IJCSC 0973-7391 Volume 7 • Number 2 March 2016 - Sept 2016 pp. 126-135 available online at www.csjournalss.com

Enhancing the Performance of Audio Visual Speech Recognition Using Deep Learning Techniques

Ankan Dutta, SharadaValiveti, Gaurang Raval
Dept. of Comp. Sc. & Engg. Nirma University, Ahmedabad, India
14mcei03@nirmauni.ac.in, sharada.valiveti@nirmauni.ac.in, gaurang.raval@nirmauni.ac.in

Abstract: Speech recognition is a very challenging task as there is huge requirement of audio visual speech analysis inmultimedia content generation and multimedia forensics. With increasing use of audio and video morphing for tampering theindividual's reputation, the need for performing the analysis of the given multimedia content has gained importance among the researching fraternity. This paper focuses on the concepts of machine learning and applicability of deep learning for speech recognition systems like using Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Deep De-Noising Auto-Encoders, Convolutional Neural Networks (CNNs), and Multi-Stream Hidden Markov Models (MSHMMs). An Audio Video Speech Recognition (AVSR) system architecture is proposed. Integrating the De-noising Auto-Encoder and Convolutional Neural network increases the reliability and robustness of the recognition system. This architecture is implemented in two phases. In this paper, tools, libraries and dataset required for audio speech recognition portion and visual speech recognition portion are discussed and then integrate these trained models using MSHMMs to build a (noise) robust and highly reliable audio video speech recognition system.
Index Terms: Audio feature extraction, Visual feature extraction, Audio visual speech recognition, Deep Learning, Convolutional Neural Network, Deep De-noising Auto-Encoder, Multi Stream Hidden Markov Model.

I.INTRODUCTION

A. Present Scenario

Its is a very difficult and challenging task to build an Automatic Speech Recognition System with high reliability and robustness. There is high variance in speech signals as different speakers have different accents, pronunciations and dialects [18]. Speakers speaking at different rates and emotional states add on to the difficulty of recognition process. Additional variability in speech signals are due to environmental noises and different recording devices. Figure 1 shows the general block diagram of the Automatic Speech Recognition (ASR) systems [18].

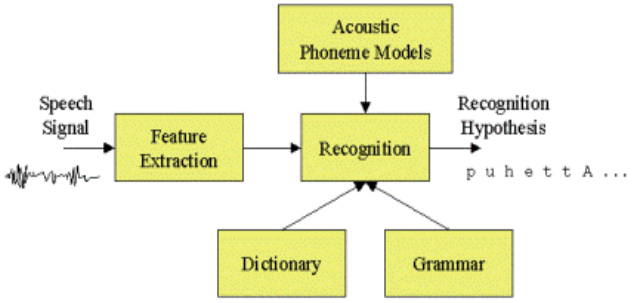

Fig. 1. General Block Diagram of ASR Systems

B. Motivation

The main idea for developing AVSR system is that we will use visual features which are derived from speaker's lip motion movements to correct corrupted audio speech features. Its isvery important to choose the audio features and visual features Fig. 1. General Block Diagram of ASR Systems carefully, because the performance of the system is affected by it [18]. Recently, in the Machine Learning domain, it is observed that Deep Learning techniques have huge impact in improving the efficiency and robustness of the underlying application. Using Deep De-noising Auto Encoders in audio feature extraction has increased the efficiency of the speech recognition systems. When Convolutional Neural Networks (CNN) are used in visual feature extraction phase, efficiency and robustness of the recognition engine improves substantially. So, knowing this fact, a new architecture is designed using both the models integrated using GMM-HMMs and MSHMMs[18].

C. Performance Evaluation

Performance of speech recognition systems is measured by Word Error Rate E, defined as,

$$E = \frac{S + I + D}{\mathrm{N}}$$

where, N is the total number of words in the test set; S, I and D are the total number of substitutions, insertions and deletions respectively [18].

