



Engaging Content
Engaging People

CAN DNNs LEARN TO LIPREAD FULL SENTENCES ?

George Sterpu, Christian Saam, Naomi Harte

Trinity College Dublin, Ireland 



MIT
Technology
Review

Topics+ The Down

Intelligent Machines

AI Has Beaten Humans at Lip-reading

A pair of new studies show that a machine can understand what you're saying without hearing a sound.


by Jamie Condliffe November 21, 2016

EMERGING TECH

Lip reading AI smashes humans at interpreting silent sentences

SHARE

LipNet: How easy do you think lipreading is?



Luke Dormehl
@lukedormehl

POSTED ON
11.25.16 - 3:57PM

DIGITAL TRENDS

Emergent Tech ▶ Artificial Intelligence

The Register
Biting the hand that feeds IT

Is that you, HAL? AI can now see secrets through lipreading – kinda

LipNet's got potential but also a loooong way to go

By [Katyanna Quach](#) 8 Nov 2016 at 00:59

14 SHARE ▼

TECH \ ARTIFICIAL INTELLIGENCE \

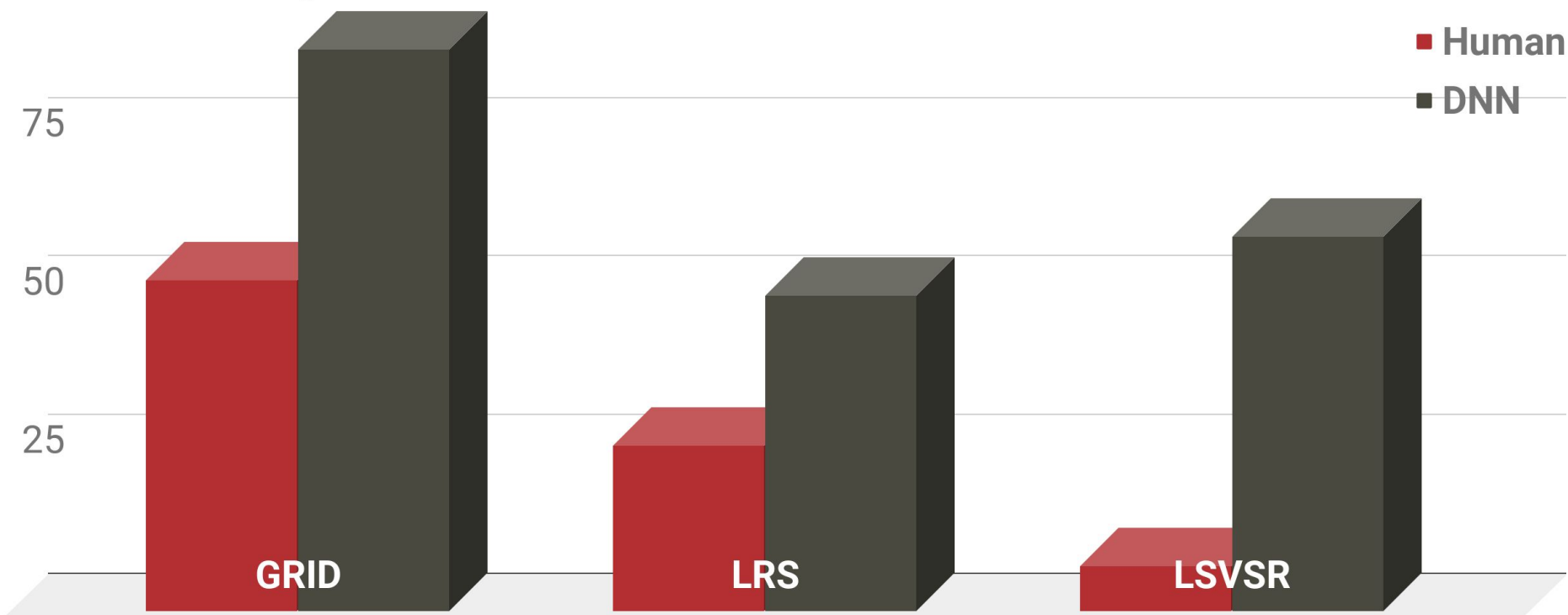
THE VERGE

Can deep learning help solve lip reading?

