VI-04

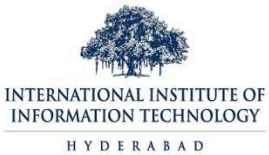
Towards Automatic Face-to-Face Translation

Originally published in ACM Multimedia, 2019

"Towards a translation system for the audio-visual age"

K R Prajwal[*+], **Rudrabha Mukhopadhyay**[*+], Jerin Philip[+], A. Jha[+], Vinay Namboodiri[#], C.V. Jawahar[+]

[+]IIIT Hyderabad, [#]IIT Kanpur

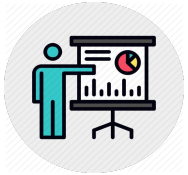


*** The first two authors have contributed equally in this work.**

Talking face videos

VI-04

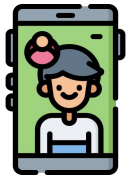
TV News broadcast and interviews



Thousands of lecture and MOOC videos

Thousands of movies in local languages

Making characters talk in animated movies



Video conferencing

½ of the smartphone users make video calls[1]

300 hrs of videos uploaded to YouTube per minute[1]

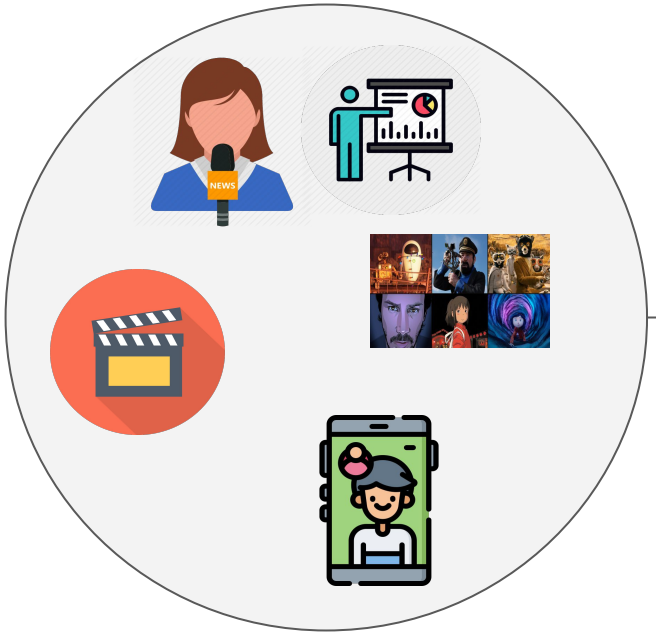
[1] NPD Press release, 2016 (link)

All images are taken from Google Images for illustration purposes only.

