Towards Improved Continuous Hindi Speech Recognition: Challenges and Advancement

Ankit Kumar[a] and Tripti Chaudhary[b]
[a]*Department of Computer Science & Engineering*
I.M.S. Engineering College,
Adhyatmik Nagar, Ghaziabad, U.P., India.
[b]*Department of Electronics & Communication Engineering*
Visveshwarya Group of Institution,
Greater Noida,U.P., India.
[*]anketvit@gmail.com

Abstract

State-of-the-art Automatic Speech Recognition (ASR) system follow a well-established statistical paradigm, that of parameterization of speech (feature extraction) at front-end and likelihood evaluation of feature vectors at back-end. This paper presents the major advancement in Automatic Speech Recognition (ASR) made in last one decade and various challenges in Hindi ASR with their most probable solutions to provide the highly efficient Hindi ASR system. ASR is the challenging task due to many sources of variability associated with the speech signal such as contextual variability, inter-speaker and intra-speaker variability and environment variability. The objective of this review paper is to summarize the technological development in ASR which may be help to identify the research topic. After year of research and development of Hindi ASR, the author identifies the major hurdles in existing ASR and described the various techniques to overcome those problems.

Keywords-Automatic Hindi speech recognition, Automatic speech recognition, Pattern recognition, Continuous Speech Recognition

Introduction

Speech is one of the most important tool for communication between man and his environment. Therefore manufacturing of high accuracy ASR system is the need of human being all the time. Many times keyboard acts as a barrier between computer and the user. Despite of significant technological development in the last one decade, ASR could not able to achieve the human perfection in adverse condition. In this paper various difficulties are described which are arise during the development of Hindi ASR system and their possible solutions are also described. There are many technological developments in ASR which is yet to be implemented in Hindi ASR.

Refinement and advancement in language modeling and acoustic modeling is the hot research area in ASR. At back-end of ASR, the statistical framework of Hidden Markov Model (HMM) has been widely used for classification since 1980's. Continuous- Density Hidden Markov Model (CD-HMM) is used in most existing ASR system in which parameters are estimated using Maximum- Likelihood (ML) estimation. Assumption of independence in which speech frames are assumed to be independent is not

Copyright 2016 IMSEC

completely true, poor discrimination due to Maximum- Likelihood (ML) estimation instead of Posteriori- estimation (MAP), and weak duration modeling are the limitations of conventional HMM model. Many researchers have proposed many refinement and advancement for HMM based acoustic model to overcome its limitation (Aggarwal, 2014).

In ASR, n-gram language model is the most popular choice from the last four decade. There has been a long struggle to move beyond n-gram language model by various statistical models: (i) Random forest language model (Xu, 2004), (ii) Non- parametric language model (The, 2006), (iii)Discriminative language model (Roark, 2006), and (iv) Class- based language model (Chen, 2009). Recently, weighted finite state transducer (WFST) is used for language model and pronunciation model, in which all the ASR component are integrated in to a joint graph with additional optimization (Mohri, 2002). WFST provides an elegant interface for spoken language understanding (Deoras *et al.,* 2013). Now- a- days, recurrent neural network (RNNLM) is introduced (Mikolov *et al.,* 2013).

Technological Advancement In ASR:

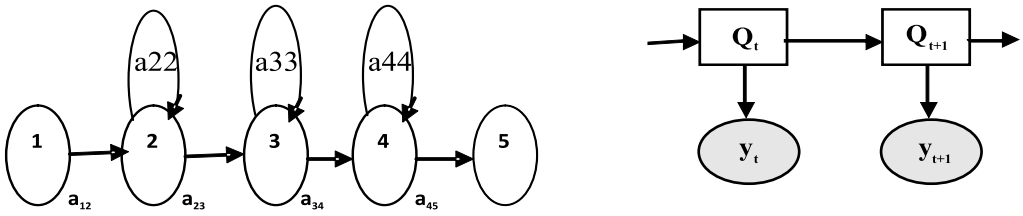
Automatic speech recognition is a challenging task from the last three decades due to various sources of variability's such as environment variations, speaker variability and contextual variability. Despite the diligent efforts of researchers to fill the gap between man and machine performance for speech recognition, the performance of state of art ASR system are not so perfect for real time applications. Major progress is being recorded regularly on technological advancement of ASR systems. Current research in ASR dealing various deficiencies presents to implement the ASR system. Researches have proposed some advanced techniques to handle the deficiencies are as follows:

Refinement of HMM

Hidden Markov model (HMM's) provide an effective framework for acoustic modeling of speech recognition. Today, almost every speech recognition systems are based on HMM. It's weak duration modeling, simplifying assumptions and restrictions on feature extraction imposed by frame based observation would result of system poor accuracy and unacceptable adaptation to changes in environment (Aggarwal *et al.,* 2011). Thus, current ASR system requires considerable refinement to handle these deficiencies.

1. Dynamic Bayesian Network

Dynamic Bayesian network is an alternative, complementary and graphical representation of basic HMM model which is useful for describing the extension to HMM model and emphasis the conditional dependencies (Gales andYoung, 2007).



[HMM] [DBN]

Figure 1: HMM network (up) and DBN (down) network

Dynamic Bayesian network represents a set of variables (discrete/ continuous etc.) on a directed acyclic graph in the form of nodes and maps the conditional independence between these variables. In DBN, squares denotes discrete variables, circles represents the continuous variables, shading in the circle/ square denotes observed variable and no shading represent an unobserved variables. Each DBN is the combination of variables which represents the observations and hidden states of that process. Generally, DBN composed of a sequence of T hidden state variable (Q_t , Q_{t+1}) and sequence of T observable variable (y_t, y_{t+1}) where T shows the time limit of that particular process.

The advantage of DBN over HMM is that in DBN, we can easily show the extension of the model in term of modifying the conditional independence assumptions (Gale and Young, 2007). By adding additional unobserved variables and adding additional dependency arcs between variables, we can extend the basic HMM structure. We can calculate the conditional probability P ($y_{1:T}$) of occurrence of the sequence of observation $y_{1:T}$ by:

$$P(y_{1:T}) = \Sigma [{}_{t=1}\Pi^{T-1} P(Q_{t+1}/ Q_t) P(y_{t+1}/ Q_t) P(Q_{T+1}/ Q_T)] \quad \ldots(1)$$

2. *HMM Duration Modeling*

HMM duration modeling plays an important role in the comprehension of speech. The duration modeling in speech recognition is not noticeable due to the HMM's because of its poor phone duration modeling. In classical HMM's, state duration modeling in which probability for duration 'd' in state 'i' is represented as:

$$P_i(d) = P(Q/\Theta, q_l = S_i)$$
$$= (a_{ij})^{d-1}(1-a_{ii}) \quad \ldots(2)$$

where , d= 1, 2…..with Q= $\{\underbrace{S_i, S_i \ldots..S_i, S_i}_{\text{'d' times}}\}$ $j \neq i$

Here, a_{ii} is the self-transition probability of state S_i. Thus, in standard HMM models, the probability of state duration decreases exponentially with time and clearly this is a poor model of duration because it is based on only one parameter a_{ii} (Gales and Young, 2007, Rabiner and Juang, 2004).

So, the weakness of self- transition probability in the standard HMM is replaced by an explicit probability distribution P_i (d). Poisson distribution and gamma distribution are the popular choices. When self- transition geometric distribution is replaced with an explicitly probability distribution then state segment durations model more accurately but Markov assumption no longer holds. This Baum welch and Viterbi algorithm for finding optimal path may heavily depends on these HMM assumption, so they are changed from their basic form (Pylkkonen, *et al.,* 2004). This extended HMM model is called a Hidden- Semi Markov model (HSMM) represented as:

$$\alpha_t(rj) = \sum_T \sum_{i, i\neq j} \alpha_{t-T}(ri)\, a_{ij}\, P_j(d) \prod_{l=1}^{T} b_j\,(y_t - r_{T+1}) \quad \ldots(3)$$

this approach add a considerable overhead to the computational complexity because by using

Copyright 2016 IMSEC

this complete state probability history is needed instead of state probability of one time step for Viterbi algorithm. By defining the maximum state duration 'D' this computational overhead is release because by this way state probability history is needed for 'D' time stamp only.

