READ MY LIPS

Abhinav Reddy
Ankita Agarwal
Arjun Surendran
Nazim Shaikh
Suchismita Sahu

LIPNET - PRESENT STATE-OF-THE-ART

LIPNET IN APPLICATION

DESCRIPTION

The goal of this project is to recognize phonemes being uttered by a person using video frames as input.

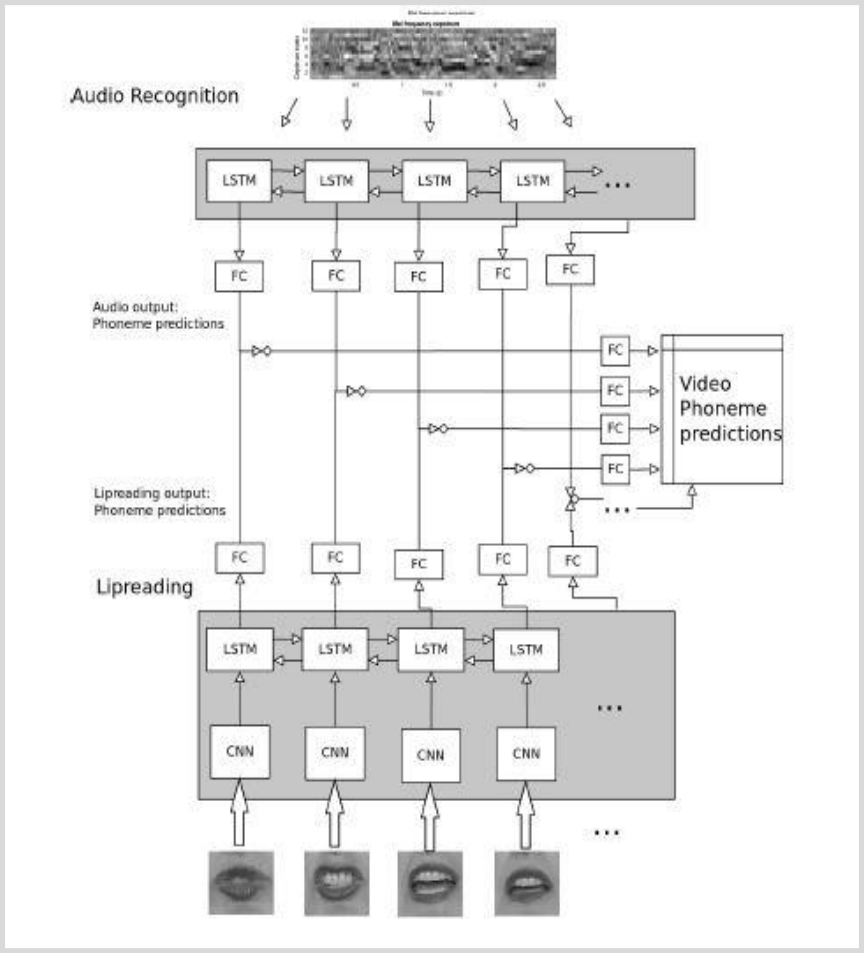
Phase 1: We will create a model that can identify correct phoneme(s) spoken from audio using LSTM & FC Networks.

Phase 2: We will simultaneously train CNN-LSTM on video frames to predict visemes

End Goal: To detect phonemes from sequential images

Reference : “Design, implementation and analysis of a deep convolutional-recurrent neural network for speech recognition through audiovisual sensor fusion” by Matthijs Van keirsbilck

BASIC NETWORK ARCHITECTURE



ARCHITECTURE FOR VIDEO/IMAGE PROCESSING

The lipreading network uses the WLAS CNN with 2 bidirectional LSTM layers on top.