Translating talking face videos

VI-04

Language A

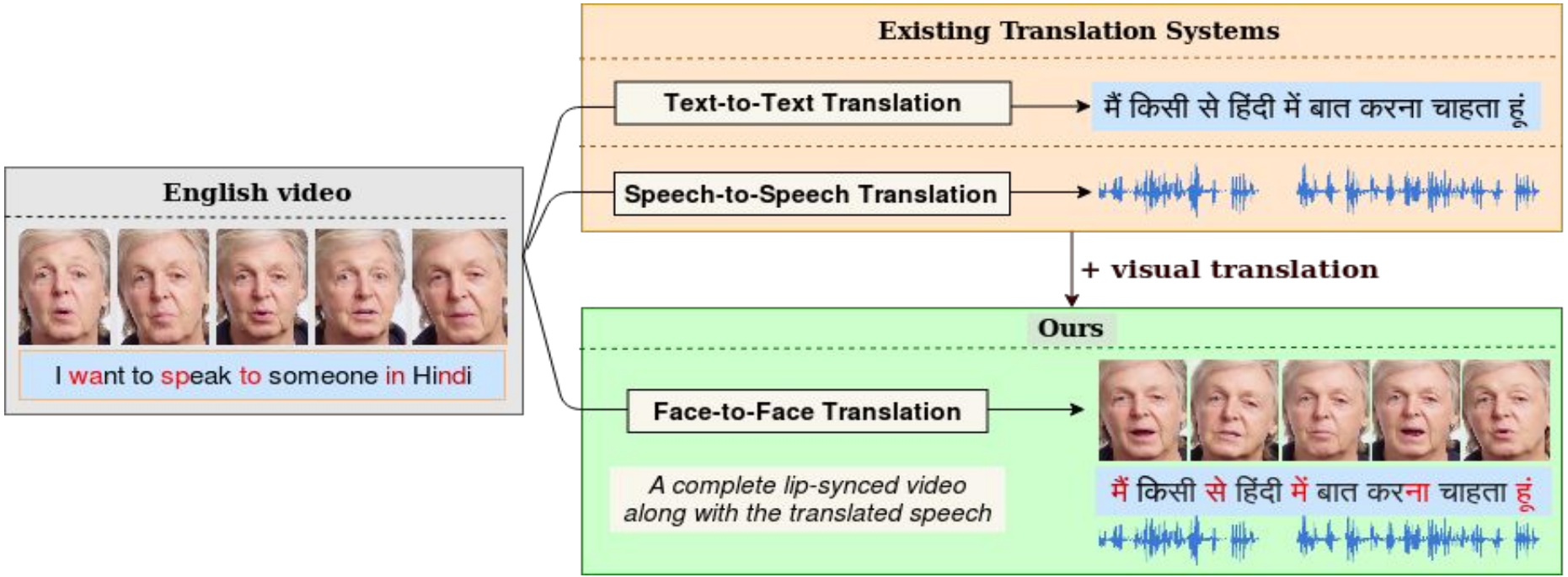


Viewer who knows only

Language B

Evolution of translation systems

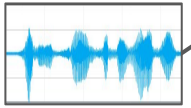
VI-04



Face-to-Face Translation Pipeline

VI-04

Original Video
in Language L1



Speech in
Language L1

Speech-to-Speech Translation

Automatic Speech Recognition

Machine Translation

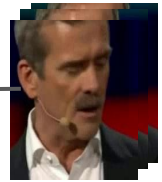
Text-to-Speech System + Voice Transfer

Speech in
Language L2

Lip Synchronization

LipGAN

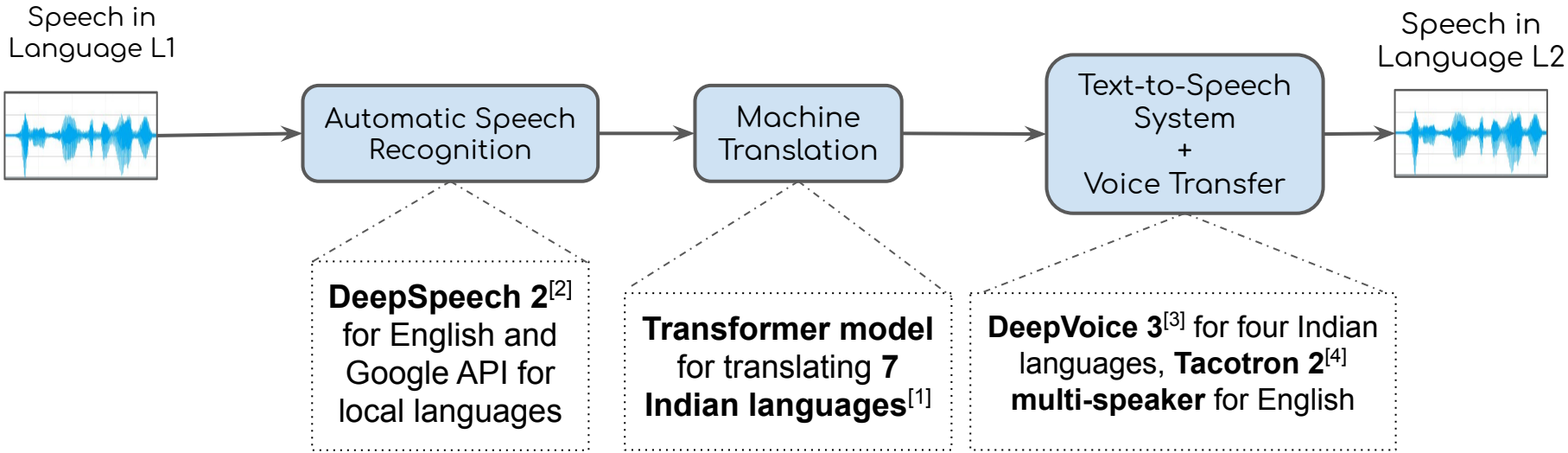
Lip-Synced Video in
Language L2



Original Video
in Language L1

Speech-to-Speech Translation

VI-04



[1] CVIT-MT Systems for WAT-2018, arXiv 2019
[2] Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, ICML 2015
[3] Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning, ICLR 2017
[4] Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, arXiv 2017

