International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 1, January - February 2017 **ISSN 2278-6856**

A Review of the Secure Offline Synchronization Methods of Web and Mobile users

Jagadish R M[1], Dr. L Swarna Jyothi[2]

[1]Faculty of Engineering College, Dept. of Computer Science, BITM, Ballari, Karnataka, India

[2]Professor of Engineering College, RRCE, Bangalore, Karnataka, India

ABSTRACT

In this era, a large number of people exchange their personal data and most important information. With the development of the certain computer applications that allow the user to transfer and store their personal data in the form of image, video, text files etc. directly on the web server by using readily available data synchronization techniques. Only the users with having the specific password for the user account will be granted access to the resources. In this paper, the ways of data synchronization through which one can synchronize their data on the web server and can replace their old data with recent updates of the data has been described. This paper will also cover the information regarding offline data synchronization along with the data deduplication technique that helps in compression and removal of redundant data. Deduplication techniques also improve the storage utilization. In this research work various different approaches of the data deduplication with encryption are reviewed. At last, the introduction to the basic structure of the speech recognition, types of speech recognition, need and implementation of the speech recognition, along with the methods and working of the efficient models of the speech recognition are included in this paper.

Keywords: Data de-duplication, Synchronization, speech recognition

1. INTRODUCTION

In computer science, Synchronization refers to one of two distinct but related concepts: synchronization of processes and synchronization of data. Process synchronization refers to the idea that multiple processes are to join up or hand shake at a certain point, in order to reach an agreement or commit to a certain sequence of action.

Data synchronization refers to the idea of keeping multiple copies of a dataset in coherence with one another, or to maintain data integrity. Data synchronization is the process for the consistency of the data from a source to destination data storage and vice versa, and to establish a continuous harmonization of data over time. Data synchronization technology to synchronize a single record between two or more devices, changes is automatically copied from one page to another. For example, the user's contact list can be synchronized of one mobile device with other mobile devices or devices. Data synchronization may be a local synchronization, in which the device and the computer are adjacent and transmitted data or a remote synchronization occur when a user is mobile and the data is synchronized through the mobile telephone network.

The process of data synchronization typically increases the security by insisting on the use of strong passwords that cannot be easily decrypt. It also reduces the number of password-related requests for help, which are reported as biggest demand on help desk resources.

Cloud computing has recently emerged as a popular business model for utility through data synchronization methods [2]. Cloud storage is one of the services in cloud computing which provides virtualized storage on emend to customers. Customer can access the resources from cloud storage anytime through Internet from anywhere without worrying about any maintenance or management of actual resources. This service allows the users to wirelessly back-up their data from devices to Sky Drive and retrieve them. It can be used in many different ways. For example, customers can use cloud storage as a backup service, or as opposed to maintaining their own storage disks. Organisations can move their archival storage to the cloud which they can achieve more capacity at the low-cost, rather than buying additional physical storage. Applications running in the cloud also require temporary or permanent data storage in order to support the applications.

2. DATA DEDUPLICATION

Data deduplication is basically a compression technique for removing redundant data. It improves the storage utilization and can also be designed to work on primary storage as well as on secondary storage.

Data Deduplication, is a file system feature that only saves unique data segments to save space. It has been most popular and successful for secondary storage systems (backup and archival). Its basic principle is to filter the data block to find the same data block, and the only instance of a pointer to point to replace. It is a technology to identify duplicate data, eliminate redundancy and reduce the need to transfer or store the data in the overall capacity.Deduplication is widely is used various applications like backup, metadata management, primary storage, etc. for storage optimization. [1]

Nowadays, data deduplication is widely used by various cloud storage providers like Dropbox, Amazon S3, Google Drive, etc. Data once deployed to cloud servers, it is beyond the security premises of the data owner, thus most of them prefer to outsource their in an encrypted format. Data encryption by data owners eliminates cloud service providers chance of de-duplicating it, since encryption and deduplication techniques have conflicting strategies, i.e., data encryption with a key convert's data into an unidentifiable format called cipher text thus encrypting, even the same data, with different keys may result in different cipher texts, making deduplication less feasible. However, performing encryption is essential to make data secure, at the same time, performing deduplication is essential for achieving optimized storage. Therefore, deduplication and encryption need to work in hand to hand to ensure secure and optimized storage.

