1. **Paper Name - END-TO-END VISUAL SPEECH RECOGNITION WITH LSTMS.**

In this paper they are present an end-to-end visual speech recognition system which jointly learns to extract features directly from the pixels and perform classiﬁcation using LSTM networks. Results on subject independent experiments demonstrate that the proposed model achieves state-of-the-art performance on the OuluVS2 and CUAVE databases when compared with models which use a similar visual front end. The best provided baseline result, 74.8%, is achieved by HMMs in combination with DCT features .The best overall result is achieved by the end-to-end 2-stream model, with a classification accuracy of 84.5%.

1. **Paper Name - Name - End-to-End Multi-View Lip reading**

In this work, they are present an end-to-end multi view lip reading system which jointly learns to extract features directly from the pixels and performs classification using BLSTM networks. The proposed model achieves state-of-the-art performance on the OuluVS2 without using external data for training or even data augmentation. The provided mouth ROIs are well cropped and this might not be the case when automatic tools for mouth ROI detection are used. The model can be easily extended to multiple streams so we are planning to add an audio stream in order to evaluate its performance on audiovisual multi-view speech recognition. The best three-view model results in a 10.5% absolute improvement over the current multi-view state-of-the-art performance on OuluVS2, without using external databases for training, achieving a maximum classification accuracy of 96.9%

1. **Paper Name - SPEAKER IDENTIFICATION BY LIPREADING**

A novel approach fix speaker identification has been described. Based ca spatial and temporal analysis of the mouth. Facial features are extracted from image sequences which represent the shape and intensity of the lips. The features are of low dimension and invariant to scale, translation and rotation. Considering the small training and test duration. Results are encouraging and demonstrate that lip intonation is an important cue for person idmt5catim. Further experiments are necessary to evaluate the performance of the method for a large number of subjects and to investigate the benefit of combining it with other approaches like speaker recognition and fact recognition.

1. **Paper name - A real-time lip localization and tacking for lip reading**

In this paper, we describe a new robust approach to improve lip localization and tracking. The first part of our proposed algorithm is lip location based on OpenCV. From experimental results, our proposed method can successfully detect the lip region. Results from the lip location accuracy allowed more accurate lip region segmentation. In the second part, a new method called *a* component of Lab color space is proposed to accurately extract lip shape and track lip region. Overall, our implemented approach has shown high reliability and is able to perform robustly under various conditions.

1. **Paper’s Name - Lip localization technique towards an automatic lip reading approach for Myanmar consonants recognition**

This paper proposed the technique to localize lip region for the Myanmar consonants recognition. The experimental system demonstrates that this technique performs lip motion sequences in video. In our experiment, we can localize all of the test lip movement successfully and the results were perceived to be acceptable for lip reading. For future work, we will intend to explore more observable features and recognition phase by applying Support Vector Machine classifier (SVM) method for remaining three syllable and four syllable Myanmar consonants recognition based on lip movements to produce the good recognition result. We hope that this study will help to a new teaching and learning method for Myanmar language education.

1. **Paper’s name - LCANet: End-to-End Lip reading with Cascaded Attention-CTC**

In this paper, we have presented LCANet, an end-to-end deep neural network architecture for machine lip-reading. To accurately predict visemes, LCANet introduced a cascaded attention-CTC decoder which effectively compensates the defect of the conditional independence assumption of the CTC approach. In addition, LCANet stacks highway net- work layers over 3D-CNN layers to further improve per-formance. Extensive experiment results show that LCANet achieves the state-of-the-art accuracy as well as faster con- vengeance. The experimental results show the proposed system achieves a 1.3% CER and 3.0% WER on the GRID corpus database, leading to a 12.3% improvement compared to the state-of-the-art models. The proposed cascaded attention-CTC model is better. Accuracy of 97.4% for AH-CTC (LCANet).

1. **Paper name -Real Time Bengali Speech to Text Conversion using CMU Sphinx**

In this paper they are presented a model for voice to text conversion method for Bengali language using various techniques to refine the data and adapt it for the complexities of implementing Bengali characters. An open source frame work called CMU Sphinx 4 was used to generate Bengali UNICODE font. A digital audio workstation called Audacity was used to manipulate the recorded data. The performance of the proposed model was tested using a dataset where both male and female voices were recorded. The proposed model showed around 75% accuracy for the tested dataset. The best work done in this arena was using Microsoft SAPI which obtained a 78% accuracy But that detection was only on a word by word basis which was a major limitation and the accuracy was for a specific data. This system not only recognizes the word but also the sentence due to our better training model as well as the newer Sphinx 4 framework.