II. LITERATURE REVIEW

In this section the research work contributed by the researchers across the globe in the domain of audio visual speech recognition is discussed. This technique generally involves audio feature extraction, visual feature extraction and integration of audio and visual systems. The architecture also suggests the same.

A. Mechanisms for Audio Features Extraction

Following factors are the barriers in creating an efficient and accurate ASR system :

- Limited computational power with GPUs and CPUs were
- unable to maintain multiple threads to increase speed by
- parallel processing.
- Systems have limited memory.
- Less use of large scale and distributed databases.
- Lack of advance machine learning algorithms and optimization
- strategies [18].

These days the above problems are not critical. Deep Learning algorithm are used in automatic speech recognition system. Mel Frequency Cepstral Coefficients (MFCCs) are the widely used standard for ASR systems for many years. When normal audio features are compared to denoised MFCC audio features, it is found that denoised MFCCs when applied gives higher word recognition rate gain [18].

Following are the two approaches of ASR system design:

- Using Deep Neural Network Hidden Markov Models(DNN-HMMs) in place of conventional Gaussian MixtureModel Hidden Markov Models (GMM-HMMs) in theASR systems significantly increase the efficiency andaccuracy of the system. But the problem with DNNHMMis that it is very computation intensive. Now-a-dayssince we have advanced high performance computingsystem, we can use DNN-HMMs [12], [9].

- Few of them also suggested the use of DNN for ASRsystems by using deep denoising auto-encoder in audiofeature-extraction phase of the system. These extractedaudio features are then used to predict audio speech byusing GMM-HMMs [8], [11], [23].

From the above two approaches, the second approach is used the AVSR system. GMM-HMM is used instead of DNN-HMM because, it easily gets integrated with MSHMM for multimodal integration with CNN [26], [17].

A. Mechanisms for Visual Features Extraction

If we are able to extract visual features from the speaker's lip movement and use these extracted features in AVSRsystems, it will make system more robust and would provide more accurate prediction of spoken words where audio speech is corrupted by different kinds of noises in the environment [18].

The approaches to extract accurate visual features are as follows:

- Firstly derive the Active Shape Models (ASMs) and Active Appearance Models (AAMs) from actual mouth area images. Then, during training period, we have to provide appropriate labelled data to make accurate lipspace models. Its a very tedious task to make ASMs and AAMs. It becomes an overhead sometimes and we mayget diverted from our main purpose [15].
- Research also suggested to derive visual features from the image directly rather than constructing ASMs. Wecan achieve this by using Discrete Cosine Transforms (DCTs), Principal Component Analysis (PCA) and dimensionality reduction algorithm. These techniques derive low level features but problem with these approaches is that they are very vulnerable to rotation of the image and lighting conditions [21].
- We can use CNN to derive low-level features, as CNNs are not vulnerable to rotation and lighting conditions of the image [6], [14].

From the above proposed approaches, the third approach will be used in the AVSR system. The visual low-level features can be derived directly from the image itself and also this approach is resistant to any kind of variance in the image. These derived features can be directly fed into the GMM-HMM for further processing [18].

B. Mechanisms for integration of Audio and Visual systems

Multimodal speech recognition system increases the accuracyand robustness of the system. So our aim is to implementmultimodal integration over audio and video features. Thereexist a few approaches, out of which we have to choose themost efficient one, the one which can be seamlessly integratedwith previously chosen model for audio and video featuresextraction [26].

Following are the approaches for integration:

- Feature Fusion Approach: In this approach we merge theextracted features from different modalities and then it isconverted to multimodal features. These transformationsto multimodal features are done using DNNs and DeepBelief Networks(DBNs). But the problem in using deepnetworks in this case is that we have to explicitly selectfeatures according to the information gain we can getfrom each of its multimodal sources [17].