Table 1: Various Approaches for Data Deduplication Implementation with Encryption

Approach	Encryption Scheme	Deduplication Strategy used
Message-locked encryption and secure deduplication	Message locked encryption	File level
BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication	Block Level Message locked encryption	Dual level: File level and Block level
HEDup: Secure Deduplication with Homomorphic Encryption	Homomorphic encryption	File level
DupLESS: Server-Aided Encryption for Deduplicated Storage	Enhanced Message level encryption to support security against Brute force attack	File level
ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage	Convergent encryption with added access control	File level
Secure Deduplication with Efficient and Reliable Convergent Key Management	Convergent encryption	Block level
Twin clouds: An architecture for secure cloud Computing	Convergent encryption	File level
A hybrid cloud approach for secure authorized deduplication	Convergent encryption	File level
Secure Data Deduplication	Convergent encryption	File level
A secure data deduplication scheme for cloud Storage	Symmetric encryption on data categorized based on popularity	File level
Secure Distributed Deduplication Systems with Improved Reliability	Deterministic secret sharing scheme	File level and fine grained block level

SecDep: A User-Aware Efficient Fine-Grained Secure Deduplication Scheme with Multi-Level Key Management	User aware convergent encryption	File level and chunk level

2.1 DATA DEDUPLICATION LEVELS

At present there are fundamentally two main levels of the data duplication techniques that can be optimized for storage capacity. These levels are defined briefly as:

A. File Level Data De-duplication Strategy

In this level, the deduplication is performed over a single file and it eliminates the duplicate copies of the same file. The file checking function is based on their hash values. The hash numbers are comparatively easier to generate so it does not require more processing power. Based on those hash numbers duplicated files are identified. I.e. if two or more files having the same hash values, they assumed to have the similar contents and only one copy of file is to be stored. The searches for any files that are exactly alike and stores only one copy, where placing "pointers" in place of the other copies said to be more efficient than no deduplication whatsoever. Even a single slight alter to the file will effect in a supplementary replica being stored.

B.Block Level Data de-duplication Strategy

In this level of deduplication, it firstly divides the files into the blocks and stores only one copy of each block. It may also use fixed-sized blocks otherwise variable-sized chunks. They compute hash values for each block for examining duplication blocks and then it eliminates duplicate blocks of data that occur in non-identical files. As the name indicates it performed over blocks and analyzes entire blocks of data and then Allows for granularity without being overly time consuming and resource intensive. While compared it to with whole file, the block level deduplication eliminates the tiny unnecessary chunk of data. The equivalent deduplication algorithm is used by each one and all file system in block level deduplication.

3.OFFLINE DATA SYNCHRONIZATION

The evolutions in the field of computing and mobile technologies have certainly provide an immense surge to a new environment in which computers and multimedia cellphones have completely dominated the human activities in all aspects. The latest technology embedded in these devices allows an individual user to store their personal data over internet through data synchronization techniques. Instead of bulky and immovable Personal Computers (PC), there are various categories of devices that are available these days such as Notebooks, Smartphones, I-pads, etc. These devices are much smaller than the previous versions of the Computers and have the feature of portability that allows user to access their resources from anywhere around the world.

Many mobile applications are data-centric, and are designed to replace pocket atlases, dictionaries, and references, as well as create new digital pocket references for data that changes dynamically by leveraging technologies that did not exist in these form factors before.

This paper focuses solely on patterns related to data synchronization, which involves ensuring consistency among data from a mobile source device to a target data storage service and vice versa. In some of the datasets it is very difficult to load and save the recently available data that an application can grasp on the device. So, to carb this issue, a specific strategy must be applied to synchronize the necessary data and ensure the user to get a recent and updated data.