New research paper shows AI easily beating humans, but there's still lots of work to be done

By [James Vincent](#) | [@jvincent](#) | Nov 7, 2016, 12:50pm EST

Word Accuracy - Human vs AI



*28 hours
laboratory
LipNet*

*246 hours
“in the wild”
WLAS*

*3,886 hours
“in the wild”
LipNet 2*

What is lip-reading ?



What is lip-reading ?







Prior work

“end to end”

- collect lots of data
- train a large DNN
- decode **text**

“traditional”

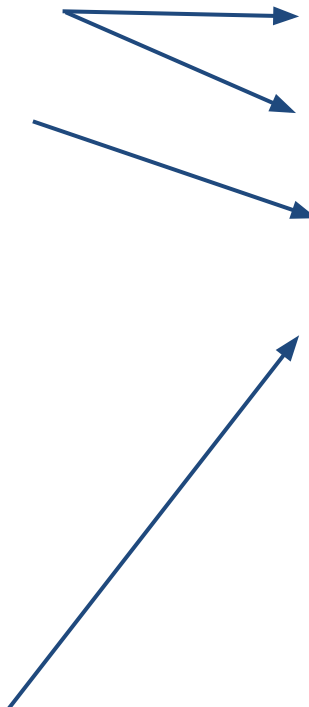
- simpler task
- DCT (etc.) + HMM
- decode **visemes**



Our work

“best of both worlds”

- large vocabulary
- full sentences
- train a small DNN
- decode **visemes**



Visemes = unambiguous units



Viseme 13

Phonemes	Word Examples
ch, sh, jh, zh	cheap, sheep



Viseme 15

Phonemes	Word Examples
f, v	fail, veil



Viseme 16

Phonemes	Word Examples
m, em, b, p	mat, bat, pat

speakers	62
sentences	6619
size	9 hours
vocabulary	6224 words
resolution	1920x1080