3. Connectionist Approach

Connectionist speech recognition system is the combination of HMM and artificial neural network (ANN), In order to take the advantages of these pattern recognition techniques to improve the performance of overall ASR system. Exact pattern classification is the strength of ANN but it's not good to handle long term speech signal (Aggarwal *et al.,* 2011). On the other side superior time wrapping and decoding techniques are the strengths of HMM approach. Therefor researchers developed a new approach (connectionist approach or hybrid approach) by combining the ANN and HMM in order to construct the ASR system with improved classification capability (Rigoll, 1998). The neural network architectures are divided in to three categories (i) feed forward architecture, (ii) feed backward architecture, and (iii) self-organized architecture. In feed forward network information can move only in one direction from input to output but in recurrent network or feed backward, the information can move in all directions and that's the reason it's used to solve the more complicated problems (Aggarwal *et al.,* 2011). These two networks are trained by supervised training while self- organized networks are trained by unsupervised training methods.

The HMM / ANN framework is used for replacing the emission probability modeling technique used in HMM and use of multilayer perceptron (MLP) for the computation of output probability is the state- of- art. The back propagation algorithm is mostly used to train the MLP. Typically, sigmoidal function is used as a activation function for all the neurons in MLP, except the neurons of output layer. This sigmoidal function is derived as:

$$f(x) = \frac{1}{1+\exp(-x)} \quad \ldots(4)$$

For the output layer of 'N' units, $f(x)$ is additionally smoothed by the *softmax* function defined as:

$$f(x_i) = \frac{1}{\sum_{n=1}^{N} \exp(x_n)} \quad \ldots(5)$$

The error between actual output (O_i) and desired output (d_i) can be expressed as:

$$\epsilon = \frac{1}{2}(d_i - O_i)^2 \quad \ldots(6)$$

For single perceptron at the output $k = 1, 2, \ldots, k$.

$$\epsilon_p = \frac{1}{2}\sum_{k=1}^{k}(d_{pk} - O_{pk})^2 \quad \ldots(7)$$

Where, p is a specific pattern $p = 1, 2, \ldots, p$.

There are number of training techniques which use the basic principal of the neural posterior probability estimation, in which time delay neural network (TDNN) and recurrent neural

network (RNN) are more popular. The MLP is replaced by RNN that avoids the use of multi-frame input (Rigoll, 1998). TDNN is the intermediate solution between the spatial input model and recurrent model. In TDNN, recurrent networks are calculated in a finite period of time by using feed forward network, in which loops were replaced by explicit use of some activation values (Bourlard, and Morgan, 1994).

4. Discriminative HMM

A further refinement is to use a discriminative training approach to HMM parameter estimation. Historically, the predominant training technique which has remained constant in the rapid evolution of speech recognition is the use of maximum likelihood estimation (MLE) (Lee *et al.,* 2000). MLE is widely used as a HMM parameter estimation technique when (i) the training data is unlimited, (ii) the language model is known ,and (iii) observations are from a known family of distribution (Vertanen, 2004).

