1. **Paper name -Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features**

In this paper, we employed three different evaluated feature sets for representing the spoken information found in the video frames recorded from the speakers, inside a speaker independent lip-reading mission. Two HMM-based visual models, including the conventional GMM-HMM and the young DNN-HMM hybrid were implemented to test and compare the introduced features. After testing the easy-to-access raw gray level ROI of the speakers’ lips, the geometric specifications of the lips (the shape features) were employed which showed a lower error rate by 20.4% relative, on average. The DBNFs which benefit from the advantages of both the former feature sets were then employed and showed a relative improvement with an average of 15.4% in comparison to the shape features, over the test data.

1. **Paper Name - Lip Reading in the Wild**

They are show that CNN architectures can be used to classify temporal sequences with excellent results. On the 333-word test set, we achieve top-1 accuracy of 65.4%, which exceeds state-of-the-art on multiple datasets that have lexicon sizes that are orders of magnitude smaller, and a top-10 accuracy of 92.3%. We also demonstrate a recognition performance that exceeds the state of the art on a standard public benchmark dataset, Oulu VS. Next steps include extending to lip reading of proﬁle views, and combining the CNNs pre-trained using this approach with LSTMs trained with a language model in order to recognize sentences rather than individual words. Of course, the visual only speech recognition method developed here can also be combined with audio only speech recognition to both their beneﬁts.

1. **Paper Name - Thai Lip-Reading CAI for Hearing Impairment Student**

We developed the CAI application that is used as a lip reading learning tool for hearing impaired student. Structure of this CAI consist of learning lesson unit and game unit. For The learning lesson unit, student will learn to understand mouth movement of each vocabulary that use in daily life. They can practice and compare their mouth shape with example mouth model. For the game unit, student plays the multiple-choice game to select the correct answer of the question. We conducted an experiment to test if the CAI can help and serve as a learning platform. The result is measured from effectiveness. The test for effectiveness showed positive results. From the pre-test and post-test process, all student scored are more after learning the CAI. Experiment data showed that the number of correct answer increased. In addition, the feedback after CAI using, the students attend to learn and would like to play the game. Therefore, these results indicated that our CAI could help students to learn in lip-reading.

1. **Paper name - Connected Bangla Speech Recognition using Artificial Neural Network**

In this work, an automatic recognizer of connected Bangla digits has been developed using BPNN and MFCC feature extraction method. Neural networks are very sensitive classifiers. A small amount of changes in the network architecture may cause significant change in the output. One of the major goals of this research experiment was to optimize the network based on network parameters such as number of hidden layers, learning rate, error threshold and number of epochs. It is also evident from the result is that the recognition accuracy varies over digits due to their phonetic traits. To acquire persistent performance over digits, BPNN based recognition system can be improved either by employing hybrid classifier and/or incorporating robust features. BPNN learning algorithm is used to train the network. The required time to train the network, number of hidden layers, error threshold and number of epochs are considered while training the network to reach the best possible recognition accuracy. This proposed system has been implemented using object oriented programming and the achieved recognition accuracy is very much satisfactory and consistent. The network has been tested for three different setups and the best recognition accuracy achieved for digit dataset is 98.46%.

1. **Paper name - Instant Bangla Speech to Text Conversion**

In this paper they are work at first task for speech to text Bangla conversion is connecting microphone perfectly to provide voice command. User speaks through the microphone. The analog sound wave coming from the microphone is changed into digital sound signal using sound card of the computer consequently. Secondly, to convert the sound signal into text phonemes, the user should have a speech engine, Microsoft SDK 5.1 likewise. The speech engine is for converting digital sound signal to text format. This text forms are written into text editor using keyboard handler function. Grammar rules are used by speech recognition (SR) process to analyze human speech input. In this process, this attempts to understand what a person is saying. In this research, the grammars are represented in a XML file in text form for the upcoming voice command. In the case of dictation, the grammars are used to indicate some words that are likely to be spoken.