- Decision Fusion Approach: In this approach, the extractedfeatures from different modalities are merged. The differencehere is that we don't transform the extracted featuresto a multimodal feature vector, rather we directly inputthe output produced by the multimodal sources to theMSHMM. It does the prediction based on the informationgain of the multi-stream features [25], [16].

The second approach is used in the proposed AVSR architectureas it is simple and easy to use [18].

III. ARCHITECTURE OF THE MODEL

The AVSR system is composed of mainly two deeplearning architectures. They are Deep De-noisingAuto-Encoders and Convolutional Neural Networks. Figure2shows the architecture of the model. Deep De-noising Auto-Encoders are used for extracting Audio features and CNN isused for extraction of visual features. The deep de-noisingauto-encoder will be trained to predict actual speech audiofeatures from the corrupted ones so as to remove the effectof the noise in audio signal. The CNN is trained to predictvisual features of phoneme label from mouth area imagesspecifically based on the movement of the lips of respectivespeakers [18], [23], [14], [8].

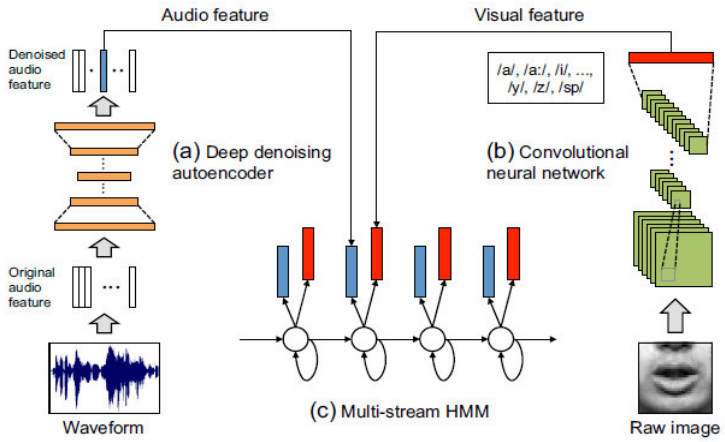

Fig. 2. Architecture of the AVSR system [18]

Also the GMM-HMMs are to be trained for recognitionof each of the individual audio and video stream using itsrespective extracted features. All the audio features should beconverted to MFCCs. Even for the GMM-HMMs used forvisual features are to be trained with MFCCs to label thephonemes from the raw mouth area images [15], [21].Actual prediction happens at the end, using MSHMMs withtwo separate streams, one with audio features from DeepDenoising Auto-Encoders and another one with visual features
from CNN [6], [26], [17].

IV. REQUIRED MACHINE LEARNING AND DEEPLEARNING CONCEPTS

In this section, various machine learning and deep learningconcepts which are used by the researchers in the domain ofAudio Visual Speech Recognition System are discussed.

A. Hidden Markov Models

It is a statistical Markov Model in which the system beingmodelled is assumed to be a Markov process with unobserved(hidden) states. Hidden units are states where the decision istaken based on the associated probabilities. Figure 3 showsthe block diagram for Hidden Harkov Model [15], [21].

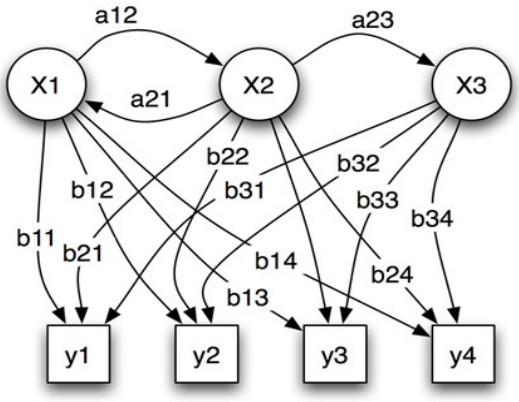

Fig. 3. Hidden Markov Model [5]

