Video to Speech Reconstruction using Deep Learning

Dilan Chavda

Imperial College London

dirac.dc@gmail.com

September 13, 2017

Overview

Introduction

Lip reading models have been the subject of significant study in recent literature with the success of deep learning.

- Applications of these systems include helping the hearing impaired and creating subtitles.
- This work however has primarily been done by creating a model which maps video of a speakers lip movements to text.

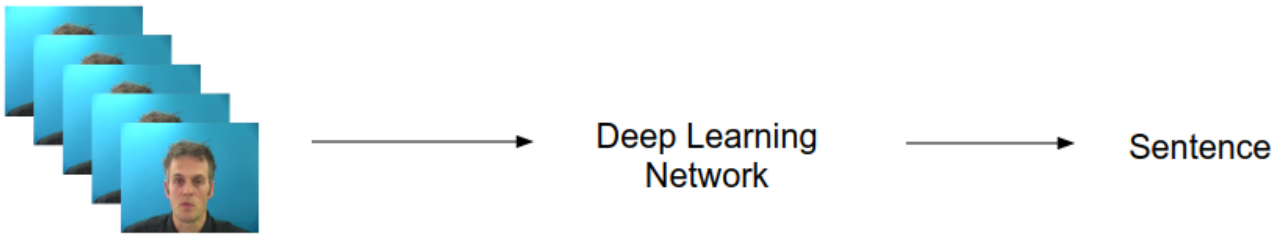
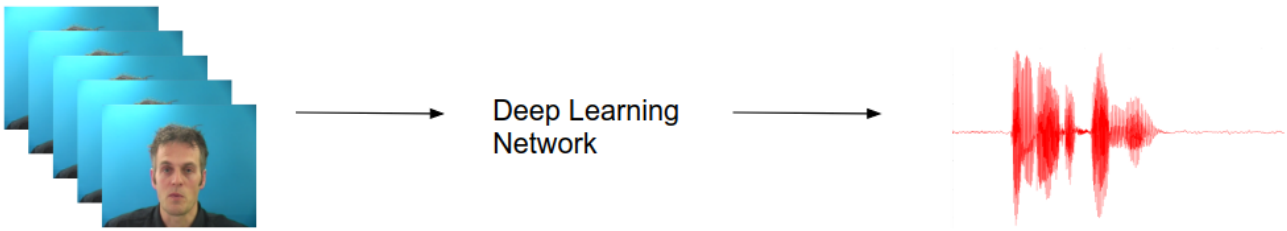


Figure: Typical Lip Reading System

Problem Description

In this paper we study an alternative approach to lip reading - convert from videos of lip movement to *speech.*

- Applications of these systems include helping surveillance, audio synchronization and computer animation.
- Very little work has been done on this problem, with only Ephrat and Peleg (2017) providing the proof of concept.



Motivation

Why study the video to speech mapping over the video to text mapping?

- It has 'natural supervision' with its synced audio track. This means no expensive or tedious labelling required.
- Can create language independent models. No need to include or learn grammar and potentially has no limit on vocabulary size.
- Speech waveforms may contain emphasis and emotion not available with text.

There are disadvantages as well:

- There are no easily interpretable classification scores without human listening tests. Intelligibility is hard to measure through analytic measures.
- Using regression output as opposed to softmax layers, so stricter modelling is required.
- Audio must be compressed to be targeted and there is a loss of information when reconstructed.

Brief Background

- The most directly related work has been in the field of *articulatory-acoustic* mapping. This involved the use of invasive sensors to collect data on articulatory organs and models to translate them into an audio representation.
- Intelligibility scores of 70-80% have been recorded.
- Artificial neural networks were also tried with decent success but the only paper to apply deep learning for video to speech production is that of Ephrat and Peleg (2017) showing intelligibility results of around 80% in line with state of the art results.

Methodology

To describe our experiments we will need to explain a few key items:

- Linear Predictive Coefficients
- Short-Time Fourier Transforms
- Recurrent Networks

It is assumed the basic structures of deep networks architectures such as neural networks, convolutional layers and max pooling are known.

