**Paper-1: Lipnet: end-to-end sentence-level lipreading**

In this paper, the authors present LipNet, a model that maps a variable-length sequence of video frames to text, making use of spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. For dataset, GRID corpus has been used which has audio and video recordings of 34 speakers who produced 1000 sentences each, for a total of 28 hours across 34000 sentences. In the LipNet architecture, a sequence of T frames is used as input, and is processed by 3 layers of STCNN, each followed by a spatial max-pooling layer. The features extracted are processed by 2 Bi-GRUs; each time-step of the GRU output is processed by a linear layer and a softmax. This end-to-end model is trained with CTC. LipNet achieves 95:2% accuracy in sentence-level, overlapped speaker split task, outperforming experienced human lipreaders and the previous 86:4% word-level state-of-the-art accuracy.

**Paper-2: Lipreading using convolutional neural network**

In this paper, they applied CNN as a visual feature extraction mechanism for VSR. They have trained a CNN with images of a speaker’s mouth area in combination with phoneme labels, the CNN acquires multiple convolutional filters, used to extract visual features essential for recognizing phonemes. To recognize words from the phoneme label sequences generated by the CNN, monophone HMMs with 8, 16, and 32GMM components are utilized. Evaluation is conducted with the 84 test words from the same speaker, yielding a closed-speaker and open-vocabulary evaluation. To compare with the baseline performance, two other visual features with similar dimensionalities are prepared. One feature has 36 dimensions, generated by simply rescaling the images to 6\*6 pixels, and the other feature has 40 dimensions, generated by compressing the raw images by PCA. Their proposed system is evaluated on an audio-visual speech dataset which is comprised of 300 Japanese words with six different speakers. The evaluation results of their isolated word recognition experiment demonstrate that the visual features acquired by the CNN significantly outperform those acquired by conventional dimensionality compression approaches, including principal component analysis. The average phoneme recognition performance is 58%. They reported that visual phoneme recognition works better for recognizing vowels than consonants. The result derives from the fact that, the mean recognition rate for all vowels is 60-100%, whereas the mean recognition rate for all other phonemes is 20-80%.

**Paper-3: Lipreading with long short-term memory**

In this paper, feedforward and recurrent neural network layers (LSTM) are stacked to form a single structure which is trained by back-propagating error gradients through all the layers. The performance of such a stacked network was experimentally evaluated and compared to a SVM classifier using conventional computer vision features (Eigenlips and Histograms of Oriented Gradients). The LSTM lipreader with a single feed-forward network learns the features automatically together with training the LSTM sequence classifier, consistently achieved almost 80% word accuracy in speaker-dependent lip-reading. They have used GRID corpus dataset which have 51 different words to classify. From their experiment they have recognized that the confusion on letters is far higher than on longer words. For this speaker and configuration the accuracy on the letters is 69.8, the accuracy on the non-letter words is 93.4%. The total accuracy is 82.0%.They reported a best word accuracy on held-out evaluation speakers of 79.6% using the end-to end neural network-based solution (11.6% improvement over the best feature-based solution evaluated).

**Paper-4: Deep inside convolutional networks: visualizing image classification models and saliency maps**

This paper is concerned about the visualisation of image classification models, learnt using deep ConvNets. They have considered two visualization techniques, based on computing the gradient of the class score with respect to the input image. The first generates an artificial image, which is representative of a class of interest. The second computes an image-specific class saliency map, highlighting the areas of the given image, discriminative with respect to the given class. They showed that such saliency map can be used to initialise Graph Cut based object segmentation without the need to train dedicated segmentation or detection models. Finally, they demonstrated that gradient-based visualisation techniques generalise the DeconvNet reconstruction procedure. Their visualisation experiments were carried out using a single deep ConvNet, trained on the ILSVRC-2013 dataset, which includes 1.2M training images, labelled into 1000 classes. Their weight layer configuration is: conv64-conv256-conv256-conv256-conv256-full4096-full4096-full1000, where convN denotes a convolutional layer with N filters, fullM – a fully-connected layer with M outputs. On ILSVRC-2013 validation set, the network achieves the top-1/top-5 classification error of 39:7%=17:7%, which is slightly better than 40:7%/18:2%, reported in for a single ConvNet.