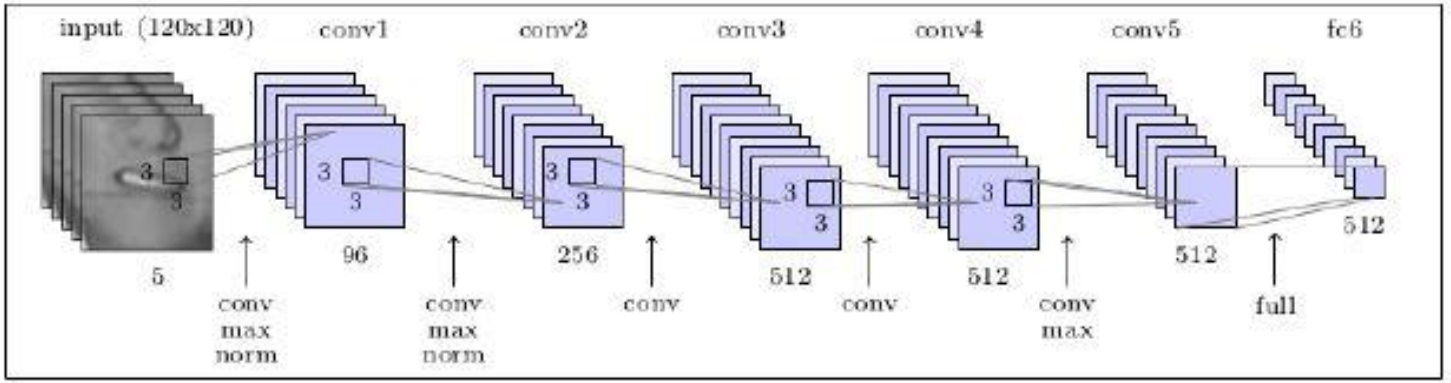


Figure 4.5: The CNN network used in WLAS

DATA PHONEME - VISEME MAPPING

Viseme	TIMIT Phonemes	Description
/V1	/ao/ /ah/ /aa/ /er/ /oy/ /aw/ /hh/	Lip-rounding based vowels
/V2	/uw/ /uh/ /ow/	"
/V3	/ae/ /eh/ /ey/ /ay/	"
/V4	/ih/ /iy/ /ax/	"
/A	/l/ /el/ /r/ /y/	Alveolar-semivowels
/B	/s/ /z/	Alveolar-fricatives
/C	/t/ /d/ /n/ /en/	Alveolar
/D	/sh/ /zh/ /ch/ /jh/	Palato-alveolar
/E	/p/ /b/ /m/	Bilabial
/F	/th/ /dh/	Dental
/G	/f/ /v/	Labio-dental
/H	/ng/ /g/ /k/ /w/	Velar
/S	/sil/ /sp/	Silence


a, e
ar
b
d
ə
ee
er
f
g
h
i
k
l
m
n
ng, nk
o
oo
oo
or
p
r
s, x
sh, ch
t
th
th
u
v
w, wh, qu
y
z

AUDIO PROCESSING

State-of-the-art performance is currently reached by recurrent networks, specifically multilayer bidirectional LSTM networks .

The hyperparameters are set as follows:

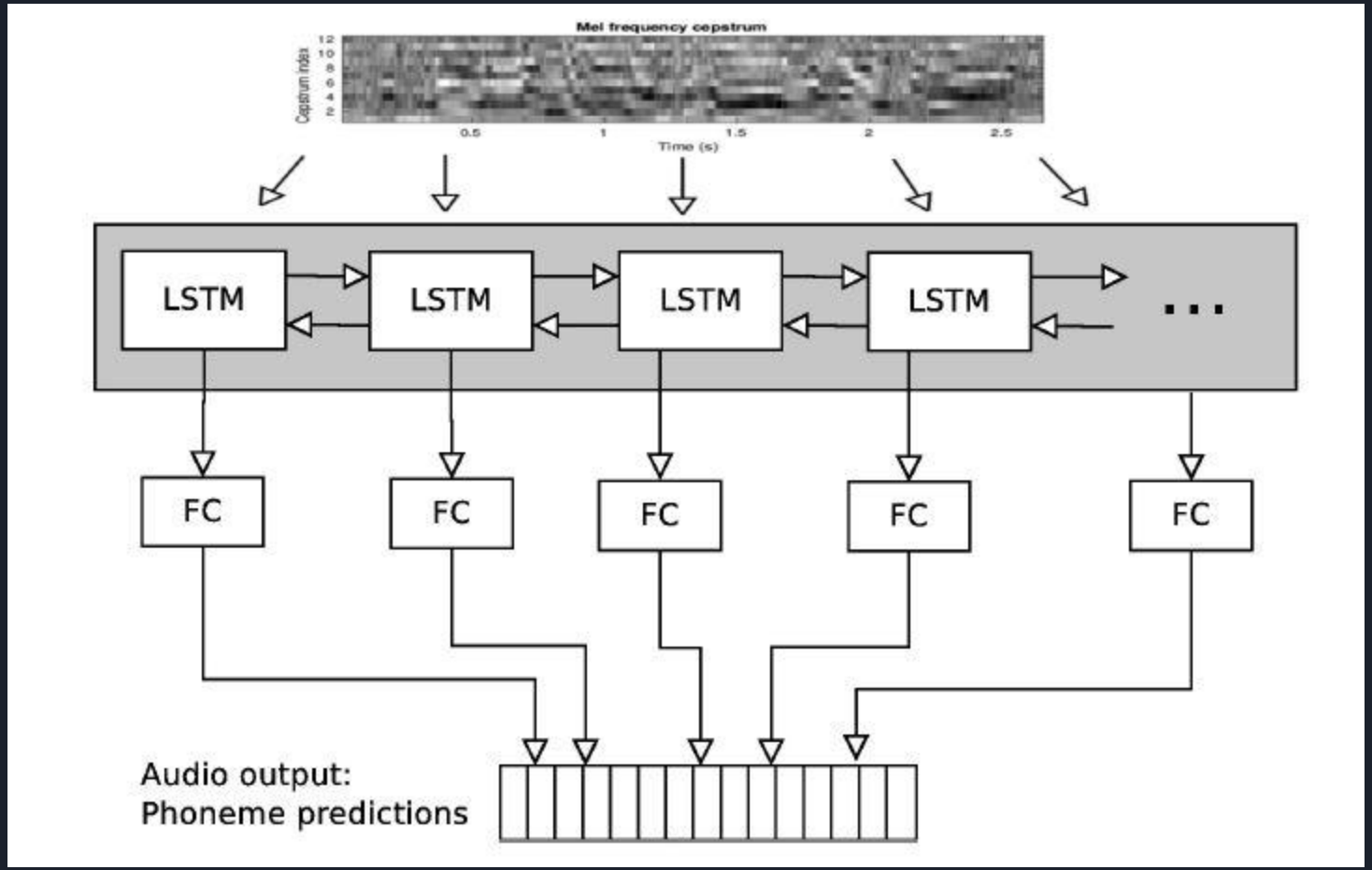
• Input features: 13 MFCC coefficients, their first and second derivatives, for a total of 39 input features.

