Audio Visual Speech Recognition using Deep Learning

S. Akshay
Dinesh M C
Karthik M A M

Dr. R S Milton (Supervisor)

SSN College of Engineering, Chennai

April 8, 2017

Problem Statement

- Use the visual information derived from a speaker's lips along with speech signals in order to improve the efficiency of a traditional speech recognition system.

Motivation

- Traditional speech recognition systems rely solely on the audio signals to predict text.
- The performance of such systems could be affected if the signal is corrupted by noise, arising due to diverse factors.
- Using visual information derived from a speakers lip movements, in addition to the audio features, alleviates the effects of noise.
- Moreover, opens up a host of applications: resolving multi-speaker simultaneous speech, dictating instructions over a phone in a noisy environment, improved hearing aids, etc.

Motivation Contd.

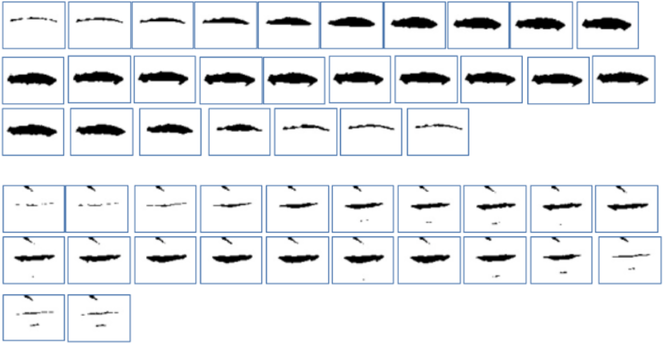

Figure: Sample Lip Movements for Letter 'a'

Data Preprocessing

- Need to preprocess the data to convert it into a form suitable for the model.
- Dataset
  - Dataset used: GRID Corpus.
  - Large multi-talker audiovisual sentence corpus.
  - Consists of high quality audio and video recordings of 1000 sentences spoken by 34 speakers.
  - Format: $\langle$command$\rangle$ $\langle$color$\rangle$ $\langle$preposition$\rangle$ $\langle$letter$\rangle$ $\langle$digit$\rangle$ $\langle$adverb$\rangle$.
  - Example: put blue at f two now.
- Audio
  - The audio files are provided as .wav files.
  - Mix noise with the audio files to increase generalisation capacity of model.
  - Extract 13 MFCC features for every 25ms window of the audio.
  - Store them in numpy files.

Data Preprocessing Contd.

- Video
  - The video frames are first extracted using ffmpeg library.
  - Blur the frames.
  - Apply Haar classifier for the Face Region of Interest.
  - Apply Haar classifier for the Mouth Region of Interest.
  - Alogrithm for width
    - Compress the image vertically, resulting in a single row.
    - Blur the row of the pixels obtained.
    - Then a density function is applied to this array, whose maximum value is the required width.
  - For height, transpose the image and apply the above procedure.
  - Store the results in a numpy array.

Data Preprocessing Contd.

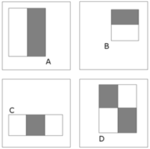

Figure: Haar Features Used

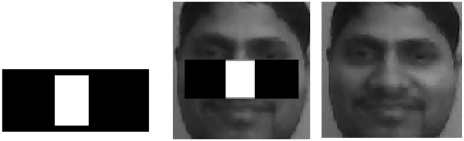

Figure: Sample Feature Application

Data Preprocessing Contd.

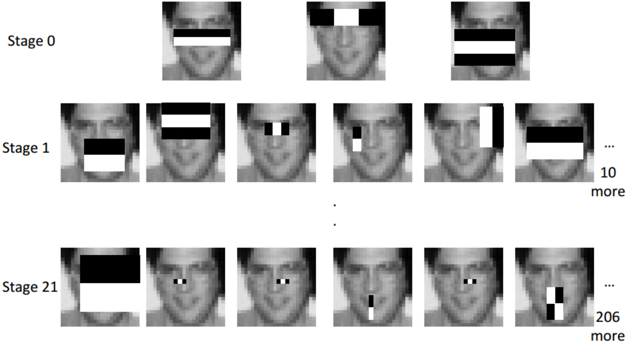

Figure: Haar Cascade Classifier

Data Preprocessing Contd.

Figure: Full Image

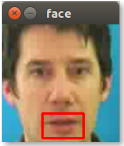

Figure: Extracted Face Image

Data Preprocessing Contd.

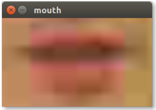

Figure: Extracted Mouth Image

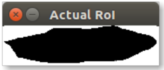

Figure: Actual Mouth Region Filtered

Data Preprocessing Contd.

- Labels
  - Transcriptions are provided as .align files.

```
0 23750 sil
23750 29500 bin
29500 34000 blue
34000 35500 at
35500 41000 f
41000 47250 two
47250 53000 now
53000 74500 sil
```

Figure: Sample Align File

  - Convert align files to text files after removing segmentation.
  - Map characters to class labels (0 for 'spaces' and 1-26 for a-z) and store them in numpy files.
  - For example, "abc xyz" is stored as [1, 2, 3, 0, 24, 25, 26].
- Training and Testing set
  - Divide the data into training and testing set to ensure that our model doesn't overfit to the data.
  - Use 80 per cent for training, and the remaining 20 per cent for testing.

Design

- Main component of our model consists of an Artificial Neural Network called a Bi-directional Long Short-Term Memory(LSTM) network.
- Neural Network
  - Learning structure consisting of a large number of simple neural units designed to mimic the function of a web of biological neurons.
  - Although successfully applied in the domain of Speech Recognition, cannot model temporal dependencies very well.