Suppose we have a set of hidden states $A_1,A_2,A_3,\ldots,A_n$and a set of visible states $O_1,O_2,O_3, \ldots O_m$ and process

moves from one state to another generating a sequence ofstates $A_{i1},A_{i2},A_{i3}, \ldots A_{ik}$. Markov chain rule states that theprobability of each subsequent state depends only on whatprevious state was,

$P(A_{ik}|A_{i1},A_{i2}, \ldots A_{ik-1}) = P(A_{ik}|A_{ik-1})$

To define a Hidden Markov Model, we have to first definethree associated probability matrices as follows,

- Transition Probabilities i.e

  $P_1 = P(A_i|A_j)$

- Observation probabilities

  $P2 = P(O_m|Ai)$

- Initial probabilities are probabilities of that state itselfsuch as

  $P_3 = P(S_i)$

$W = \text{argmax}P(O_i|_{Si})P(S_i)$

here, O_i is the acoustic vector , Si is the sequence of words.

B. Gaussian Mixture Models

Gaussian function is a data distribution function. It formsa Bell-Shaped curve as shown in figure 4. Data with similarcharacteristics will fall under this curve of Gaussian function.Gaussian Mixture Model is a probabilistic mixture model.It assumes that the underlying data will surely belong toa Gaussian function [26], [7].

Mathematical explanation ofGMM given by,

$P(x) = C_1P_1(x) + C_2P_2(x)\ldots + C_nP_n(x)$

Where, P(x), is Mixture Component.

$C_1,C_2,C_3\ldots C_n$are the Mixture Coefficients.

$P_i(x)$,are the Gaussian densities.

Gaussian mixture model uses more than one Gaussiandistribution function to describe the data distribution of thedata sets [17]. GMM represents data by multiple Gaussiandensities. GMM is used widely in speech recognition systembecause GMM has the capabilities to model arbitrary densities as shown in figure 4.

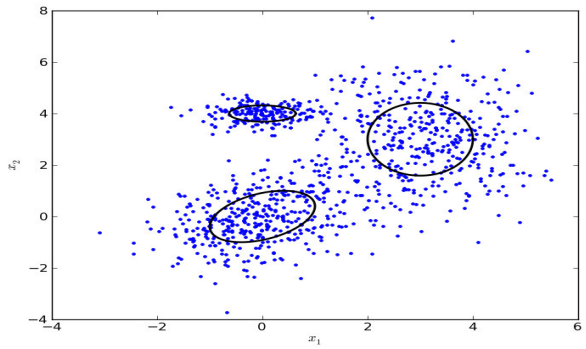

Fig. 4. Gaussian Mixture Model Distribution Plot [4]

C. Deep De-Noising Auto-Encoders

Auto-Encoders are basically a feed-forward neural networkwhose purpose is to learn the compressed distributed featuresof the data in the dataset at a conceptual level.Auto-Encoders have two portions in its structure, anencoding portion and a decoding portion. First the encodingportion is trained using back-propagation. So now eachlink between the neurons of the encoding portion of theauto-encoder receives some weight, such that the data isstudied at the conceptual level. A weight matrix of theseweights is to be created and then transpose this matrixtransposed. The weights of this transposed matrix is used by
the decoding portion of the auto-encoder[23].
Auto-encoder derives low-level features from the data-setbut it cannot be used directly. So to use this extracted low-level features, stacked auto-encoders are used to stackthem. For stacking the auto-encoders, one layer of the stackedauto encoder is to be trained at a time using back-propagationalgorithm [24].