• MFCC frame length of 10 ms, with overlapping window size of 25ms.

• Bidirectional LSTM layers

• 39 phoneme output classes

- We intend to go further and try to predict all 44 phonemes in english language.


Mel frequency cepstrum
Cepstrum index
12
10
8
6
4
2
0.5
1
1.5
2
2.5
Time (s)
LSTM
LSTM
LSTM
LSTM
...
FC
FC
FC
FC
FC
Audio output:
Phoneme predictions

General network structure of a deep (bidirectional) LSTM network used:

• K layers, with Ni LSTM units in each layer

• The LSTM units on the first layer are fully connected to the features of one frame of the input sequence.

• For bidirectional networks, each layer contains two ’sublayers’ of LSTM units,
traversing the input sequence in opposite directions.

• The output features of the last LSTM layer are fed through a softmax (FCNN) layer for classification.

DATA AVAILABILITY

- Primary Dataset : TCD_TIMIT dataset - It consists of 13826 video clips in MP4 format, yielding high-quality audio and video footage of 62 speakers reading a total of 6913 phonetically rich sentences from the TIMIT corpus.

Vowels:

iy	beet	bcl b IY tcl t
ih	bit	bcl b IH tcl t
eh	bet	bcl b EH tcl t
ey	bait	bcl b EY tcl t
ae	bat	bcl b AE tcl t
aa	bott	bcl b AA tcl t
aw	bout	bcl b AW tcl t
ay	bite	bcl b AY tcl t
ah	but	bcl b AH tcl t
ao	bought	bcl b AO tcl t
oy	boy	bcl b OY
ow	boat	bcl b OW tcl t
uh	book	bcl b UH kcl k
uw	boot	bcl b UW tcl t
ux	toot	tcl t UX tcl t
er	bird	bcl b ER dcl d
ax	about	AX bcl b aw tcl t
ix	debit	dcl d eh bcl b IX tcl t
axr	butter	bcl b ah dx AXR
ax-h	suspect	s AX-H s pcl p eh kcl k tcl t

DATA AVAILABILITY

```
Orthography (.txt):
        0 61748 She had your dark suit in greasy wash water all year.

Word label (.wrd):
        7470 11362 she
        11362 16000 had
        15420 17503 your
        17503 23360 dark
        23360 28360 suit
        28360 30960 in
        30960 36971 greasy
        36971 42290 wash
        43120 47480 water
        49021 52184 all
        52184 58840 year
```