Linear Predictive Coefficients (LPCs)

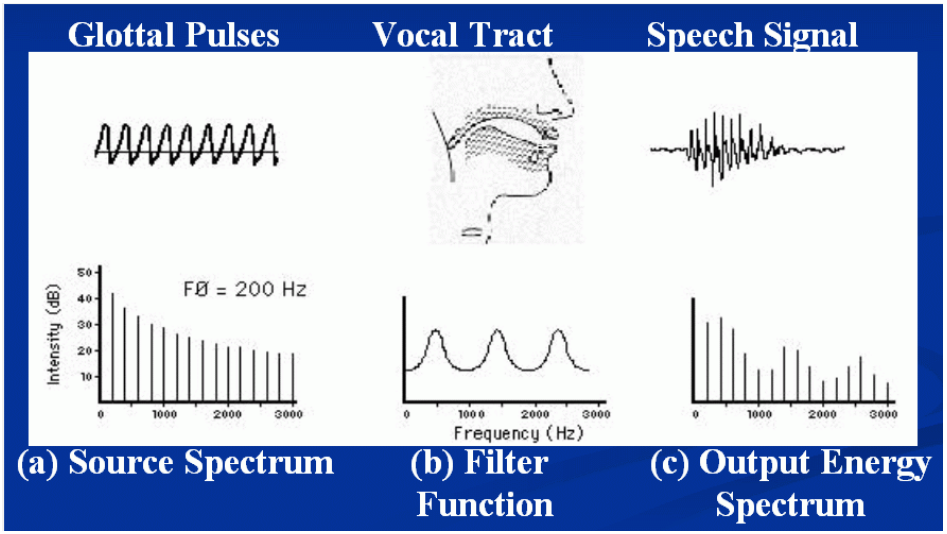


Figure: Basic LPC Model

- Throat creates an initial source of sound (different for voiced and unvoiced sounds)
- Source signal passes through mouth cavity and experiences resonance.
- Final signal is produced.
- LPCs capture information about the energy and filter shape function.

Short Time Fourier Transform (STFT)

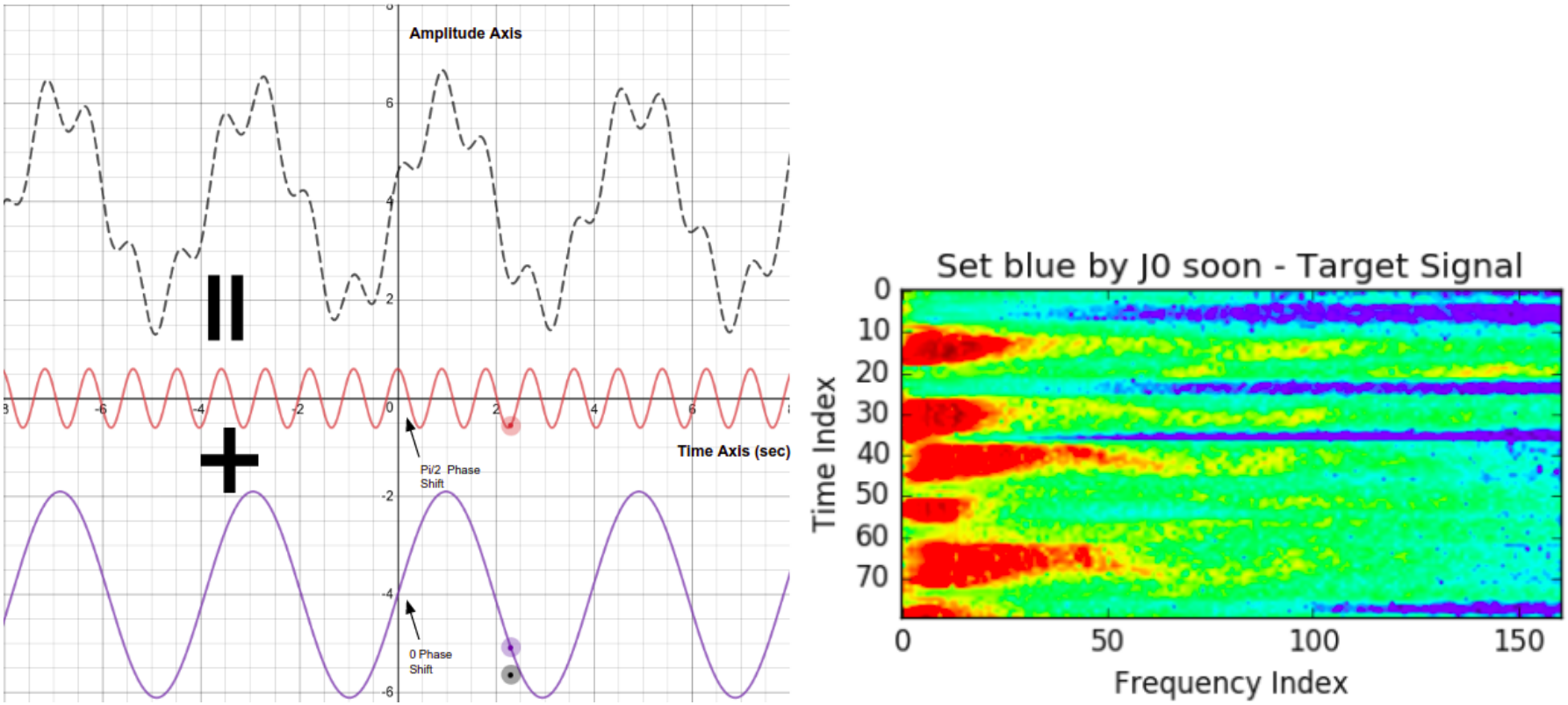


Figure: STFT example and output.