**Paper-5: Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss**

This paper is concerned about an audio-visual speech recognition system for a person with an articulation disorder resulting from severe hearing loss. They proposed a novel visual feature extraction approach that connects the lip image to audio features efficiently, and the use of CBN’s increases robustness with respect to speech fluctuations caused by hearing loss. The effectiveness of this approach was confirmed through word-recognition experiments in noisy environments, where the CBN-based feature extraction method outperformed the conventional methods. They have used utterances of one male person with hearing loss, where the text is the same as the ATR Japanese speech database A-set where 2,620 words are used as training data, and 216 words as test data. First, they prepared the input features for training a CBN from lip images and speech signals uttered by a person with hearing loss. For the audio signals, after calculating short-term mel spectra from the signal, they obtained mel-maps by merging the mel spectra into a 2D feature with several frames, allowing overlaps. The visual signals of the eyes, mouth, nose, eyebrows, and outline of the face are aligned using the point distribution model (PDM) and its model parameter is estimated by constrained local model (CLM) and a lip image is extracted. For the output units of the CBN, they used phoneme labels that correspond to the input mel-map and lip images. Audio and visual CBNs are separately trained, and the parameters of the CBN are trained by back-propagation with stochastic gradient descent, starting from random values. Following the training of CBNs, the input mel-map and lip images are converted to the bottleneck feature by using each CBN. Then these features are concatenated, and used in the training of HMMs for speech recognition. In the test stage, they extracted features using each CBN, which tries to produce the appropriate phoneme labels in the output layer. Finally, extracted bottleneck audio and visual features are simply concatenated and used as the input features of HMMs to audio-visual speech recognition. Their proposed audio-visual feature outperforms the AV BNF I in the clean environment and SNR of 20dB, where the integrated features between the audio and the proposed visual bottleneck features improved 3.3% and 3.8% compared with the AV BNFs I, respectively. However, at the SNRs of 10dB and 5dB, the integrated feature using their proposed feature could not improve the accuracy in comparison with that of the AV BNF I.

**Paper-6: Lip reading using CNN and LSTM**

In this paper, the author presented various methods to predict words and phrases from only video without any audio signal. They employed a VGGNet pre-trained on human faces of celebrities from IMDB and Google Images. The VGGNet is trained on images concatenated from multiple frames in each sequence, as well as used in conjunction with LSTMs for extracting temporal information. While the LSTM models fail to outperform other methods for a variety of reasons, the concatenated image model that uses nearest-neighbor interpolation performed well. They have used the MIRACL-VC1 dataset in their project. The dataset was created from 15 people who spoke each of ten words and ten phrases ten times leading to a total of 15 \* 20 \* 10 = 3000 instances. They proposed several new methods for performing visual speech recognition on sequences of color images with variable length. The initial methods concatenated the first k images of each sequence into a 2D grid, which was then classified by a VGGNet pre-trained on faces. One method attempted to train a smaller model on these concatenated images from scratch. The final model attempted to handle variable-length sequences with multiple LSTM layers which were given the feature vectors output from the VGGNet as input. Their best-performing model was the concatenated model that used interpolation. The model we trained from scratch did not perform well because the dataset we used is relatively small. Therefore the pre-trained VGGNet was still able to perform well on these concatenated images. They also experimented with different parameter update strategies. They noticed that SGD was incapable of training the model in reasonable time. It gave no significant improvements even after 20 epochs. In comparison, Adam showed improvements right from the first epoch. They have achieved best validation accuracy of 76%, and the test accuracy of their best model is 44:5%. They noticed that phrases have a higher accuracy as compared to words.

**Paper-7: Comparison of human and machine-based lip-reading**

In this paper, the authors have contrasted the performance of a machine-based lip-reading system with human lip-reading ability. They found that the automated system outperforms human lip-readers. For relatively simple tasks there is little improvement in recognition accuracy when adding full appearance features to the machine-based system, whereas for human lip-readers they observed significant improvements in performance. Finally, they measured the effect of ‘speaker training’ on human lip-reading ability and they found even very limited training is sufficient to improve performance. They also have looked at the contribution of shape and appearance to lip-reading ability and found that for the (6-class) task presented here, appearance is significant for human viewers, but not for the automated system. Generally we find that automatic lip-reading systems outperform human viewers. However, an exception is recognized in full, isolated, real words. To train machine-based lip-reading systems, word-level HMMs were trained from both shape-only and full shape and appearance features using a leave-one-out cross-validation framework. The topology of the HMMs was optimized and the best performing topology used in the evaluation. This was a topology of 16 states each with 3 Gaussian mixture components. To measure the baseline performance of human lip-readers, 17 computing undergraduate students from a UK University volunteered to take part in the experiment. This provided a set of 60 utterances (six letters \* five speakers \* two stimuli type). In this experiment, participants were shown, in a randomized order, three repetitions of the 60 test movies (without audio), and they were each asked to circle on an answer sheet the letter they believed was being spoken. This was a closed test where all letters A–F letters were offered as possible answers to all utterances. At both the phoneme and viseme levels, the automated system performed significantly better than the human participants (p < 0:002) and achieved recognition rates of 80.27% and 91.6% (full shape and appearance) compared to 31.6% and 35.4% respectively.