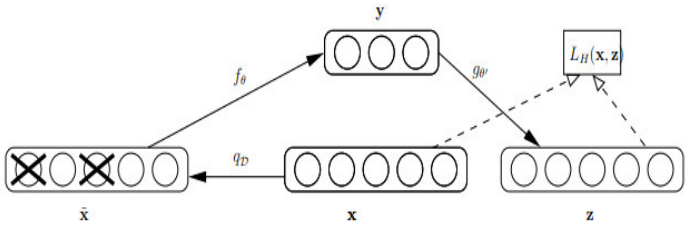

Fig. 5. Deep De-noising Auto-Encoder [2]

For using the Auto-Encoders in the speech recognition system,deep de-noising auto-encoder is used, as speech signalsare generally corrupted by noise. In deep de-noising autoencoders,the system is trained with the corrupted audio input.Figure 5 shows a model of deep de-noising auto-encoder [23],[24].

D. Convolutional Neural Networks

CNN is a feed-forward network that can extract visualproperties from an image. Convolutional neural networks aregenerally trained using back-propagation. CNN can recognizevisual features directly from the image. Patterns can be recognizedwith extreme variability [6], [14].

CNN has two main portions :-

- Convolution Layer
- Sub-Sampling Layer

Convolution Layer contains weight matrix for each ofthe associated filters. This weight matrix is then convolvedwith the image input and from the result of this process,respective feature map is made. Convolutional layer is alsovery computationally intensive.

In Sub-Sampling Layer, max-pooling is to be done onthe feature map made from convolutional layer. Hence, theextracted features can be reduced to a lower form of importantfeatures only. It will reduce the memory utilization [6], [14].

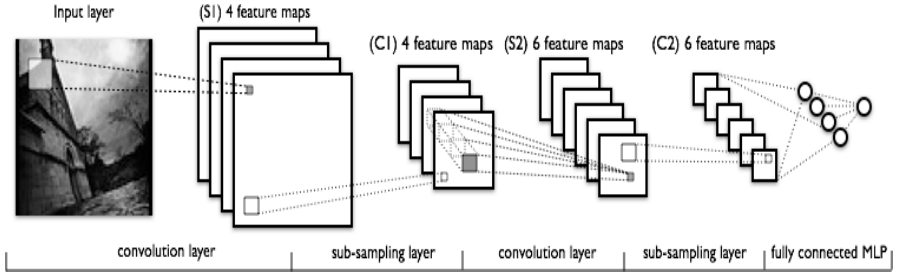

Fig. 6. Convolutional Neural Network [1]

Figure 6 shows the block diagram for Convolutional NeuralNetwork. According to the requirement of AVSR, pair ofconvolutional layer and sub-sampling layer is repeated formore robust feature extraction.Finally at the end a fully-connected multi-layer perceptronis attached to the CNN for classification purpose [6].

V. TOOLS AND LIBRARIES

In this section, the tools and libraries associated with theAVSR are discussed.

- NVIDIA CUDA 7.5 (System should have an NVIDIAGPU) [10]
- Toolkit for implementing our Automatic Audio SpeechRecognition: KALDI Speech Recognition Toolkit is usedfor the said implementation. For using KALDI, followinglibraries and tools have to be installed [20]:
  - OpenFst: Most of the compilation is done with it,and it is very heavily used.
  - IRSTLM : It is a language modeling Toolkit.
  - sph2pipe : It is for converting .sph files to .wav files.It is required for using LDC datasets.
  - sclite : It is not that important but still need may ariseas one of the dependencies, so its better to install it.
  - ATLAS :Its a linear algebra library. It will only workif the CPU throttling is not enabled.
  - CLAPACK : This also a linear algebra library. If onedoesn't have ATLAS, CLAPACK can be used as analternative.

- Toolkit for implementing our Automatic Visual SpeechRecognition:
  - OpenCV: It is an open source library for computervision.
  - Caffe : It is a deep learning framework dedicatedtowards image recognition [13].

VI. DATASETS

LIBRISPEECH dataset is used here. It is an ASR corpus based on public domain audio books. It consists of 1000 hours of speech sampled at 16000 kHz and all the speech samples are in .wav format [19].
For the full-fledged implementation, VidTIMIT dataset [22] can be used as it comprises of video and corresponding audio recordings of 43 people reciting short sentences. It is most suitable for audio visual speech recognition. The first two sentences for all persons are the same, with the remaining eight generally different for each person. Here even head rotation movement is performed in each session. So training on this system will give us a robust system.