1. **Paper name- Isolated and Continuous Bangla Speech Recognition: Implementation, Performance and application perspective**

In this paper they are concentrated on the research and development of a Bangla Speech Recognizer using the appropriate technique and tools. This work is the first reported attempt to recognized Bangla speech using HMM Technique with the assist of stochastic language model. They are create a regular grammar and convert it to an intermediate form of decoding network using the HParse tool. Networks are specified using the HTK Standard Lattice Format (SLF). In the grammar the legal word sequences explicitly for translation, automotive speech recognition, dictation, hands-free computing: voice command recognition computer user interface, home automation, interactive voice response, medical transcription, mobile telephony, pronunciation evaluation in computer-aided language learning applications and robotics. The isolated speech recognition for commands & control, data entry, mobile telephony and home automation task. On the other hand continuous speech recognition can be used for speech to text conversion. Recognizing continuous speech with ANN classifier has average accuracy rate of 73.36% for three layer Back Propagation Neural Network the maximum accuracy rate is 86.67% and spoken letter recognition by measuring Euclidian distance, which can recognize only the vowels, has an 80% accuracy rate In comparison, the recognizer presented in this paper has an average accuracy rate of 85%.

1. **Paper name - Implementation of Speech Recognition System for Bangla**

In this paper they are use Building Acoustic model, Utterance, Speaker Dependence, Vocabularies, HMM model and language model .Speech recognition (SR) in terms of machinery is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. It is a broad term which means it can recognize almost anybody's speech, but to make the machine independent of voice, huge training data is required. There are basically two types of SR: 1. Isolated speech recognition - ISR 2. Continuous speech recognition - CSR .They tried to test the system by varying speaker, environment, microphone etc. Testing the system is done using audio inputs of two test speaker and live test from the microphone in different environment (speaker profiles are given in Appendix section). Tests are done using two different decoder: 1. Pocket Sphinx 2. Sphinx 4. Sampling rate of the audio: 16 kHz • Bit rate (bits per sample) : 16.Then after the process is completed we will be getting an output like: MODULE: DECODE Decoding using models previously trained Decoding 529 segments starting at 0 (part 1 of 1) 0% This step had 2 ERROR messages and 45 WARNING messages. Please check the log file for details. Aligning results to find error rate SENTENCE ERROR: 2.8% (15/529) WORD ERROR RATE: 1.0% (23/2319).

Test results with Pocket Sphinx

TOTAL Words: 2292 Correct: 2276 Errors: 16 TOTAL Percent correct = 99.30% Error = 0.70% Accuracy = 99.30%

TOTAL Words: 142 Correct: 134 Errors: 9 TOTAL Percent correct = 94.37% Error = 6.34% Accuracy = 93.66%

TOTAL Words: 392 Correct: 381 Errors: 12 TOTAL Percent correct = 97.19% Error = 3.06% Accuracy = 96.94%

Average Accuracy Rate = 90.65%

Input Type: Microphone-

TOTAL Words: 48 Correct: 43 Errors: 5 TOTAL Percent correct = 89.58% Error = 10.42% Accuracy = 89.58%

TOTAL Words: 128 Correct: 118 Errors: 11 TOTAL Percent correct = 92.19% Error = 8.59% Accuracy = 91.41%

The described above system is in its preliminary level, all the tools, data used to train and build the system has been discussed throughout the whole report. To make the system more natural, lots of improvement in case of data, tools and its parameters is required, but again no speech recognizer till now has 100% accuracy.

**10. Paper name - LIP-READING VIA DEEP NEURAL NETWORKS USING HYBRID VISUAL FEATURES**

This study has discussed some issues of the automatic lip-reading systems and explored an overcoming approach with two major parts of visual feature extraction and classification. In the feature extraction part, hybrid features were extracted in a proposed well-designed arrangement of certain function blocks. As the final classifier, a single DBN with a well-designed topology, inside a DBN-HMM structure, was established and employed for all speakers. Experiments were applied on the CUAVE database, in two separate tasks of the MS and SI conditions. The proposed method was basically evaluated on the phoneme-level and considerably outperformed the HMM baseline recognizer. Here, application of a properly structured Deep Belief Network (DBN)-based recognizer is highlighted. Multi-speaker (MS) and speaker-independent (SI) tasks are performed over CUAVE database, and phone recognition rates (PRRs) of 77.65% and 73.40% are achieved, respectively. The best word recognition rates (WRRs) achieved in the tasks of MS and SI are 80.25% and 76.91%, respectively. Resulted accuracies demonstrate that the proposed method outperforms the conventional Hidden Markov Model (HMM) and competes well with the state-of-the-art visual speech recognition works.