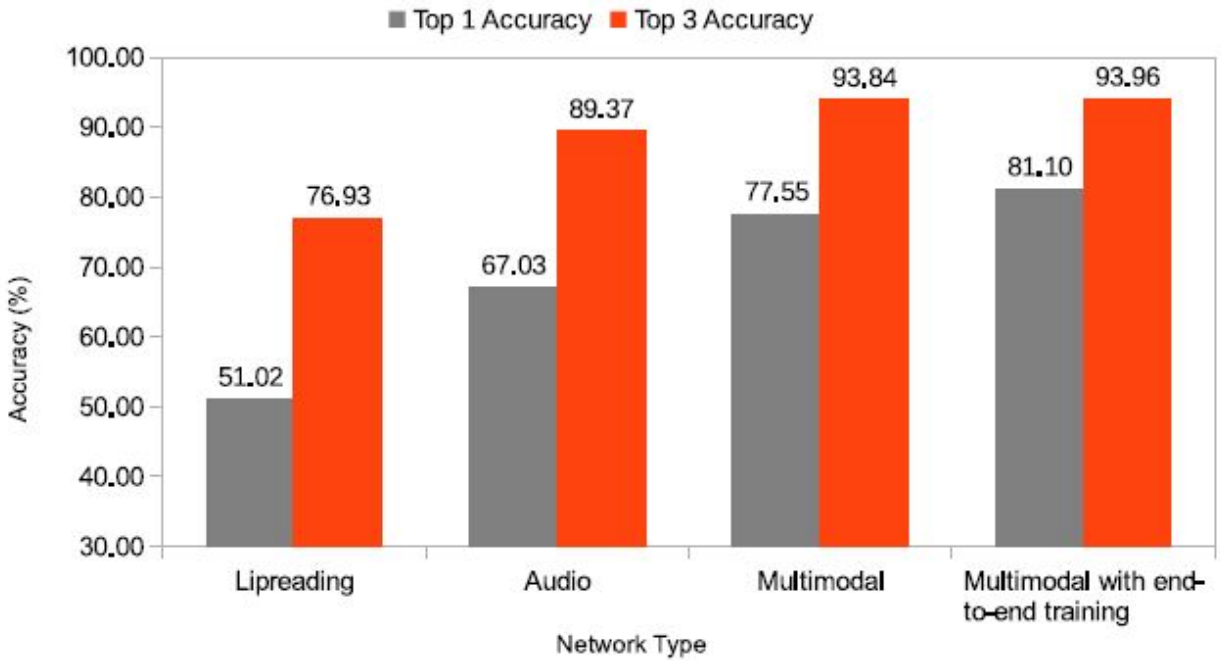
For each spoken sentence in the audio-video dataset, the dataset also contains time-aligned words and phonemes .

```
Phonetic label (.phn):
(Note: beginning and ending silence regions are marked with h#)
        0 7470 h#
        7470 9840 sh
        9840 11362 iy
        11362 12908 hv
        12908 14760 ae
        14760 15420 dcl
        15420 16000 jh
        16000 17503 axr
        17503 18540 dcl
        18540 18950 d
        18950 21053 aa
        21053 22200 r
        22200 22740 kcl
        22740 23360 k
        23360 25315 s
        25315 27643 ux
        27643 28360 tcl
        28360 29272 q
        29272 29932 ih
        29932 30960 n
        30960 31870 gcl
        31870 32550 g
        32550 33253 r
        33253 34660 iy
        34660 35890 z
        35890 36971 iy
        36971 38391 w
        38391 40690 ao
        40690 42290 sh
```

FEASIBILITY STUDY

- The TCD-TIMIT dataset contains 23GB of audio-visual data.
- We are expecting around 1-2 days of training to train all the networks (Phase 1 and Phase 2). We will be splitting the data in batches and training.

RESULTS FROM PAPER



PRESENT STATE-OF-THE-ART

Method	Dataset	Size	Output	Accuracy
Fu et al. (2008)	AVICAR	851	Digits	37.9%
Hu et al. (2016)	AVLetter	78	Alphabet	64.6%
Papandreou et al. (2009)	CUAVE	1800	Digits	83.0%
Chung & Zisserman (2016a)	OuluVS1	200	Phrases	91.4%
Chung & Zisserman (2016b)	OuluVS2	520	Phrases	94.1%
Chung & Zisserman (2016a)	BBC TV	> 400000	Words	65.4%
Gergen et al. (2016)	GRID	29700	Words*	86.4%
LipNet	GRID	28775	**Sentences**	**95.2%**

Note from source: Existing lipreading datasets and the state-of-the-art accuracy reported on these.

Source: Assael et al. (2016, p. 3)

FURTHER SCOPE

- We intend to further implement Attention Networks for implementing sequence to sequence prediction for word-level speech recognition on top of the phoneme predictions for scalability of the project.

REFERENCES

- “Lip Reading sentences in the wild” https://www.robots.ox.ac.uk/~vgg/publications/2017/Chung17/chung17.pdf
- “Read My Lips” cs230.stanford.edu/files_winter_2018/projects/6940477.pdf
- “Lip reading using CNN and LSTM” cs231n.stanford.edu/reports/2016/pdfs/217_Report.pdf
- “Multimodal speech recognition using lipreading (with CNNs) and audio (using LSTMs)” https://github.com/matthijsvk/multimodalSR
- Harte, N.; Gillen, E., "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," Multimedia, IEEE Transactions on , vol.17, no.5, pp.603,615, May 2015 doi: 10.1109/TMM.2015.2407694