VII. IMPLEMENTATION

The goal of the work is to implement an AVSR system capable of efficiently recognizing audio from the given multimedia file. Following are the steps involved in developing the desired AVSR:

- As mentioned in the previous sections, KALDI [20] is installed. Many dependencies arise apart fromthose mentioned above. Those mentioned are themajor ones.
- KALDI provides us with deep learning models andmachine learning model required for developingspeech recognition system.
- So as to develop our own system, some shell scripts are written to configure our AutomaticSpeech Recognition system, according to proposedarchitecture.

- After modeling our ASR portion of the architecture,the model is trained with LIBRISPEECH dataset.
- Then the trained model is tested against three audiosamples of the LIBRISPEECH dataset [19] whichwere sampled at 16000 kHz and was of .wav format.In this case, the accuracy was almost 100%.
- An English song of the same format was alsotested for. But the results were not accurate. Therecognizer was not able to differentiate accuratelybetween lyrics and the music components of thesong. Hence it was very inefficient.
- After analysing the results produced by the system,it seems that the system is getting over-fitted overthe training data. That is why, the accuracy is sogood over training data and poor on other test dataof that format.
- Currently, the said AVSR is partially implemented.The results of this partial implementation are shown infigures 7 and 8. Figure 7 shows the outcomewhen File 1 (contains the audio file bearing story) is givenas the testing parameter. The recognizer is able to convertthe audio into text.Figure 8 shows the output when the English song isbeing passed as the input parameter.

