An Overview about Lip Reading & Audio-visual Speech Recognition

Chen Chen

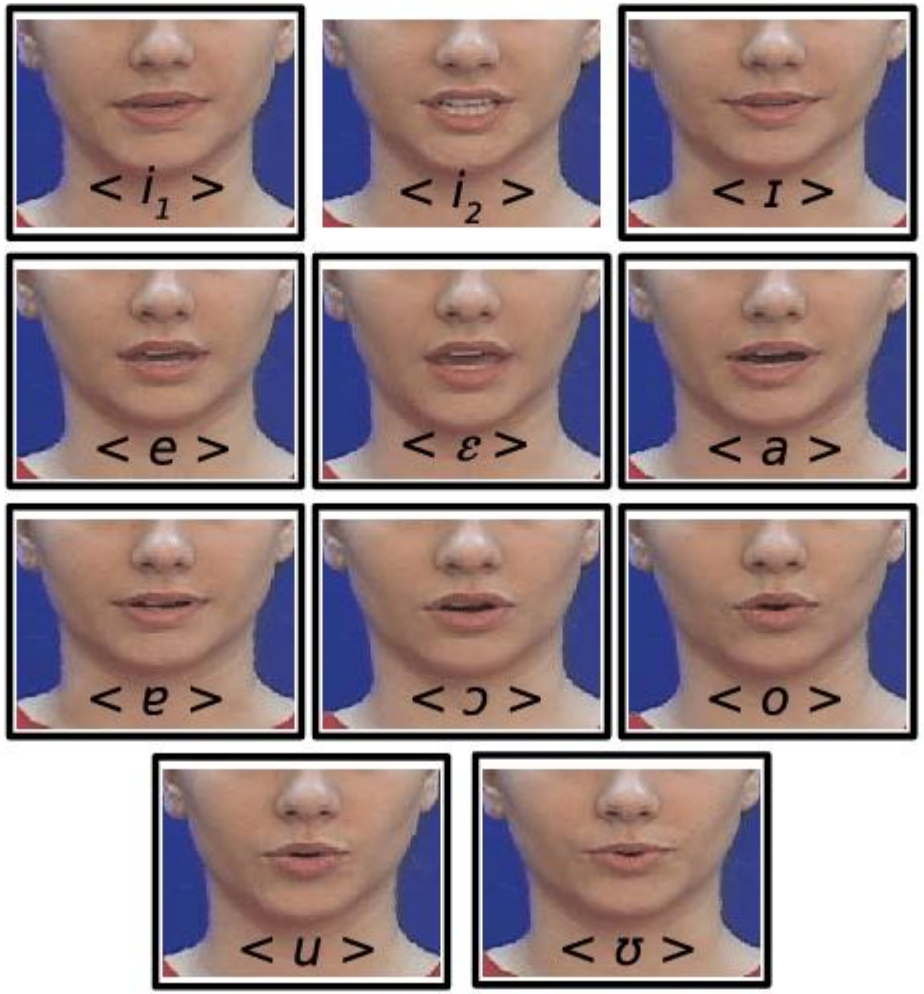
2022/03/18

Catalog

- Introduction
- Pipeline
- Datasets & Performance Evaluation
- Methods

Definition

- Lip Reading
    - recognize what is being said from visual information alone
- Audio-visual Speech Recognition
    - recognize what is being said from both audio and visual information