Sync Framework is a synchronization stage that empowers engineers to add synchronization capacities to applications, administrations, and devices. It takes care of the issue of how to synchronize any kind of information in any store utilizing any convention over any topology. The strength of this framework is that it can connect different frameworks into one frame. Sync Framework can also be used to access the database offline. Synchronization can be used for one-to-one or one-to-many units, which function offline. The synchronization service is incorporated in VS2008 and VS2010.

The versatile applications have the capacity work in the offline mode i.e. disconnected from the net. For any information based application, it is very essential to run or to be executed online as well as offline and can be accessed anytime from anywhere without any inconvenience related to the server. The synchronization is the most important technique which is utilized in the process that involves the upgradation of the application on the web server. It makes the application fundamentally more helpful. For instance, exchanging information to a web server makes a helpful reinforcement, and exchanging information from a server makes it accessible to the client although when the gadget is logged off. [6]

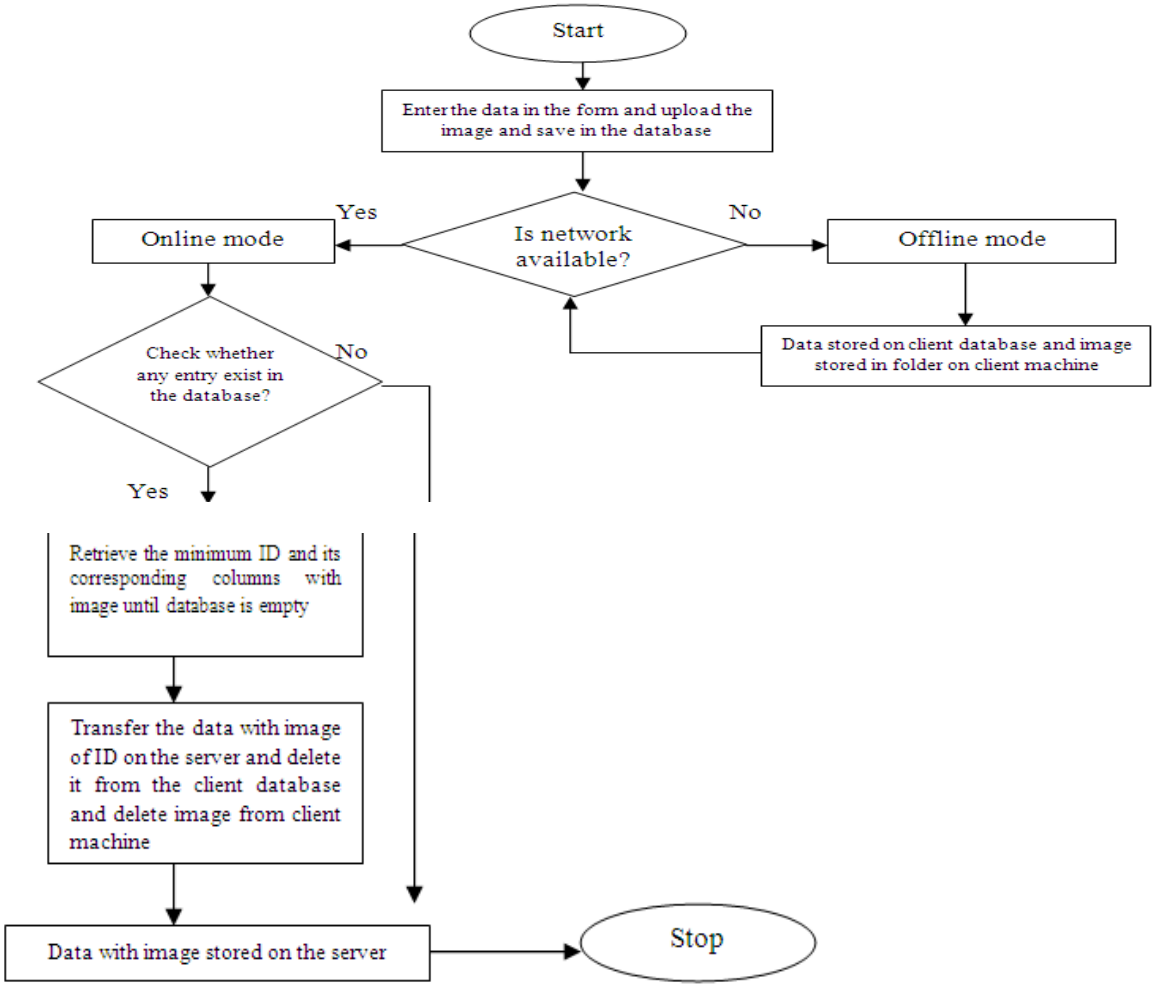

Figure 1: Flowchart of Data Synchronization

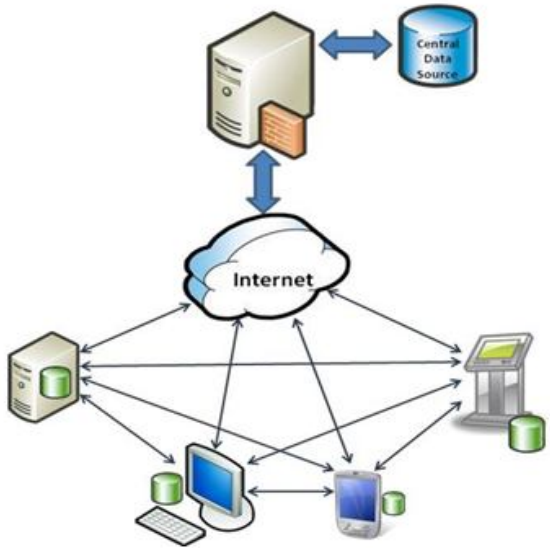

Figure 2: Sync Framework data synchronization

For building and managing a reliable synchronization through the web servers, a potential algorithm design is used which besides the online operations of the application also allow to work in the offline mode, in the absence of Internet or due to some power related issues and make these interruptions unnoticed by the user.The two major examples of the most widely used offline data synchronization technique are as follows:

i. **YouTube Offline:**YouTube is the most popular software application among the youngsters. It provides a feature that allows the users to save the videos in their user accounts and then saved videos can be seen any time in the future.

ii. **Cloud Network:** The cloud networks provides a feature that allows offline data synchronization at the IAAS Layer. Cloud network also ensure the high end security of the saved data to avoid any kind of discrepancy. Only the individuals that have the password to the users account are provided the access to the data.

4.SPEECH RECOGNITION TECHNIQUE

The speech is one of the most important and primary mode of communication among the humans. It is natural and efficient techniques that enable humans to exchange information to each other. Speech Recognition is the ability of machine or computer generated program to identify various sounds of the words and phrases that are present in a human speech and convert them into a machine readable format.

To make a real "intelligent computer", it is important that the machine can hear, understand, and act upon spoken information, and also speak to complete the information exchange. Speech Recognition (is also known as Automatic Speech Recognition (ASR) or computer speech recognition. It is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer.

4.1 Structure of Basic Speech Recognition System

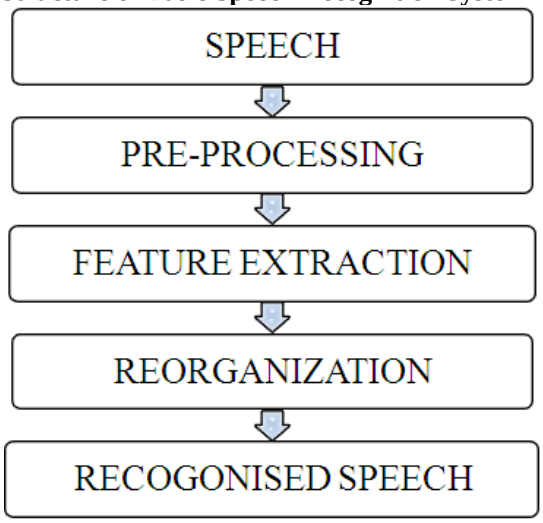

Figure 3: General Structure of Speech Recognition System

i. Speech Capturing

The input speech can be captured with the help of microphone. The sound card that is already installed in the computer captures the speech signal and turn the analogue signals into digital signal. The sound card then makes the understandable format of speech for the computer and record the sound. Now using the MCI commands, the frequency and size of the speech can be control at instance.