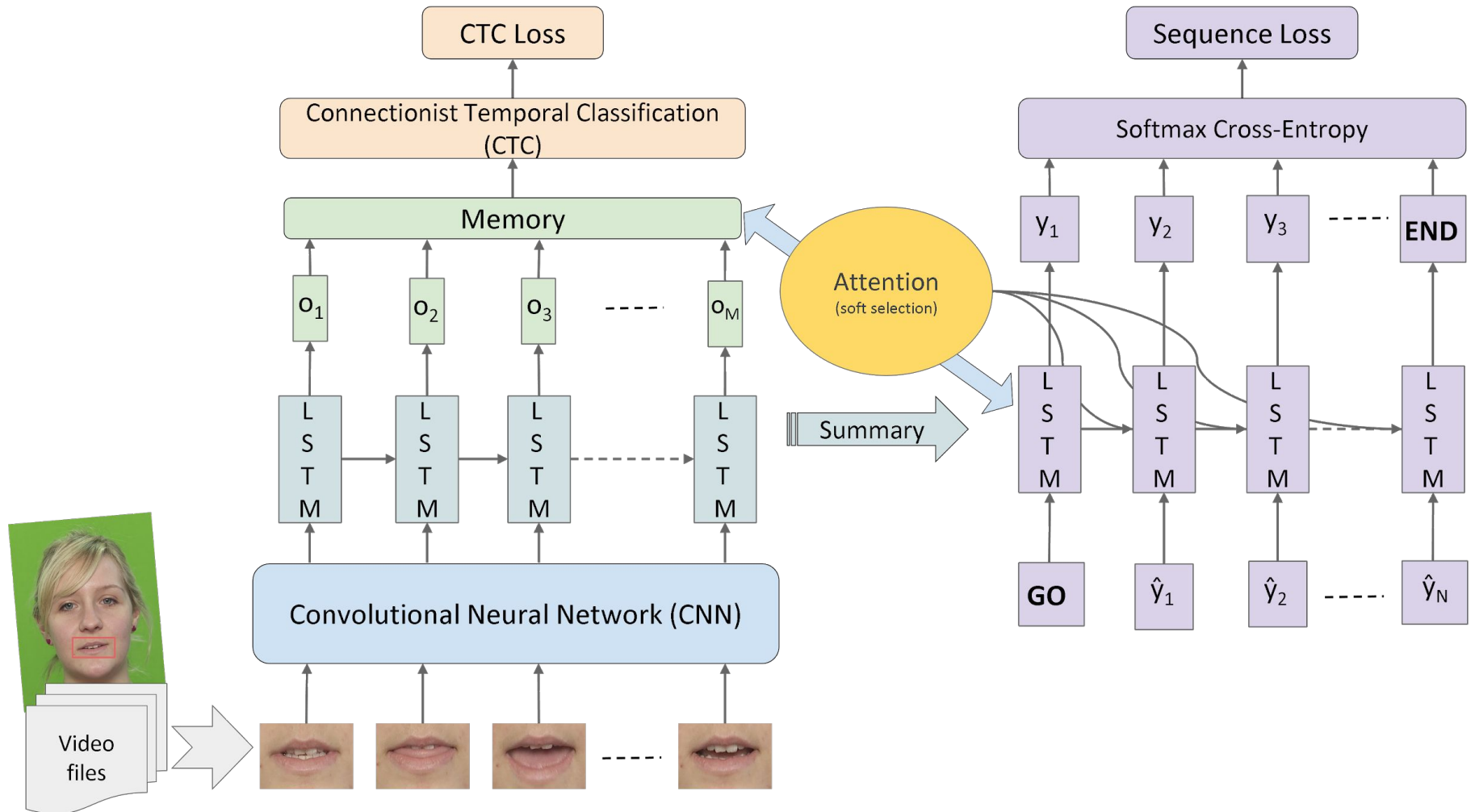


Publicly available:
sigmedia.tcd.ie/TCDTIMIT

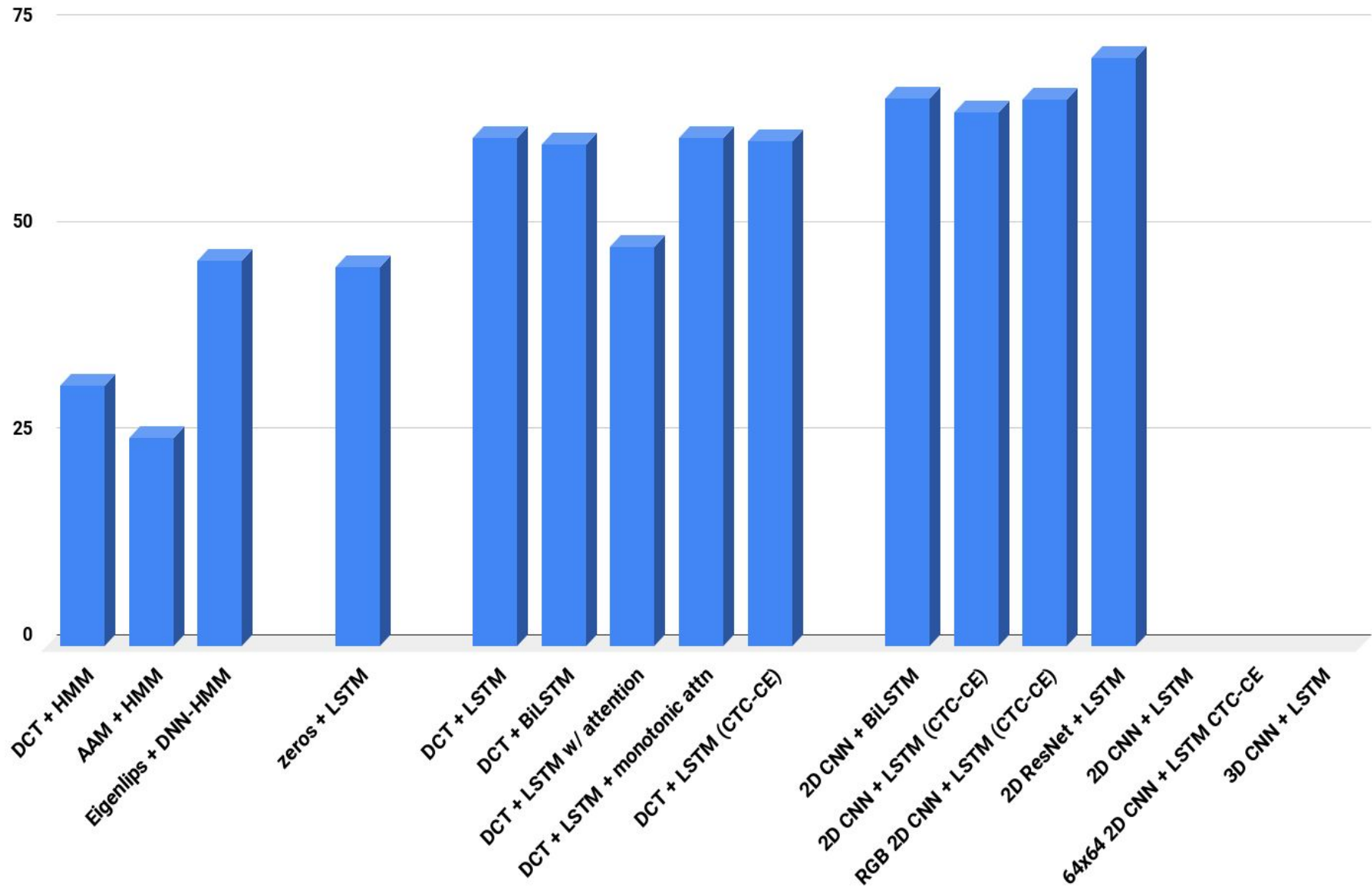
A few examples:

- *he took his mask from his forehead and threw it unexpectedly across the deck*
- *civilization is what man has made of himself*
- *the mayan neoclassic scholar disappeared while surveying ancient ruins*

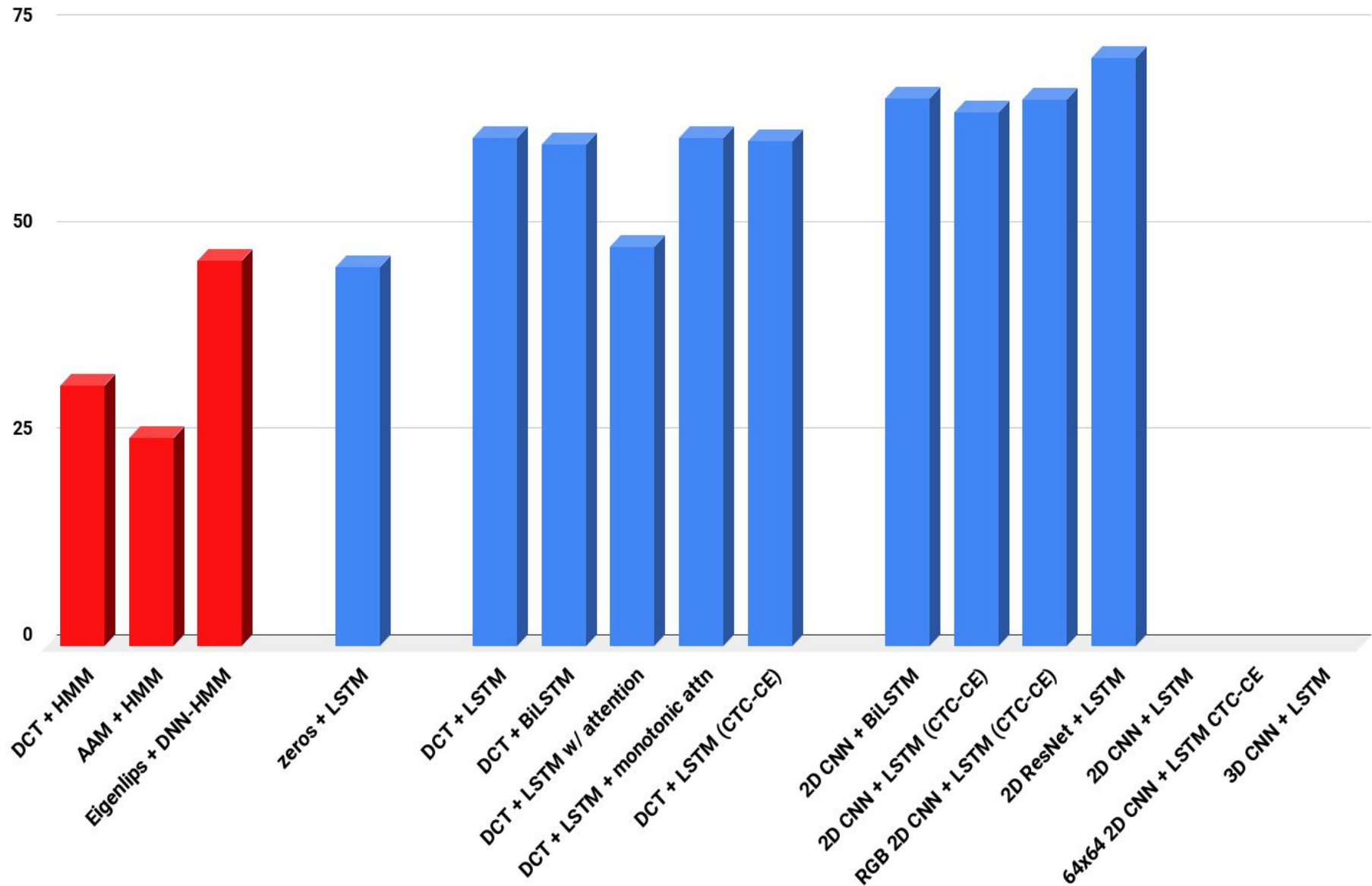
CNN + Seq2seq architecture



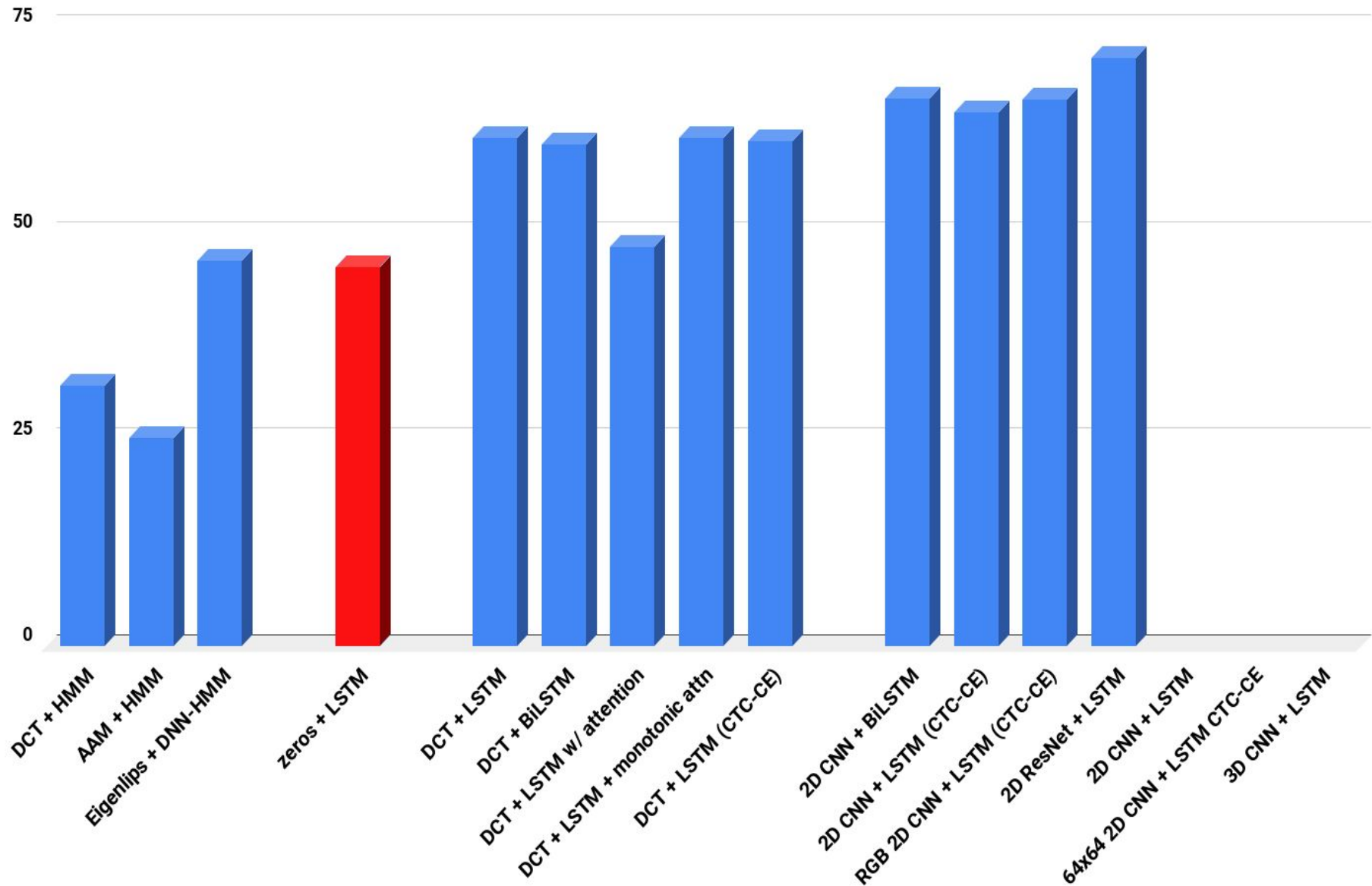
Viseme recognition accuracy [%]



