindi Speech Recognition System", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 4, No. 4, 2011, pp. 157-170.
[34] B.W. Gawali, S. Gaikwad, P. Yannawar and S. C. Mehrotra, " Marathi Isolated Word Recognition System using MFCC and DTW Features", ACEEE International Journal on Information Technology, Vol. 01, No. 01, 2011, pp. 21-24.
[35] B.W. Gawali, S. Gaikwad, P. Yannawar, S.C. Mehrotra, "Marathi Isolated Word Recognition System using MFCC and DTW Features", In proc. Of International Conf. on Advances in Computer Science, 2010, pp. 21-24.
[36] F. Hassan, M. R. A. Kotwal, M. S. A. Khan and M. N. Huda, "Gender independent Bangla automatic speech recognition", In proc. Of on Informatics, Electronics & Vision (ICIEV), 2012, pp. 144-148.
[37] G. Muhammad, A. Yousef Alotaibi and M. N. Huda, "Automatic speech recognition for Bangla digits", *In proc. Of 12th IEEE Conf. on Computers and Information Technology, ICCIT*'09, 2009, pp. 379-383.
[38] C. Vimala and V. Radha, "Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM", Procedia Engineering, Vol. 30, 2012, pp.1097-1102.
[39] K.V.N. Sunitha and N. Kalyani, "Isolated Word Recognition using Morph Knowledge for Telugu Language", International Journal of Computer Applications, Vol. 38, No. 12, 2011, pp. 47-54.
[40] H. Sarika, K.K. Achary and S. Shetty, "Isolated word recognition for Kannada language using support vector machine", Wireless Networks and Computational Intelligence, Springer Berlin Heidelberg, 2012, pp. 262-269.
[41] K. Vimal, V.R., Babu and P. Anto, "Features of wavelet packet decomposition and discrete wavelet transform for Malayalam speech recognition", Recent Trends Engineering,Vol.1,No.2,2009, pp. 93–96.
[42] K. M. Smrithyn, "Shreshta Bhasha - Malayalam Speech Recognition using HTK", International Journal of Advanced Computing and Communication Systems (IJACCS), Vol. 1, No. 1, 2014, pp. 1-5.
[43] H. N. Patel and P.V. Virparia, "A Small Vocabulary Speech Recognition for Gujarati", International Journal of advanced Research in Computer Science, Vol. 2, No. 1, 2011, pp. 1-5.
[44] M.K. Deka , C.K. Nath , S.K. Sarma and P.H. Talukdar, "An Approach to Noise Robust Speech Recognition using LPC-Cepstral Coefficient and MLP based Artificial Neural Network with respect to Assamese and Bodo Language", International Symposium on Devices MEMS, Intelligent Systems & Communication (ISDMISC), 2011, pp. 23-26.
[45] A. Ahad, A. Fayyaz and T. Mehmood, "Speech recognition using multilayer perceptron", *In proc. Of IEEE Students Conf*., 2002, pp. 103-109.
[46] M. A. Hasnat, J. Mowla and M. Khan, "Isolated and continuous bangla speech recognition: implementation, performance and application perspective", In proc. of International Symposium on Natural Language Processing (SNLP), Hanoi, Vietnam, 2007, pp. 1-6.
[47] M. M. pour and F. Farokhi, "A new approach for Persian speech Recognition", *In proc. Of IEEE conf. on Advance Computing Conference (IACC*2009), 2009, pp.153-158.
[48] R.L.K. Venkateswarlu and R. VasanthaKumari, "Novel approach for speech recognition by using Self-Organised Maps", In proc. Of Conf. on Emerging Trends in Networks and Computer Communications (ETNCC*)*, 2011, pp. 215–222.
[49] M. Dua, R.K. Aggarwal, V. Kadyan and S. Dua, "Punjabi Automatic Speech Recognition Using HTK", International Journal of Computer Science Issues, Vol. 9, No. 4, 2012, pp.359-364.