ii. Pre-Processing

After the process of capturing the sound sample is completed the speech sample can be available continuously at any instance. The Next step after the sound capturing includes the pre-processing of the captured samples to make it available for feature extraction and recognition. This process involves the following steps

a. Background Noise and Silence Removing
b. Pre-emphasis Filter
c. Blocking into Frames
d. Windowing

iii. Feature Extraction

Feature extraction is mainly defined as the process of parameterization of the available speech sample in the terms of vectors that can be utilize for purpose of recognition while maintaining the discriminating power of the signal. These feature vectors should not change with the speaker i.e. the features should be same for the same statement by different speakers. These features can be extracted by using several methods, for example digital filter, Fourier Transformation or Linear Predictive Coding

Table 2: List of Techniques with their properties for Feature Extraction

Sr.No	Method	Property	Procedure for Implementation
1	Principal Component analysis(PCA)	Non-linear feature extraction method, Linear map, fast, eigenvector-based	Traditional, eigenvector base method, also known as karhuneu-Loeve expansion; good for Gaussian data
2	Linear DiscriminateAnalysis(LDA)	Non-linear feature extraction method, Supervised linear map; fast, eigenvector-based	Better than PCA for classification
3	Independent ComponentAnalysis (ICA)	Non-linear feature extraction method, Linear map, iterative	Blind course separation, used for de-mixing non- Gaussian

		non- Gaussian	distributed sources(features)
4	Linear Predictive coding	Static feature extraction method,10 to 16 lower order coefficient,	It is used for feature Extraction at lower order
5	Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope[7]
6	Mel-frequencyscale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a Subjective frequency Scale i.e. Mel-frequency Scale.
7	Filter bank analysis	Filters tuned required frequencies	
8	Mel-frequency cepstrum(MFFCs)	Power spectrum is computed by performing Fourier Analysis	This method is used for find our features
9	Kernel based feature extraction method	Non-linear transformations	Dimensionality reduction leads to better classification and it is used to redundant features, and improvement inclassification error.[8]
10	Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform
11	Dynamic feature extractions i)LPC ii)MFCCs	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and MFCCs coefficients	It is used by dynamic or runtime Feature
12	Spectral subtraction	Robust Feature extraction method	It is used basis on Spectrogram[9]
13	Cepstral mean subtraction	Robust Feature extraction	It is same as MFCC but working on Mean statically parameter
14	RASTA filtering	For Noisy speech	It is find out Feature in Noisy data
15	Integrated Phoneme subspace method (Compound Method)	A transformation based on PCA+LDA+ICA	Higher Accuracy than the existingMethods[10]

iv. Recognition

This stage of the speech recognition process is further divided in two different phases, i.e. the Training Phase and the Testing Phase.

The training phase is simply defined as the process that is similar to learning process of children. A child should experience a phenomenon many times and with a wide variability before being able to recognize it. The current speech recognition technology does not allow real-time implementation of models comparable to human

complexity. This means that the variability of speech must be limited to achieve proper results.
On the other hand, in the testing phase, an unknown utterance is scored over the reference patterns. The word corresponding to the reference pattern closest to the unknown pattern is the word recognized.