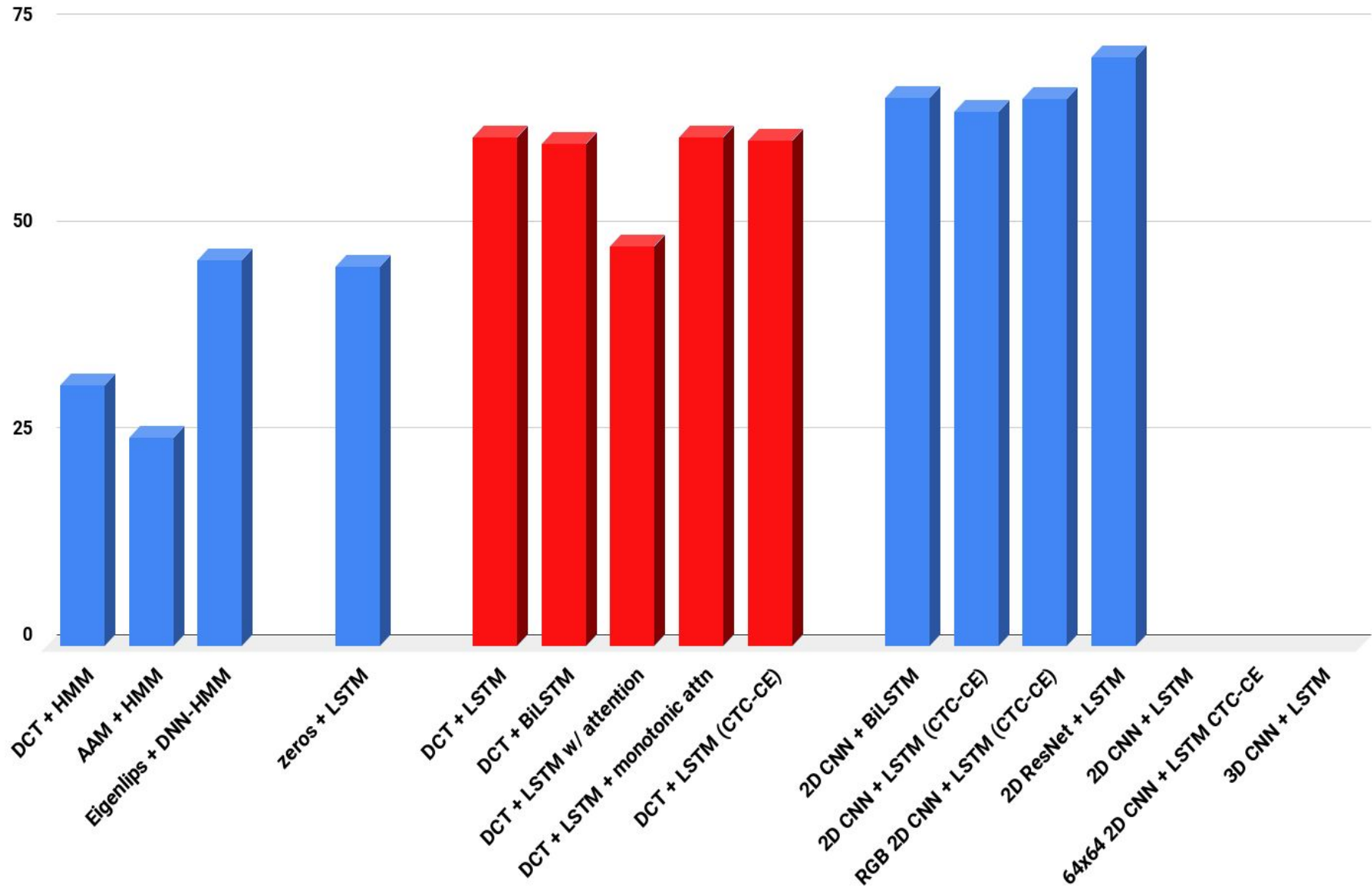
Viseme recognition accuracy [%]