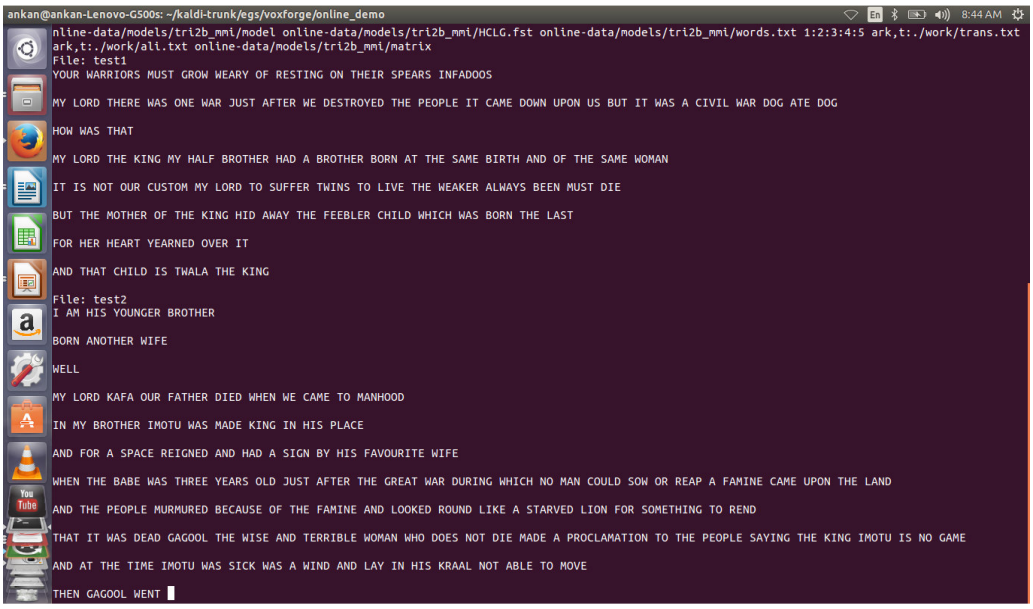

Fig. 7 Implementation Screenshot 1

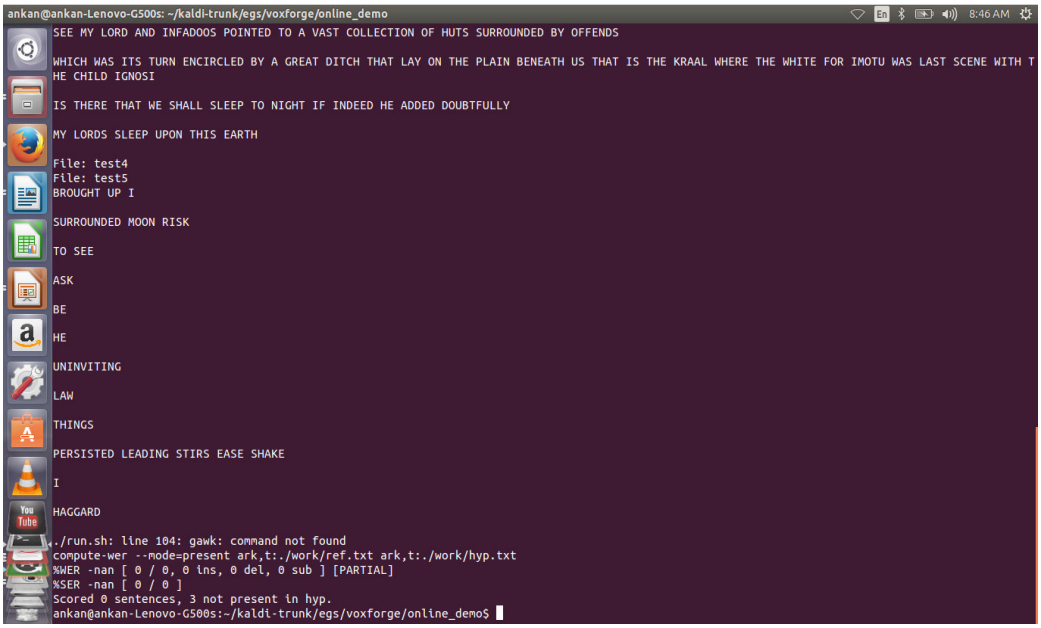

Fig.8 Implementation Screenshot 2

- The work is under progress and currently the focus is onimplementing the full fledged AVSR system for accurateaudio video recognition.

VIII. CONCLUSIONS

Video files from VIDTIMIT dataset [22] will be used to extract the video features and convert into text file. To further, improve the efficiency, related audio files will be converted into text file. Thereby, the video converted into text file and audio converted into text file will be integrated together to improve efficiency of recognition. The future work of this work is as under:

- The problem of over-fitting which at present is associated with this ASR system will be resolved.
- Generating audio and video text Automatic Visual Speech Recognition System using Caffe[13](deep learning framework) and OpenCV (libraries associated to image processing).
- VidTimit [22] dataset will be used in implementation.
- The final integration of the two portion will be done using MSHMM[17], [7].

REFERENCES

[1] Convolutional neural network, http://deeplearning.net/tutorial/lenet.html, 2016, (Date last accessed 22-January-2016).
[2] Deep autoencoder, http://eric-yuan.me/dae/ , 2016, (Date last accessed 22-January-2016).
[3] Gaussian mixture model, http://www.maths.adelaide.edu.au/matthew.roughan/code.html, 2016, (Date last accessed 22-January-2016).
[4] Gaussian mixture model, http://pypr.sourceforge.net/mog.html, 2016, (Date last accessed 22-January-2016).
[5] Hidden markov model, https://en.wikipedia.org/wiki/HiddenMarkovmodel, 2016, (Date last accessed 22-January-2016).
[6] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, Convolutional neural networks for speech recognition, Audio, Speech, and Language Processing, IEEE/ACM Transactions on 22 (2014), no. 10, 1533–1545.
[7] Herv´eBourlard, St´ephaneDupont, and Christophe RisMartignyValais Suisse, Multi stream speech recognition, (1996).