4.2 Types of Speech Recognition
The Speech Recognition can be further categorised into various different classes by describing that specific types of utterances they have the ability to recognize. These classes are classified as following:

i. Isolated Words
Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

ii. Connected Words
Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allows separate utterances to be 'run-together' with a minimal pause between them.

iii. Continuous Speech
Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. (Basically, it's computer dictation). Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries

iv. Spontaneous Speech
At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

4.3 Models of Speech Recognition

i. Hidden Markov Model (HMM)
This is a mathematical way to recognize speech. It was observed by a different set of probability processes that generate a sequence of observations, although it would not be possible to directly observe (hidden) the underlying stochastic process and the double-buried stochastic process. Probabilistic Modelling includes the use of probabilistic models to handle uncertain or incomplete information. Recognition in speech leads to uncertainty and incompleteness of many sources; For example, the change in speaker sound contextual effect disagree word confusion. Therefore, a probability model for speech recognition is particularly suitable. A hidden Markov model is a collection of connected states with transitions. Each transition gives two probability contributions: when the transition occurs, the finite state of probability of the emission of each output symbol of the alphabet defines the transition probability and gives the opportunity to take the transition output probability. The problem of HMM evaluation decodes the problem learning problem.

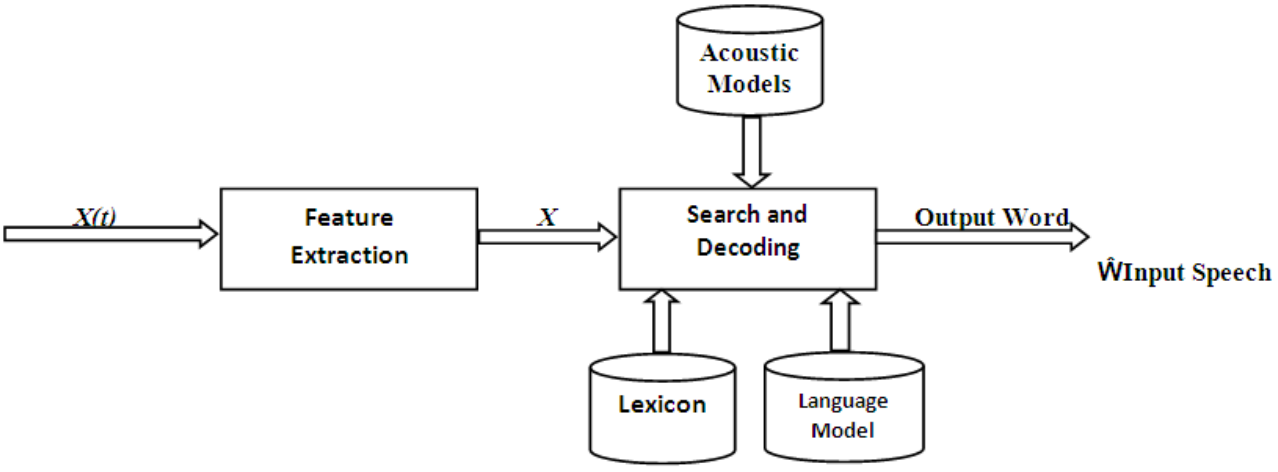

Figure 4: HMM Method Working

i. Dynamic Time Wrap (DTW)

DTW is a method that allows a computer to find an optimal match between two given sequences .It is a template-based approach. [11]. In order to understand DTW, two concepts need to be dealt with[12]:

i. **Features: -** the information in each signal has to be represented in some manner.
ii. **Distance: -** some form of metric has to be used in order to obtain a match path. There are two types of distances
   a) **Local Distance:** The computational difference between a feature of one signal and a feature of the other is called Local distance.
   b) **Global Distance:** The overall computational difference between an entire signal and another signal of possibly different length is called Global distance.

Dynamic Time Wrap is based on the two concepts defined as:

Symmetrical DTW
Language is a time-dependent process. Multiple representations of the same word are likely to have different durations, and different expressions of the same word at the same time, distinguish them because different parts of the words are possible at different rates.
• The accompanying path cannot go back in time;
• Each frame of input uses the appropriate path;
• A combination that provides a total distance value for a local distance value.

It is known as dynamic programming (DP). A dynamic time warp (DTW) that is frequently called when applied to speech recognition based on a template. The DP guarantees to find the minimum distance path through the matrix while minimizing the amount of computation.

Asymmetrical DTW
Each frame in the input frame is used only once. This means that the length of the template fails to normalize and you do not need to add two diagonal crosses to the local distance. This method is called asymmetric dynamic programming.