**Paper-8: Improved speaker independent lip reading using speaker adaptive training and deep neural networks**

In this paper, it is shown that the application of SAT appears to have considerable promise in speaker-independent lipreading. For data they have used an audiovisual corpus of twelve speakers, seven male and five female, each reciting 200 sentences selected from the Resource Management Corpus. Kaldi speech recognition toolkit was used to train the visual speech models (phonemes and visemes units) and decode the test data. The HMM/GMM systems that they built are: (i) monophone and monoviseme systems with and features (Mono), (ii) triphone and triviseme systems with LDA ((Tri:LDA)) (iii) triphone and triviseme systems with LDA+MLLT ((Tri:LDA+MLLT)), (iv) triphone and triviseme systems with LDA+MLLT+SAT (Tri:LDA+MLLT+SAT).In the experiment, firstly the visual features are considered in a block of 7 frames. They are then decorrelated and forced to a dimensionality of 40 using Linear Discriminant Analysis and further decorrelated using maximum likelihood linear transform. SAT is then applied using feature-space maximum likelihood linear regression of 40 \* 41. The 40-dimensional speaker adapted features are then spliced across a window of 9 frames and applying LDA to decorrelate the concatenated features and reduce dimensionality to 250. They found that word recognition accuracy is always higher when phonemes are used as the modelling units rather than visemes.

**Paper-9: Listening with your eyes: towards a practical visual speech recognition system using deep boltzmann machines**

This paper presents a novel feature learning method for visual speech recognition using Deep Boltzmann Machines (DBM). The experimented method is able to explore both acoustic information and visual information to learn a better visual feature representation. During the test stage, only the videos are used to generate the missing audio features, and both the given visual and audio features are used to produce a joint representation. The experimental results show that the proposed techniques outperform the performance of handcrafted features and previously learned features. The data corpus used in this paper was collected through an Australia wide research project called AusTalk. It is a large-scale audio-visual database of spoken Australian English, including isolated words, digit sequences, and sentences, recorded at 15 different locations in all states and territories of Australia. In the experiment, the visual feature learned by the Deep Boltzmann Machine (DBM) is concatenated with Discrete Cosine Transform (DCT) feature vector, followed by a Linear Discriminant Analysis (LDA) to decorrelate the feature and reduce the feature dimension. Then, the Gaussian Mixture Model- Hidden Markov Model (GMM-HMM) is used as a classifier for visual speech recognition. Their proposed method showed the accuracy of 69.1%.

**Paper-10: Towards End-to-End Speech Recognition with Recurrent Neural Networks**

This paper has demonstrated that character-level speech transcription can be performed by a recurrent neural network with minimal preprocessing and no explicit phonetic representation. They have also introduced a novel objective function that allows the network to be directly optimized for word error rate, and shown how to integrate the network outputs with a language model during decoding. Finally, by combining the new model with a baseline, they have achieved state-of-the-art accuracy on the Wall Street Journal corpus for speaker independent recognition. The experiments were carried out on the Wall Street Journal corpus (available as LDC corpus LDC93S6B and LDC94S13B). The RNN was trained on both the 14 hour subset ‘train-si84’ and the full 81 hour set, with the ‘test-dev93’ development set used for validation. For both training sets, the RNN was trained with CTC. The network had five levels of bidirectional LSTM hidden layers, with 500 cells in each layer, giving a total of 26.5M weights. It was trained using stochastic gradient descent with one weight update per utterance, a learning rate of and a momentum of 0.9. The system achieves a word error rate of 27.3% on the Wall Street Journal corpus with no prior linguistic information, 21.9% with only a lexicon of allowed words, and 8.2% with a trigram language model. Combining the network with a baseline system further reduces the error rate to 6.7%.

**Paper-11: Bangla speech-to-text conversion using SAPI**

The aim of this study is to investigate Speech-to-Text (STT) conversion using SAPI for Bangla language. SAPI is a middleware that provides an API and a device driver interface (DDI) for speech engines to implement. Microsoft Windows 7 operating system supplies default recognition and synthesis speech engines. The speech engines are either speech recognizers or synthesizers. This study used English language as a middleware to manage SAPI for Bangla STT conversion. To recognize Bangla pronunciation, an xml grammar file for SAPI is generated with English character combinations for each Bangla word. The xml grammar file is then loaded into the SAPI to recognize the spoken words. SAPI returns an English character set when a spoken word is matched. Using this character set Bangla words are collected from a dictionary and write into a file. In this study, they employed several possible English character sets for a particular Bangla word that improves performance a certain level with respect to best possible English character set. The recognition rate was found 78% when the system is tested for an article of a newspaper. In the present study, Bangla speech is recognized word by word basis. A person should speak with a proper break in each word so that system writes the word if match occurs.

.