- Any finite signal can be looked as the linear combination of sinusoidal waves of different phases,amplitudes and increasing frequencies.

Recurrent Neural Networks

These are neural networks with a memory. A closed loop allows information of the systems previous state to the next state.

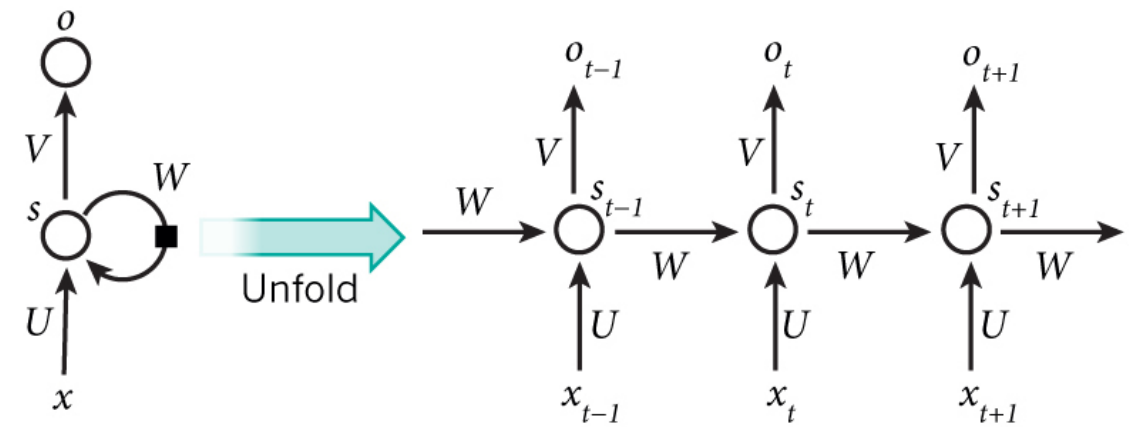


Figure: Typical Basic RNN Structure

Experiments

We first expanded on the deep architecture used by Ephrat and Peleg (2017):

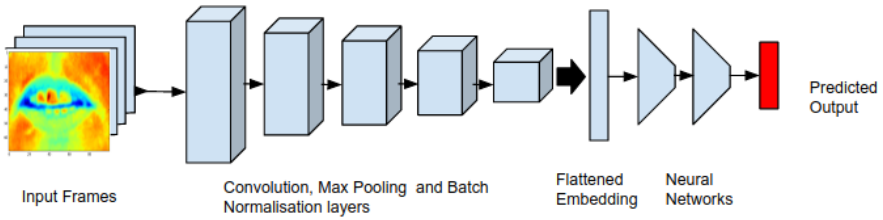


Figure: Ephrat and Peleg Architecture

- Strictly convolutional.
- Targets LP coefficients.
- Uses mean squared loss function.

Changing Parameters and Settings

We first briefly tested the effect of changing typical parameters of their model including:

- Learning Rate Strategy and Batch Size
- Full face or lips only input data
- Window Length (Number of Time Frames used as Input)
- Number of convolutional layers

Our results suggested the following:

- RMSProp and a smaller batch size provided better results.
- Lip Only data performs better than full face data.
- Larger length windows lead to more predictive output.
- More convolutional layers improves performance.

Adding LSTM networks

Ephrat and Pelegs model was strictly convolutional leaving scope for the use of recurrent networks. We tested the addition of recurrent networks including Stateful,Stateless and Bidirectional LSTMS and GRUs integrated using the typical architecture:

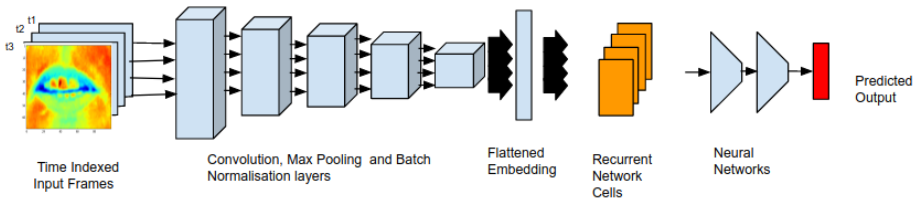
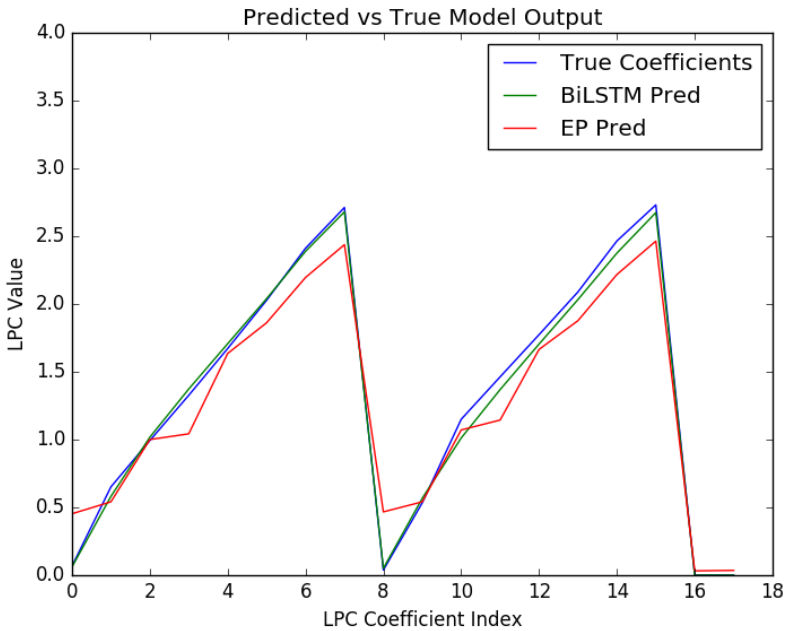


Figure: Typical Recurrent Architecture.

We found that a network involving bi-directional LSTMs provided the best results.

Comparison of our Model to Ephrat and Peleg Model

Our final model using a bidirectional LSTM to that of the original Ephrat and Peleg model and showed an 8% reduction in test error.



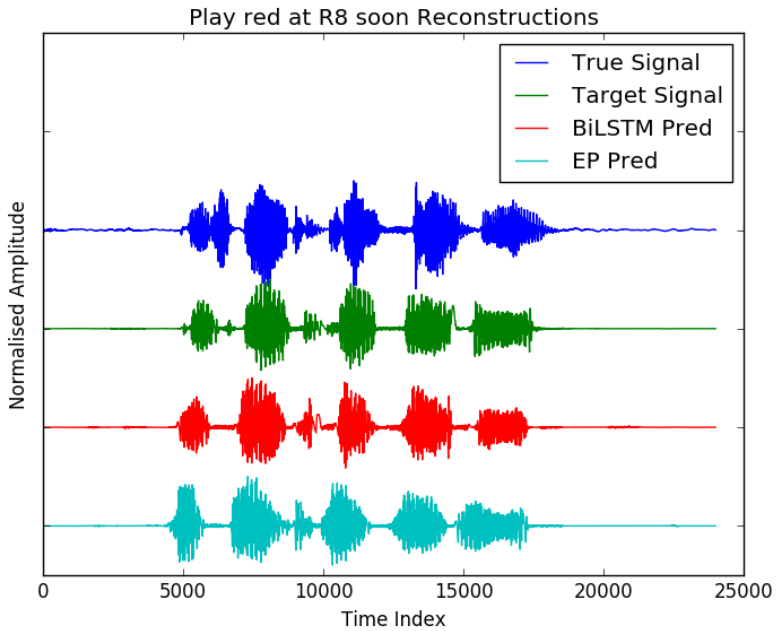


Figure: Ephrat and Peleg vs Bidirectional LSTM model predicted coefficients.

Bidirectional LSTM Model Example

‘Lay Blue by R Zero Again’