Recent works in Talking Face Generation

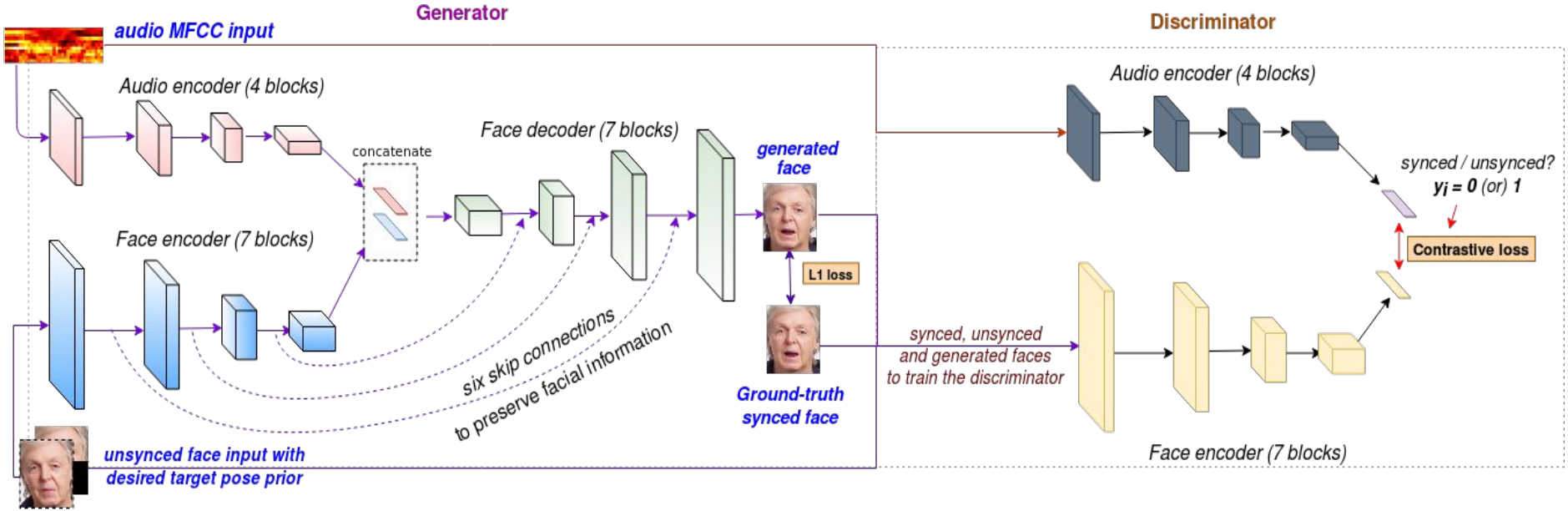
VI-04

Paper	Works for any face	Cross-language	Self-supervised	Seamless blending	Extra supervision for accurate lip-sync
Synthesizing Obama [1]	x	x	✓	✓	x
ObamaNet [2]	x	x	x	✓	x
Talking Face [3]	✓	x	x	x	✓
You Said That? [4]	✓	✓	✓	x	x
LipGAN (ours)	✓	✓	✓	✓	✓

[1] Synthesizing Obama: Learning Lip Sync from Audio, *SIGGRAPH 2017*
[2] ObamaNet: Photo-realistic lip-sync from text, *NeurIPS 2017 Workshop*
[3] Talking Face Generation by Adversarially Disentangled Audio-Visual Representation, *AAAI 2019*
[4] You Said That? , *BMVC 2017*

LipGAN

VI-04


Generator
audio MFCC input
Audio encoder (4 blocks)
Face encoder (7 blocks)
concatenate
Face decoder (7 blocks)
generated face
L1 loss
six skip connections
to preserve facial information
Ground-truth synced face
unsynced face input with desired target pose prior
Discriminator
Audio encoder (4 blocks)
synced / unsynced?
y_i = 0 (or) 1
Contrastive loss
synced, unsynced and generated faces to train the discriminator
Face encoder (7 blocks)

Results

VI-04

Face-to-Face Translation

LipGAN on Animated Characters

Thank you

VI-04

We thank Google for providing a travel grant which allowed me to travel to France and attend ACM Multimedia, 2019

For more information please visit,
https://cvit.iiit.ac.in/research/projects/cvit-projects/facetoface-translation

Please contact

{prajwal.k, radrabha.m} @ research.iiit.ac.in.