1. **Paper name - A BENGALI HMM BASED SPEECH SYNTHESIS SYSTEM**

The paper presents the capability of an HMM-based TTS system to produce Bengali speech. In this synthesis method, trajectories of speech parameters are generated from the trained Hidden Markov Models. A final speech waveform is synthesized from those speech parameters. In our experiments, spectral properties were represented by Mel Cestrum Coefficients. Both the training and synthesis issues are investigated in this paper using annotated Bengali speech database. Experimental evaluation depicts that the developed text-to-speech system is capable of producing adequately natural speech in terms of intelligibility and intonation for Bengali. The speech database employed here for training purpose is originally developed by C-DAC. The output of the Bengali-HTS is also compare with previously developed Epoch Synchronous Non Overlap Add (ESNOLA) based concatenated speech synthesis technique. The total average score for the original sentences is 4.66 and the ESNOLA based synthesis sentence is 2.34HTS is 3.6.

1. **Paper name - Lip-reading by neural networks: Visual preprocessing, learning and sensory integration**

In this paper they formed a hybrid speech reading system consisting of two time delay neural networks (video and acoustic) and integrated their responses by means of independent opinion pooling - the Bayesian optimal method given conditional independence, which seems to hold for our data. The video pre-processing presented here represents a first pass at reducing the amount of visual data to a manageable level in order to enable on-line processing. The results indicate that even these straightforward, computationally tractable methods can significantly enhance speech recognition. This hybrid system had an error rate 25% lower than that of the acoustic subsystem alone on a five-utterance speaker-independent task, indicating that video can be used to improve speech recognition.

1. **Paper name - Lip Reading Sentences in the Wild**

The goal of this work is to recognize phrases and sentences being spoken by a talking face, with or without the audio. They used (1) a ‘Watch, Listen, Attend and Spell’ (WLAS) network that learns to transcribe videos of mouth motion to characters; (2) a curriculum learning strategy to accelerate training and to reduce over fitting. The WLAS model trained on the LRS dataset surpasses the performance of all previous work on standard lip reading benchmark datasets, often by a significant margin. This lip-reading performance beats a professional lip reader on videos from BBC television, and they also demonstrate that if audio is available, then visual information helps to improve speech recognition performance. They got Word error rates 23.8% and 3.0% respectively for LRW dataset and GRID dataset.

1. **Paper name - Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory**

In this paper, the convolutional neural network (CNN) used to image feature extraction is combined with a recurrent neural network (RNN) based on attention mechanism for automatic lip-reading recognition. They proposed method for automatic lip-reading recognition can be divided into three steps. Firstly, they extract key frames from their own established independent database (English pronunciation of numbers from zero to nine by three males and three females). Then, they use the Visual Geometry Group (VGG) network to extract the lip image features. It is found that the image feature extraction results are fault-tolerant and effective. Finally, they compare two lip-reading models: (1) a fusion model with an attention mechanism and (2) a fusion model of two networks. The accuracy of the proposed model is 88.2% in the test dataset and 84.9% for the contrastive model. Which is 3.3% higher than the general CNN-RNN.

1. **Paper name - Deep Learning for Lip Reading using Audio-Visual Information for Urdu Language**

In this paper, they have tried to train two different deep-learning models for lip-reading: first one for video sequences using spatiotemporal convolution neural network, Bi-gated recurrent neural network and Connectionist Temporal Classification Loss, and second for audio that inputs the MFCC features to a layer of LSTM cells and output the sequence. Apart from implementation and investigation of models, they contributed a small Urdu language corpus for lip-reading. Corpus is consisting of 10 words and 10 phrases, each spoken by 10 users 10 times, in total they recorded and pre-processed the 1000 videos of dataset. In terms of this paper LSTM based network performs better than DNN. For word, accuracy of LSTM based model and DNN model were respectively 62% and 56%. For digits, accuracy of LSTM based model and DNN model were respectively 72% and 64%.