True Audio

Predicted Audio

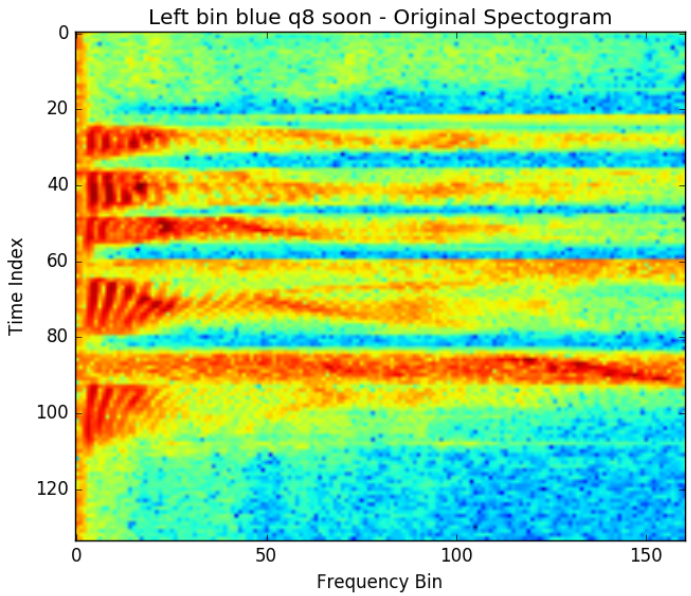
Changing Target Features

We tried using STFT features as opposed to LPC coefficients, with the same bidirectional LSTM architecture.

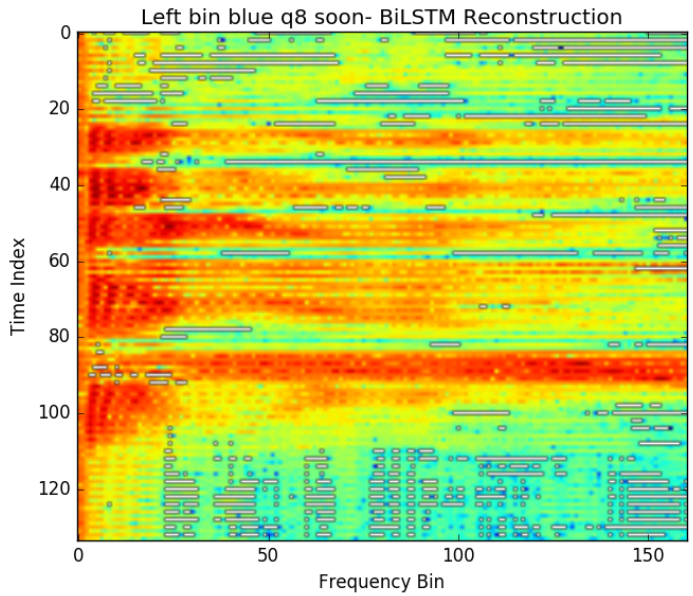
- Models could not converge with phase information so used magnitude.
- Used three signal recoveries from the predicted magnitude spectrograms:
  - No phase recovery
  - True phase recovery
  - Reconstructed phase recovery (using algorithm proposed by Griffin and Lim (1984))