5.CONCLUSION

In this research work we studied various techniques, models, types, etc. for the purpose of defining the data deduplication, data synchronisation (online and offline) and speech recognition Techniques. Deduplication is a method available in cloud storage for saving bandwidth and storage capacity. But, deduplication is less feasible with encrypted data since, different key encryptions convert same data into different formats. In this paper various methods are discussed where deduplication methods are carried out on encrypted data in a large storage area. Most of the methods studied here work on the basis of convergent encryption, which is a simple approach that makes deduplication compatible with encrypted data. In this information dense world, we cannot compromise on both security and duplication of data across storage areas. An efficient Strategy should be proposed which will enhance storage optimization without negotiating on encryption method; by providing deduplication technique in data storage servers where the available data is encrypted.

The objective of the research is to provide an algorithm to solve the problem that when all clients are reliant on a single server. If that database becomes unavailable due to planned server downtime or from server failures, all of the remote workers will be disconnected from their data. Data is stored on their system (user system). When the user connected to the internet data automatically sink from their client system to the server in serial order. In the future, application developers should work with serial of data sink should be in order between offline and online application.

Speech is the primary, and the most convenient means of communication between people. Whether due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech and speaker recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades. In whole architecture of speech recognition, HMM is just one block which help in creating search graph. It work in tandem with other block such as front-end, language model, lexicon to achieve desired goal. The purpose of HMM is to map feature vector to some representable state and emit symbol, concatenation of which gives desired phoneme sequence.

REFERENCES

[1]. Amrita Upadhyay, Pratibha R Balihalli, Shashibhusha Ivaturi and Shisha Rao, “Deduplication and Compression Techniques in Cloud Design”, In International systems Conference (*SysCon*), 2012, IEEE 978-1-4673-0750-5/12, pp. 1-6.
[2]. Sudha S and Brindha K, “Data Synchronization Using Cloud Storage”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 11, ISSN: 2277 128X, November 2012, Available online at: www.ijarcsse.com
[3]. Bolosky WJ, Corbin S, Goebel D, Douceur JR, “Single instance storage in Windows 2000”, In Proceedings of the 4th Usenix Windows System

Symp. Berkeley: USENIX Association, 2000, pp. 13-24.

[4]. Jorge Guerra, Luis Useche, Medha Bhadkarnkar, Ricardo Koller, and Raju Rangaswami. "The Case for Active Block Layer Extensions", ACM Operating Systems Review, Vol. 42 No. 6, October 2008.

[5]. Austin Clements, Irfan Ahmad, Murali Vilayannur, and Jinyuan Li, "Decentralized deduplication in san cluster file systems", In Proceedings of the USENIX Annual Technical Conference, June 2009, pp. 101-114.

[6]. Naveen Malhotra, Anjali Chaudhary(2014), "Implementation of Database Synchronization Technique between Client and Server ",International Journal Of Engineering And Computer Science, Volume 03 Issue 07, Page No. 7070-7073

[7]. M.A.Anusuya , S.K.Katti "Speech Recognition by Machine: A Review" International journal of computer science and Information Security 2009.

[8]. Kenneth Thomas Schutte "Parts-based Models and Local Features for Automatic Speech Recognition" B.S., University of Illinois at Urbana-Champaign (2001) S.M., Massachusetts Institute of Technology (2003).

[9]. W. M. Campbell_, D. E. Sturim W. Shen D. A. Reynolds_, J. Navr´atily "The MIT- LL/IBM Speaker recognition System using High performance reduced Complexity recognition" MIT Lincoln Laboratory IBM 2006

[10]. Sannella, M Speaker recognition Project Report report" From http://cs.joensuu.fi/pages/tkinnu/research/index.html Viewed 23 Feb. 2010

[11]. D.G. Bhalke, C.B.R. Rao, and D. S. Bormane, "Dynamic Time Warping Technique for Musical Instrument Recognition for Isolated Notes", IEEE International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), 2011, Tamil Nadu, pp.768-771.

[12]. L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, 2010, Vol. 2, Issue 3, pp. 138-143.