[8] Adam Coates, Brody Huval, Tao Wang, David Wu, Bryan Catanzaro, and Ng Andrew, Deep learning with cots hpc systems, Proceedings of the 30th international conference on machine learning, 2013, pp. 1337–1345.
[9] George E Dahl, Dong Yu, Li Deng, and Alex Acero, Contextdependentpre-trained deep neural networks for large-vocabulary speech recognition, Audio, Speech, and Language Processing, IEEE Transactions on 20 (2012), no. 1, 30–42.
[10] ETSI/SAGE, Specification of the 3GPP Confidentiality and Integrity Algorithms 128-EEA3 & 128-EIA3. Document 1: 128- EEA3 and 128-EIA3 Specification.
[11] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, AbdelrahmanMohamed, NavdeepJaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, Signal Processing Magazine, IEEE 29 (2012), no. 6, 82–97.
[12] Po-Sen Huang, Kush Kumar, Chaojun Liu, Yifan Gong, and Li Deng, Predicting speech recognition confidence using deep learning with word identity and score features, Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE, 2013, pp. 7413–7417.
[13] YangqingJia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093 (2014).
[14] YuxuanLan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong, and Richard Bowden, Improving visual features for lip-reading., AVSP, 2010, pp. 7–3.
[15] JuergenLuettin, Neil Thacker, Steve W Beet, et al., Visual speech recognition using active shape models and hidden markov models, Acoustics, Speech, and Signal Processing, 1996. ICASSP-96.Conference Proceedings., 1996 IEEE International Conference on, vol. 2, IEEE, 1996, pp. 817–820.
[16] N Morgan and H Bourlard, Connectionist speech recognition: a hybrid approach, 1994.
[17] JiquanNgiam, AdityaKhosla,Mingyu Kim, Juhan Nam, HonglakLee, and Andrew Y Ng, Multimodal deep learning, Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 689–696.
[18] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata, Audio-visual speech recognition using deep learning, Applied Intelligence 42 (2015), no. 4, 722–737.
[19] VassilPanayotov, GuoguoChen, Daniel Povey, and SanjeevKhudanpur, Librispeech: an asr corpus based on public domain audio books, Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.
[20] Daniel Povey, ArnabGhoshal, Gilles Boulianne, Luk´aˇsBurget, Ondˇrej Glembek, NagendraGoel, MirkoHannemann, PetrMotl´ıˇcek, YanminQian, Petr Schwarz, et al., The kaldi speech recognition toolkit, (2011).
[21] Steve Renals, Nelson Morgan, Herv´eBourlard, Michael Cohen, and Horacio Franco, Connectionist probability estimators in hmm speech recognition, Speech and Audio Processing, IEEE Transactions on 2 (1994), no. 1, 161–174.
[22] Conrad Sanderson, The vidtimit database, Tech. report, IDIAP, 2002.
[23] Pascal Vincent, Hugo Larochelle, YoshuaBengio, and Pierre- Antoine Manzagol, Extracting and composing robust features with denoisingautoencoders, Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 1096–1103.
[24] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, YoshuaBengio, and Pierre-Antoine Manzagol, Stacked denoisingautoencoders: Learning useful representations in a deep network with a local denoising criterion, The Journal of Machine Learning Research 11 (2010), 3371–3408.
[25] Takami Yoshida, Kazuhiro Nakadai, and Hiroshi G Okuno, Automatic speech recognition improved by two-layered audio visual integration for robot audition, Humanoid Robots, 2009. Humanoids 2009. 9th IEEE-RAS International Conference on, IEEE, 2009, pp. 604–609.
[26] ZhihongZeng, Yuxiao Hu, Ming Liu, Yun Fu, and Thomas S Huang, Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition, Proceedingsof the 14th annual ACM international conference on Multimedia, ACM, 2006, pp. 65–68