STFT Model Results

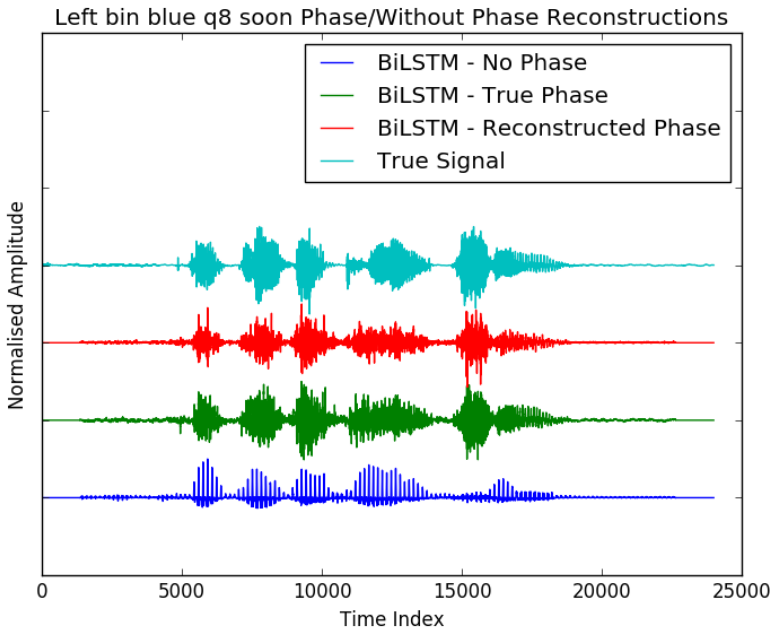
True



Predicted



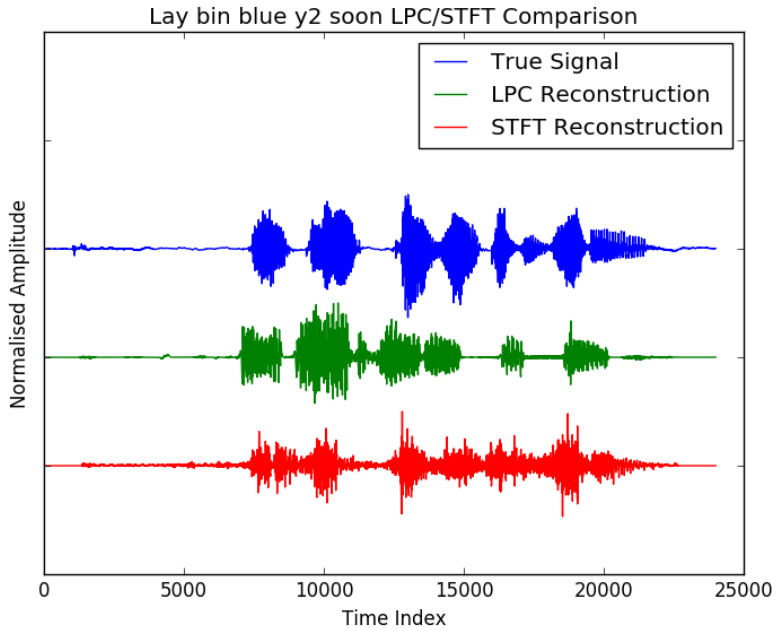
STFT Model Results

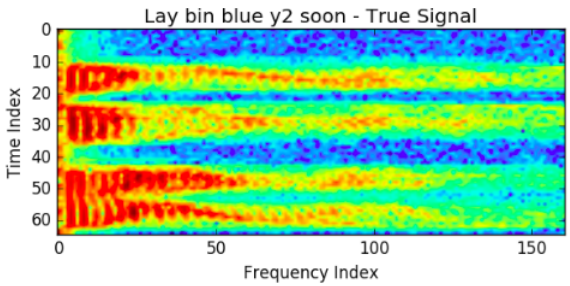


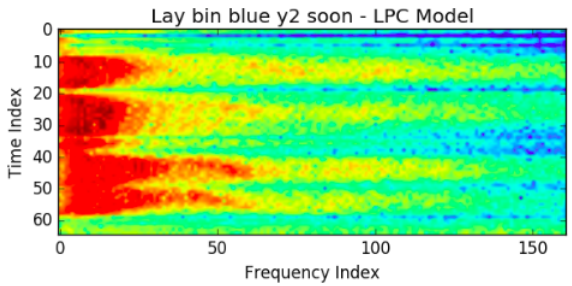
‘Lay Blue by R Zero Again’

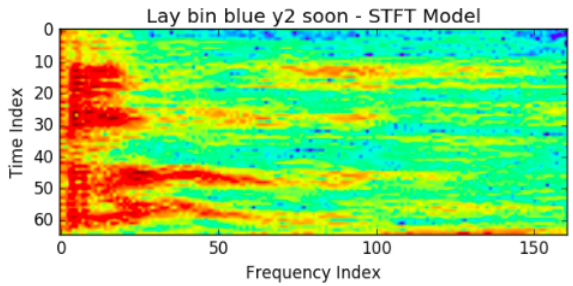
No Phase True Phase Reconstructed Phase

LPC vs STFT Comparison









‘Lay Blue by R Zero Again’

LPC Signal STFT Signal

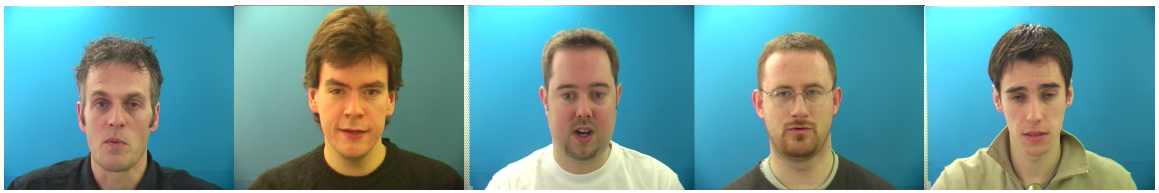
Adapting to Multiple Speakers

Adapting these models for training/testing on multiple speakers is **not** a trivial task.

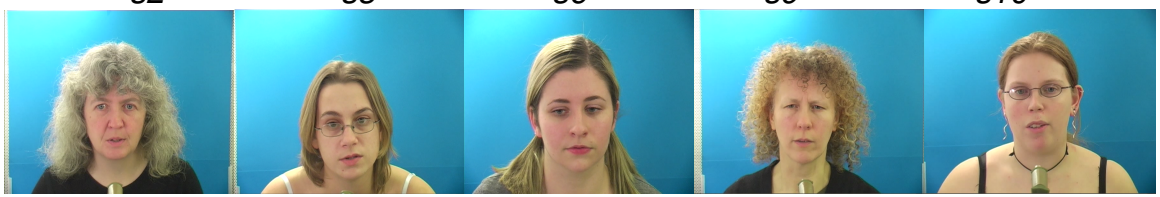
- What representation should be used?
- What voice does the final output have?