The reason behind the probability of MLE is its ability to quickly train the ASR system using convergent Baum-Welch algorithm but MLE is not guaranteed to produce the optimal result. In discriminative training methods, model parameters are optimized based on a training data set by directly formulate the class posterior probability to minimize the classification error rate. The main goal of discriminative training is to correctly discriminate the observations for the best classification results (Lee *et al.,* 2000). Discriminative training increase the considerable amount of recognition accuracy and increasingly used by the large vocabulary speech recognition systems (Young *et al.,* 2002). The aim of maximum mutual information estimation (MMIE) is to maximize the mutual information between word sequence and observation sequence. During MMI training, the objective function is defined as:

$$I = \sum_{u=1}^{U} \log P_{\lambda}(w(u)/O(u)) \quad \ldots(8)$$

The expression of mutual information (MI) is expressed as:

$$MI = \sum_{u=1}^{U} log\left[\frac{P_{\lambda}(O(u),w(u)}{P^{\bullet}(w(u)),P_{\lambda}(O(u))}\right] \quad \ldots(9)$$

$$= \sum_{u=1}^{U} \log P_{\lambda}(w(u)/O(u)) \quad -\log P^{\bullet}(w(u))$$

$$= I - \sum_{u=1}^{U} \log P^{*}(w(u)$$

Instead of maximize the mutual information; we can directly minimize the error on training data set by using minimum classification error (MCE) algorithm. The main objective of MCE algorithm is to minimize the error and it gives better result than MMI algorithm for small vocabulary tasks. However, implementing MCE on large vocabulary task is quit challenging. The objective function of MCE can be defined as

$$F_{MCE}(\lambda) = \sum_{r}^{R} = \frac{P^{k}(Sr)P^{k}(Xr|Sr)}{1\sum_{S}P^{k}(S)\,P^{k}(X_{r}|S)} \quad \ldots(10)$$

Where X_r is the speech data for r^{th} training sentence and k is the scaling factor typically less than one. The minimum phone error (MPE) algorithm uses the I-smoothing technique to increase the

Copyright 2016 IMSEC

phone level accuracy. The MPE algorithm gives the better test set performance. The objective function used in MPE algorithm, for maximization, is

$$F_{MPE}(\lambda) = \sum_{r=1}^{R} \frac{\sum_S P^k(S)\, P^k(X_r|S)\, A(SrS)}{\sum_S P^k(S)\, P^k(X_r|S)} \quad \ldots(11)$$

Where denotes a variable which may take all possible labels in a hypothesis space. The probability of observation for the given sequence denoted by $p^k(X_r|S)$.

5. *Margin Based HMM*

First time margin based classification technique was incorporated in to Hidden Markov modeling for speech recognition at York university by Jaing *et al.,* 2006. As discussed earlier, the negative aspect of MLE training algorithm could be removed by discriminative training such as maximum aposteriori (MAP) that produce the lower error rate but unfortunately, discriminative approach is limited to very specific applications (Ganapathiraju *et al.,* 1998). To overcome these limitations, researchers tend to margin based HMM which use the support vector machine (SVM) framework. SVM gets the researchers attention due to their discriminative learning and generalization capabilities. Risk minimization is the basic concept behind the SVM. SVM is a linear classifier, in which classes are separated by hyper planes but this can be used in non-linear classification problems by using kernel-trick. In which, SVM is not directly apply to the input space R^n rather apply to high dimension feature space $\mathcal{F}$, which is non-linearly related to input space: $\phi : R^n \rightarrow \mathcal{F}$ (Kriiger *et al.,* 2005).Accordingly to the literature, marginal based approaches are divided in to two categories (i) large margin approaches, (ii) soft margin approach. As show in figure 1, SVM produce the hyper plane that maximizes the margin between the two classes.