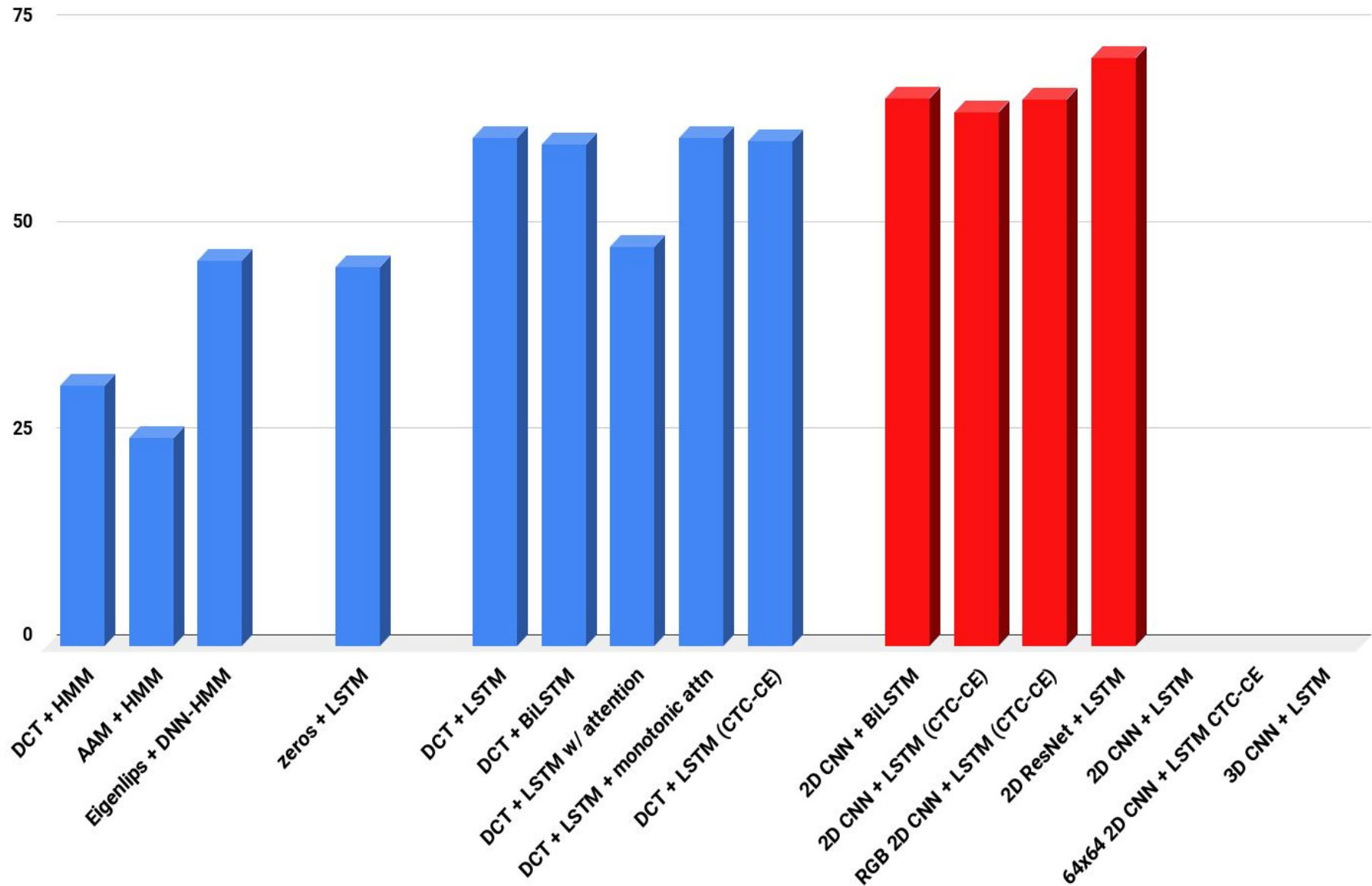
Viseme recognition accuracy [%]



Viseme recognition accuracy [%]



Viseme recognition accuracy [%]






Our work

“Visual Sentence to Visemes DNN”

- Pros:
 - easier to train
- Cons:
 - manually defined visemes
 - not for Video-only applications
- Potential:
 - Audio-Visual fusion
 - automatically learn visual speech units

- Our code is publicly available on Github
 -  [github.com / georgesterpu / **Sigmedia-AVSR**](https://github.com/georgesterpu/Sigmedia-AVSR)
 - TensorFlow-based seq2seq network for **Speech Recognition**
 - Audio only
 - Visual only
 - Audio-Visual fusion

 @John_Tukey