Four models were trained *s2*, *Male*, *Female* and *Mix*.

OBSERVED



s2 *s3* *s6* *s9* *s10*

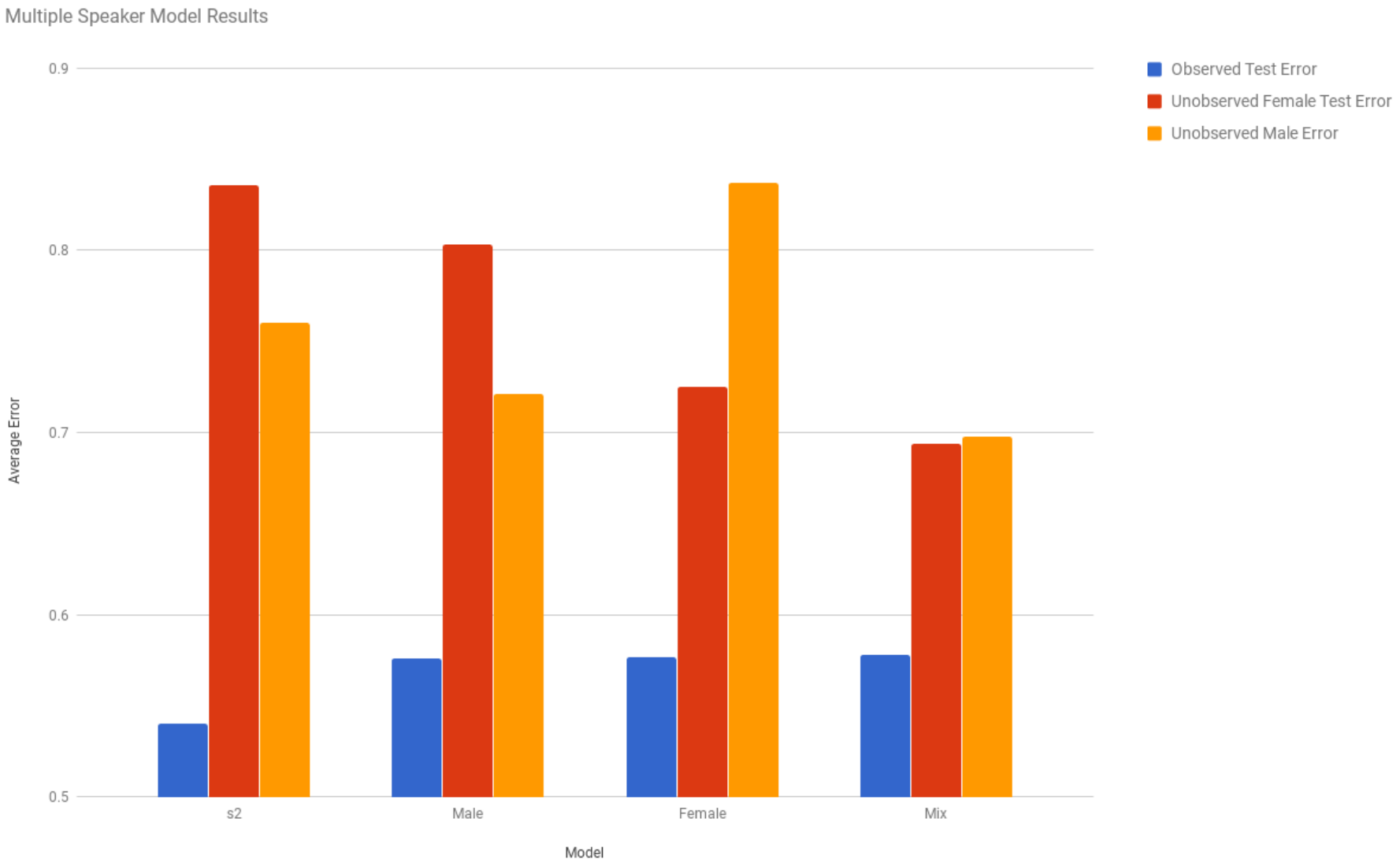


s4 *s7* *s11* *s15* *s16*

UNOBSERVED

s12 *s25*

Multiple Speaker Results



Key Limitations

- Our analytic measures such as correlation and Signal to Noise ratios do not capture intelligibility accurately.
- Trained on relatively small dataset.
- Better phase reconstruction algorithms are possible.
- Deeper architectures were limited by computer memory.