Maximize the margin is an optimization problem by using lagrangian multiplier á$_i$, which results only few training samples (support vector) affecting the hyper plane (Stadermann and Rigoll, 2011). Large margin estimation can be expressed by the following equation:

$$\tilde{\lambda}_{LME} = argmax_{\lambda} \quad min_{X_r \varepsilon D}\, d(X_r|\lambda) \quad \ldots(12)$$

Where D = $\{X_1, X_2, \ldots\ldots, X_R\}$ denotes the training data set

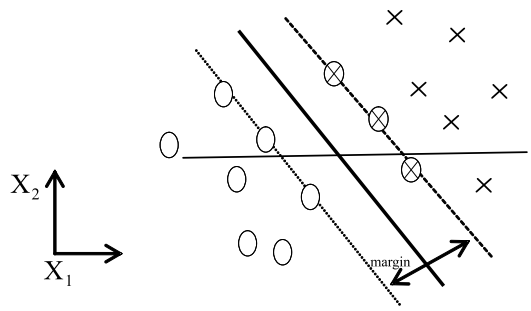


Figure 2: Linear separable problem with two class and the support vectors are circled

In LME, we first define a separation margin for each training data X_r as

$$d(X_r|\lambda)=log[P(S_r).P(X_r|S_r)]-\max_{S,S=S_r} log[P(S).P(X_r|S)]d(X_r|\lambda)$$
$$= log[P^k(S_r).\ P^k(X_r\ |S_r)] - \ldots$$
$$\ldots log[\textstyle\sum_{S,S=S_r} P^k\ (S)\ P^k\ (X_r\ |S)] \quad \ldots(13)$$

Using the margin definition in Eqn. (5.12), the LME criterion can be represented by

$$\tilde{\lambda}_{LME} = argmax_{\lambda} \quad min_{r=1..R} ln \frac{P^k(Sr)P^k(Xr|Sr)}{\sum_{S,S=S_r} P^k(S)\ P^k(Xr|S)} \quad \ldots(14)$$

Soft margin estimation (SME) was proposed by Li et al. in 2006 and shows the better generalization capability. In general, there are two methods of SME. In first method, we jointly optimize the soft margin and HMM parameters. In second method, we firstly set the soft margin and then after find the optimal number of HMM parameter (Aggarwal *et al.,* 2011).

Adaptation techniques

Acoustic model adaptation is the process of changing the acoustic model parameters used for speech recognition in such manner that ASR system recognizes the different target user by using a few utterances from the target user (Shinoda, 2011). Speech dependent recognition systems have high accuracy for one speaker but this system become worse for different target user. By using speaker adaptation technique we can improve the speech recognition accuracy by using few utterances of target user. In speaker dependent system is based on ML estimation using the EM algorithm, which cannot precisely estimate the model parameters. As a result, recognition accuracy would be much worse. *Speaker adaptation* aims to overcome these problems. An adaptation model should improve recognition accuracy with a small amount of data. Maximum *a posteriori* (MAP) estimation is used in statistical modeling and particularly, used for speaker adaptation. It estimates model parameters more robustly than ML estimation in a condition when the amount of data is small, and its estimates asymptotically approach ML estimations as the amount of data increases (Shinoda, 2011). Maximum likelihood linear regression (MLLR) based on a linear mapping between the acoustic feature spaces of different target speakers. MLLR is one of the most popular adaptation methods because it is easy to use and it performs well in most cases. Generally, MLLR is robust and well suited to unsupervised incremental adaptation. There are two main variants of MLLR: unconstrained and constrained. In MLLR, the mean vectors of the Gaussian distributions in the HMMs, $\mu=(\mu_{1...},\mu_n)$, where n is the dimension of a feature vector, are updated according to the following transformation:

$$\mu = A\ \mu + b \quad \ldots(15)$$

Here, A is an n x n matrix, and b is a n-dimensional the human auditory system.

Feature reduction techniques

An important task for any pattern recognition problem is to find a good feature space, which should be compact and contain the richest possible discriminant information. Feature dimensions which contain less discriminant information should be discarded because their

Copyright 2016 IMSEC

existence not only slows down the classification process but also degrades the system performance. This step is aimed at incorporating the techniques which project the features into low dimensional subspace, while preserving discriminative information. To derive such a good feature subspace is through finding a linear transformation from the original feature space $\mathcal{R}''$ to a new low dimensional one $\mathcal{R}'''$ such that:

$$F_M(T) = \theta_T F_N(T) \quad \ldots(16)$$

where $F_M(T)$ is the feature vector in the transformed feature space and $F_N(T)$ is the feature vector in original feature space and M is the target feature space's dimensionality.