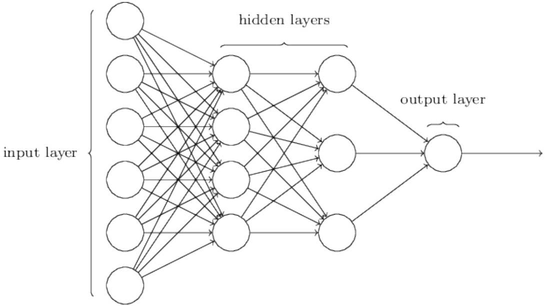

Figure: Typical Neural Network

Design Contd.

- Recurrent Neural Networks
  - Solve temporal dependency problem.
  - Networks with loops in them allowing information to persist.
  - Output at a timestep depends on the current input as well as the previous inputs, therefore they are ideal for Speech Recognition.
  - But, suffer from the problem of long term dependencies.

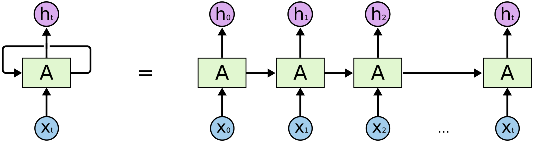

Figure: Recurrent Neural Network

Design Contd.

- Long Short-Term Memory Networks
  - Special kind of Recurrent Neural Network capable of learning long-term dependencies.
  - Have special structures called gates to store and manipulate information.

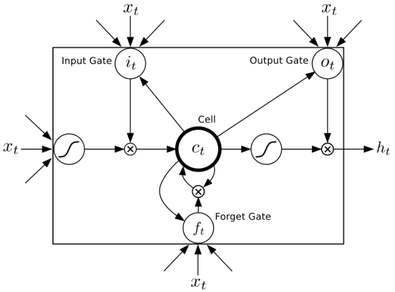

Figure: Long Short-Term Memory Cell

Design Contd.

- Bi-directional Long Short-Term Memory Networks
  - Bi-directional LSTMs are used because we can exploit future context as well.
  - Have two separate hidden layers which process data in both directions.

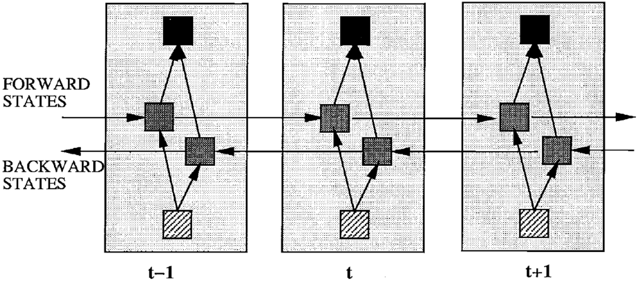

Figure: Bi-directional Network

Design Contd.

- Connectionist Temporal Classification
  - Novel method for labelling sequence data with RNNs that removes the need for pre-segmented training data.
  - Basic idea is to interpret the network outputs as a probability distribution over all possible label. sequences, conditioned on a given input sequence.
  - Has a softmax output layer with one more unit than there are labels.
  - The activations of the units are interpreted as the probabilities of observing the corresponding labels at particular times.
  - Together, these outputs define the probabilities of all possible ways of aligning all possible label sequences with the input sequence.

Architecture

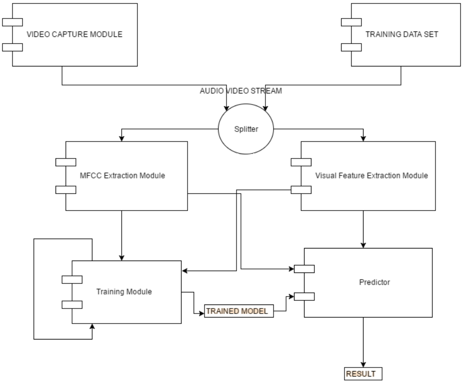

Figure: Main Architecture

Architecture Contd.

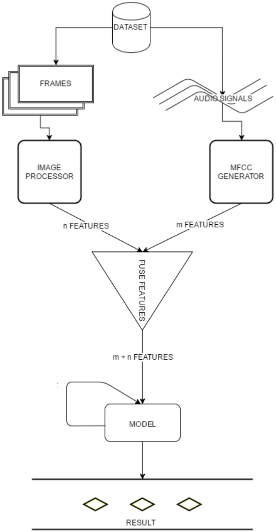

Figure: Model's Data Flow

Performance

- Metric: Edit distance

Type	Error Rate
Audio Only	0.0456
Noisy Audio	0.4396
Video Only	0.4027
Combined	0.0748

Conclusion

- Our experiments show us that if we combine the audio and video features of the speech signal, it performs much better in a noisy environment than a stand-alone audio model. It also produced better results than a stand-alone video model.

References

- Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016) ‘LipNet: Sentence-level Lipreading’, arXiv preprint arXiv:1611.01599.
- Cooke, M., Barker, J., Cunningham, S. and Shao, X. (2006) ‘An audio-visual corpus for speech perception and automatic speech recognition’, The Journal of the Acoustical Society of America, 120(5), pp.2421-2424.
- Graves, A., Fernndez, S., Gomez, F., & Schmidhuber, J. (2006) ‘Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks’, In Proceedings of the 23rd international conference on Machine learning, ACM, pp. 369-376.
- Hannun, Awni, et al. (2014) ‘Deep speech: Scaling up end-to-end speech recognition.’ arXiv preprint arXiv:1412.5567 .
- Viola, Paul, and Michael Jones (2001) ‘Rapid object detection using a boosted cascade of simple features.’, Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Vol. 1. IEEE, pp. I-I.
- http://colah.github.io/posts/2015-08-Understanding-LSTMs.