Further Research

- Study activations of layers and use ‘crop tests’ to find out which part of the mouth regions are important to improve multiple speaker models.
- Test models on a dataset with a larger vocabulary and multiple viewpoints.
- Look at developing better phase reconstructions or different target features such as wavelet transforms.
- Use more elaborate architectures including the use of attention and residual networks.

References

Ephrat, A. and Peleg, S. (2017). *Vid2speech: Speech Reconstruction from Silent Video.* arXiv:1701.00495 [cs].

Griffin, D. and Lim, J. (1984). *Signal estimation from modified short-time Fourier transform.* https://www.researchgate.net/publication/3177517_Signal_estimation_from_modified_short-time_Fourier_transform

The End

Griffin and Lim Algorithm

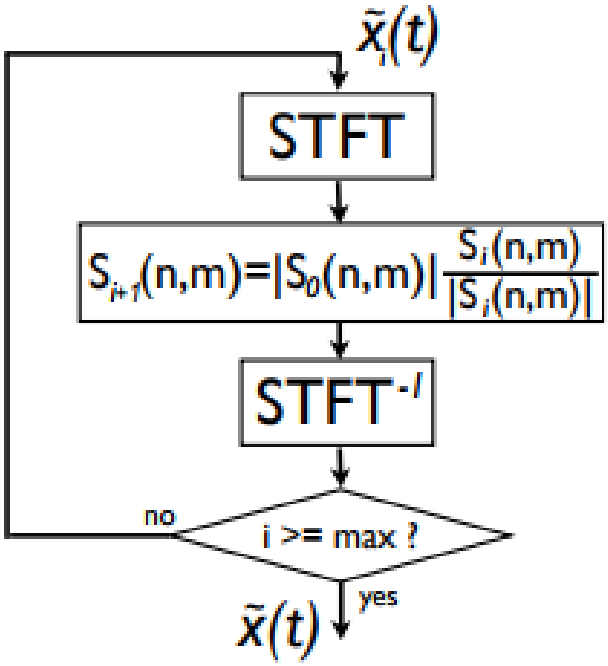


Figure: Griffin and Lim Algorithm

LSTM

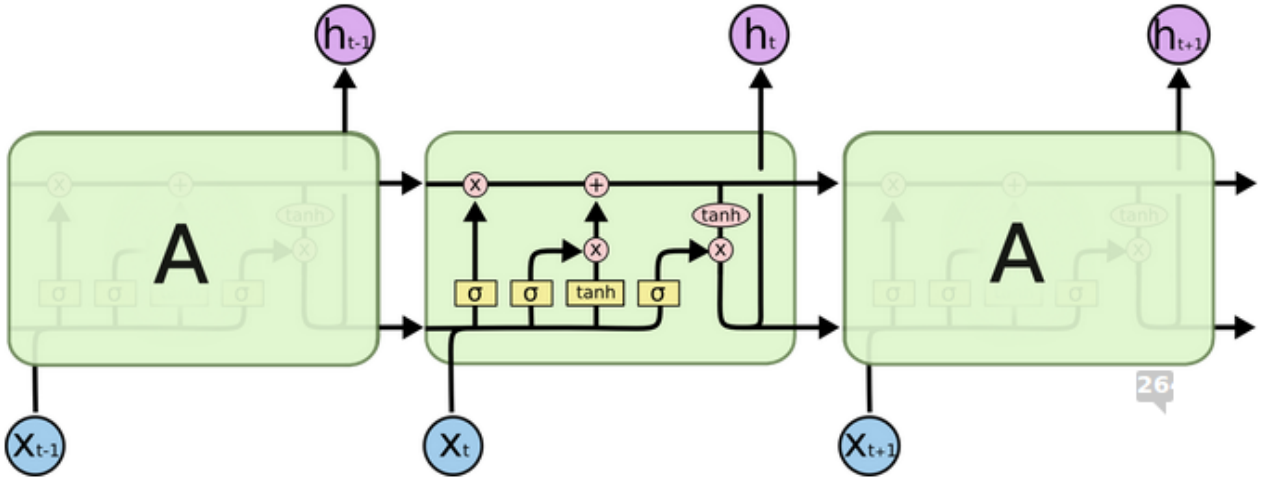


Figure: Working Diagram of a LSTM

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f); \tag{1}$$

$$\hat{\boldsymbol{C}}_t = tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c) \quad ; \quad \boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{C}_t = \boldsymbol{i}_t * \boldsymbol{C}_{t-1} + \boldsymbol{f}_t * \hat{\boldsymbol{C}}_t \tag{3}$$

$$\boldsymbol{h}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) * tanh(\boldsymbol{C}_t) \tag{4}$$