These transformations may be supervised or unsupervised. Unsupervised methods are variable directed techniques which make no assumption about the existence of classes or groupings within the data. The simplest form of this criterion is PCA, which selects the orthogonal projections of the data that maximizes the total variance in the projected subspace. Simply selected subspaces that yield large variances does not necessarily yield subspaces that discriminant between the classes. To address this problem, supervised approaches such as LDA criterion can be used. In LDA the objective is to increase the ratio of the between class variance to the average within class variance for each dimension. An alternative extension to LDA is HLDA which outperform approaches like LDA, but are more computationally intensive and require more memory. Full covariance matrix statistics for each component are required to estimate an HLDA transform, whereas only the average within and between class covariance matrices are required for LDA. In a Gaussian distribution framework, it maximizes the likelihood of all the training data in the transformed space and each training sample contributes equally to the objective function. It assumes that n-dimensional original feature space can be split into two statistically independent subspaces. The goal of HLDA is to find the transform matrix $\boldsymbol{A}$ that maps an n-dimensional feature vector $\boldsymbol{x}$ to a new $\boldsymbol{y}$:

$$Y = Ax = \begin{bmatrix} A_m x \\ A_{n-m} x \end{bmatrix} = \begin{bmatrix} y_m \\ y_{n-m} \end{bmatrix} \quad \ldots(17)$$

where A_m is a matrix consisting of first m rows of $n \times n$ matrix $\boldsymbol{A}$, A_{n-m} consists of remaining n—m rows, y_m are deemed to be those useful dimensions in the rotated space and $y_{m\text{-}n}$ are the nuisance dimensions.

To find the optimal HLDA, a model is created and its parameters are estimated in maximum likelihood framework. Let x_s denote a training sample and $g(s)$ indicates its class label. In the HLDA framework, each class is modelled in single Gaussian. The probability density of x_s in the transformed space is given as:

$$p(x_s) = \frac{det(A)}{\sqrt{(2\pi)^n |\Sigma_{g(s)}}} exp\left(\frac{-(Ax_s - \mu_{g(s)})^T \Sigma_{g(s)} (Ax_s - \mu_{g(s)})}{2}\right) \quad \ldots(18)$$

where $\mu_{g(s)}$ and $\Sigma_{g(s)}$ are the mean vector and covariance matrix of class $g(s)$ in the transformation space. The objective function is defined as the log-likelihood of all the training samples as follows:

$$\mathcal{L}(A; \{x_s\}) = \Sigma_{\forall s} \log p(x_s) \quad \ldots(19)$$

The HLDA solution maximizes the objective function

$$A = \text{argmax}_A L(A, \{x_S\}) \quad \ldots(20)$$

Challengs in Hindi ASR:

There are various challenges in automatic Hindi speech recognition. This section includes various challenges in Hindi speech recognition and theirs solutions. Some challenges are:

Lack of Hindi language database

Building the Hindi speech database is the first step toward a Hindi speech recognition system. The performance of Hindi automatic speech recognition (ASR) system is heavily depends on the speech database which is used to train the Hindi ASR system. The main reason behind the abysmal condition of Hindi speech research as compared to the research of other languages of developed countries is the lack of Hindi speech database. A very few projects are carried out by some IIT's with the collaboration of other universities but these databases are not open source for the research scholar of other universities. For example, Hindi speech database for ASR system was built by C-DAC Noida. This database contains the 26 Hr. of speech recording which contain 8567 sentences by the uniformly distributed speakers over all age groups from 17 to 60 years (Arora, *et al,.* 2010). Same as IIIT Hyderabad, IIT kharagpur, Utkal University, KIIT Bhubaneswar build Hindi speech database for research projects sanctioned by the Technology development for Indian languages (TDIL) (Godambe, *et al.,* 2011). This situation is the matter of great anxiety for the research of Hindi speech recognition. The availability of one unique Hindi speech database is must to overcome this scenario.

Variation in Hindi accent from region to region

India is the multilingual society where about 1652 native languages are used, while Hindi is a national language of India. There are 17 other official languages recognized by the constitution of India (Shrishrimal *et al.,* 2012). Clearly, there are lots of variations in Hindi accent in a multilingual society like India. This factor plays an important role in slow progress of Hindi speech research. For example, if Indian government wants to apply a Hindi ASR system in government offices then there will be need of large number of speakers from different geographical regions of our country.

Similar utterances of different phoneme

Continuous Hindi speech recognition gives better results if acoustic modeling is based on phoneme level except the whole word. This mono-phone based acoustic modeling for Hindi speech recognition has enormity advantage over word level but at the same time there is some adverse condition that gives the negative impact. In Hindi language character set there are some phoneme which has almost same utterances like (त, थ), (ड, ढ), (ण, न) etc. and this situation become worse when we want to recognize the word like डोल, ढोल, रण, रन etc. due to same utterances of these phonemes ASR system may confuse and ultimately overall degradation in the performance of ASR system.

Computational overhead

To implement the large vocabulary Hindi speech recognition system gigantic speech database are needed

Copyright 2016 IMSEC

to cover all the speaker variability by using large number of speakers from different geographical regions. To handle such massive database for real time application is a big challenge. According to the literature, it has been proved that mixture of multivariate Gaussian to represent emission distribution of HMM. A range of 2-64 mixture components have been found useful for speech recognition (Aggarwal *et al.,* 2011). In such systems, more than 50% computational load is due to use of GMM. That's why LVCSR system could not be able to apply in real time, mainly because of slower processing due to computational overhead. As the vocabulary size of ASR system increases, this overhead will also be increase.

Microphone quality and distance

The quality of microphone improves the performance of speech recognition system. The highly accurate microphone has the ability to reject the external noise. Therefore close talking good quality unidirectional microphone should be used to increase the ASR system accuracy which has the ability to ignore extraneous noises. There are various kind of speech noises which substantially change the original speech signal. Some of them are ambient noises (office machinery, human conversation, industrial plat etc.), distortion in signal and variation in articulation (Haton, 2004). So, unidirectional microphone with excellent sound quality is required to minimize the effect of these noises. Distance between speaker lips and microphone is also plays an important role to minimize the background noise, room reverberation etc. The distance should be about 10 cm from the lips of speaker.

Labeling issues in Hindi speech utterances

Although each and every researchers of speech recognition has performed some manual annotation of speech corpora but very few has been spent their effort on examining the reliability of manual annotation criteria. However, in order to train and test the ASR system manually labeled speech data are needed as a reference sets and this manually labeled speech data is used to measure the accuracy of ASR system. An abrupt change in the signal marked at starting and ending point confused whether it indicated boundary or not. If labeling is not done in an good manner then at last ASR system could not be able to recognize the words which have almost same pronunciation like ली की, राजा रजा, डीग जीत etc. so, accurate labeling of speech utterances is an important component of an ASR system. To achieve the high performance ASR system, the labeling procedure should be repeatedly followed until ASR system able to distinguish the same utterances word.

Conclusion

Speech as a communication medium between man and machine plays an important role in today's life. Despite of much technological development, human- computer interaction (HCI) is not as close to human- human interaction. The lot of efforts has to be done to fill this gap. This review paper attempts to give a current status of speech recognition. Due to resource deficiency, the research in Indian languages like Hindi, Tamil and Bengali are still in its infancy stage. Most of the work in Hindi ASR has been done in last one decade but still there is lots of work to be done. Due to its fascinating nature, ASR attracts many scientist and researchers to minimize the gap between man- machine interactions. This research paper gives the detail description of state- of- the – art approaches in the field of speech recognition and various difficulties in Hindi ASR.

References

Aggarwal, R. K. 2014. Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition. *Ph.D. thesis, Department of Computer Engineering , National Institute of Technology, Kurukshetra*.

Mohri, M. and Pereira, F. 2002. Weighted Finite State Transducers in Speech Recognition. in CSL. *Computer Speech & Language,* 16(1), 69-88.

Deoras, A., Tur, G., Sarikaya, R., Tur, D.H. 2013. Joint Discriminative Decoding of Words and Semantic Tags for Spoken Language Understanding. In *IEEE TASLP*.

Mikolov, T. et al. 2013. Joint Discriminative Decoding of Words and Semantic Tags for Spoken Language Understanding. In *IEEE TASLP*.

Aggarwal,R. K. and Dave, M. 2011. Acoustic modeling problem for automatic speech recognition system: Advances and Refinements (Part II). International Journal of Speech Technology, *Springer*, 14 (2).

Gales, M. and Young, S. 2007.The Application of Hidden Markov Models in Speech recognition. *Foundation and trends in Signal Processing*, 1 (3), 195-304.

Rabiner, L.R. and Juang, B. H. 2004. Statistical Methods for the Recognition and Understanding of Speech. *Encyclopedia of Language and Linguistics*.

Pylkkönen,J and Kurimo,M. 2004. Duration Modeling Techniques for Continuous Speech Recognition. *In the Proceedings of the 8th International Conference on Spoken Language Processing* (Interspeech 2004), 385-388, Jeju Island, Korea.

Rigoll, G. 1998. Hybrid Speech Recognition System: A real alternative to traditional approaches?. Survey Lecture, *Proceeding International Workshop Speech and Computer* (SPECOM'98).

Bourlard,H.,Morgan, N. 1994. Connectionist speech recognition: A hybrid approach. *Kluwer Academic*, Boston.

Lee, H. C., Soong, Frank, K.., Paliwal, Kuldip., 2000. Automatic speech and speaker recognition. *Kluwer academic publishers*.

Vertanen,K. 2004. An Overview of Discriminative Training for Speech Recognition. *University of Cambridge*.

Young, S., Evermann, Gunnar., Kershaw, Dan., Moore Gareth., Odell Julian., et al. 2002. The HTK Book. Available: http://htk.eng.cam.ac.uk.

Jaing, H., Li, Xinwei., Liu, Chaojun., 2006. Large margin Hidden Markov Models for speech recognition. IEEE Transactions on Audio, *Speech and Language Processing*, 14(5), 1584-1589.

Copyright 2016 IMSEC

Ganapathiraju, A. et al. 1998. Support Vector machines for speech recognition. Proceeding the of International Conference on Spoken Language Processing, 1998.

Koechlin, E., and Jubault, T. 2006. Speech recognition with support vector machines in a hybrid system. In the Proceeding of INTERSPEECH, 993-996.

Stadermann, J., Rigoll,G. 2011. A Hybrid SVM/HMM acoustic modeling approach to automatic speech recognition. *In the Proceeding of APSIPA AS*C.

Shinoda, K. 2011.Speaker adaptation techniques for automatic speech recognition. *In Procedding of APSIPA ASC*.

Arora,S. et al. 2010. Hindi ASR for Travel Domain. *In Proceedings of OCOCOSDA*, Kathmandu, Nepal. Tejas, Godambe and Samudravijaya, K. 2011. Speech Data Acquisition for Voice based Agriculture Information Retrieval. *In the Proceeding of 39th All India DLA Conference*, Punjabi University, Patiala.

Shrishrimal,P.P., Deshmukh, R.R., Waghmare, Vishal, B., 2012. Indian Language Speech Database: A Review. *International Journal of Computer Applications*, 47(5).

Haton,J. 2004. Automatic Speech Recognition: A Review. *Enterprise Information Systems* V, 6-11.