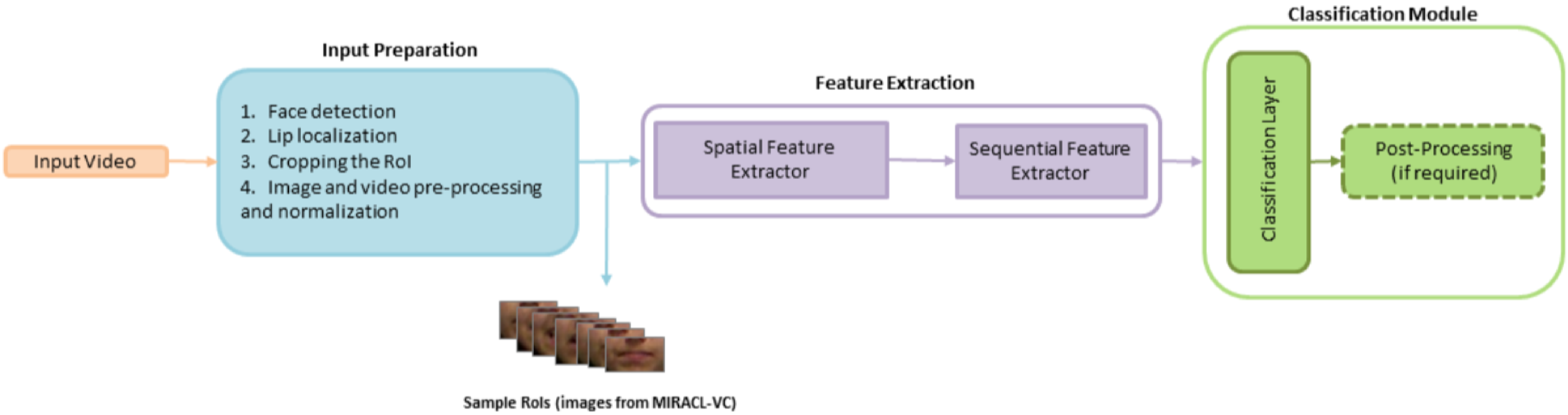


Viseme

Challenges

- Subject dependent factors
  - speaker variation
- Video quality factors
  - pose variation
  - unsynchronized audio & video
- Content-based factors
  - homophones (same viseme but different phoneme)
- Single modality dominate
- ...

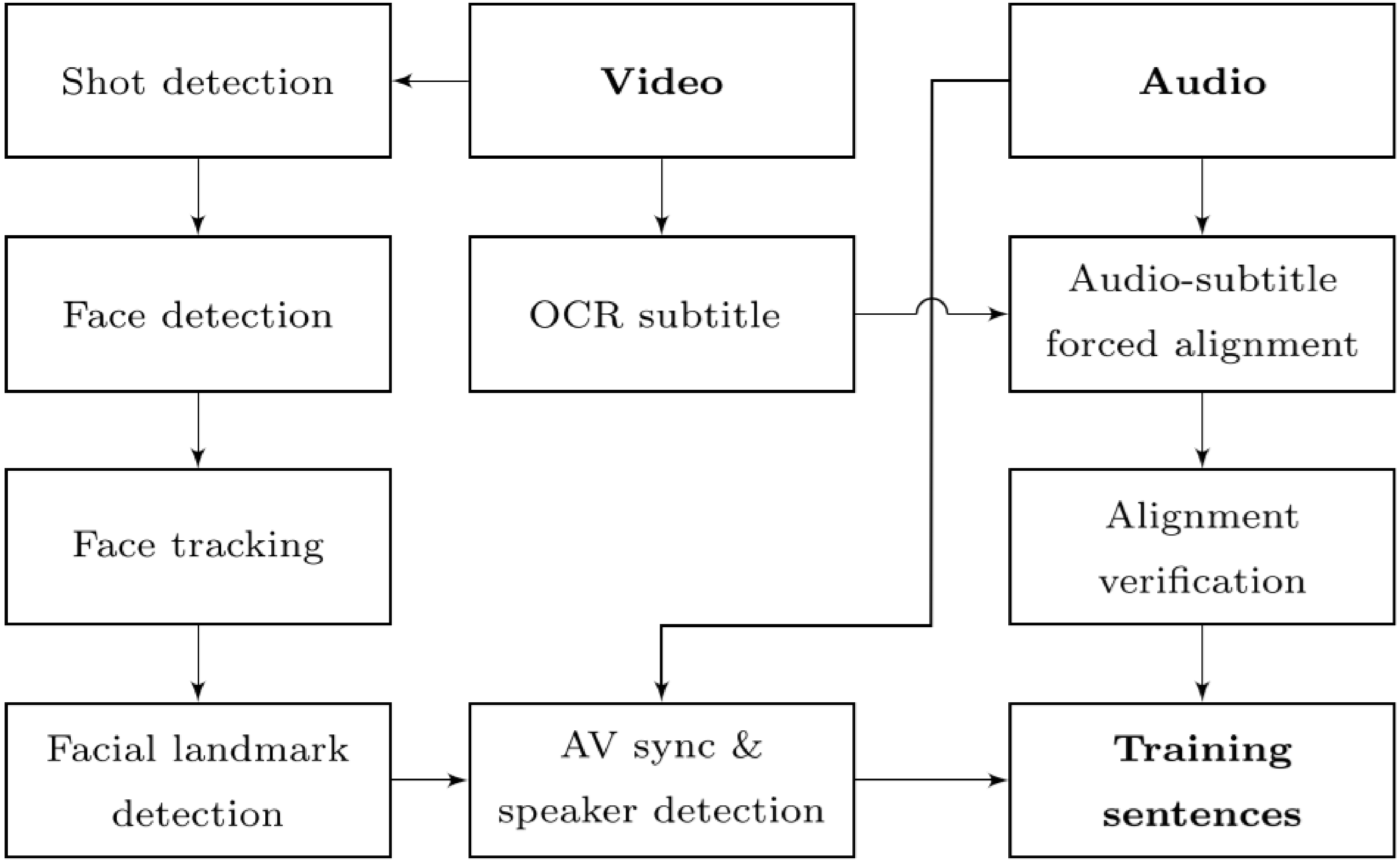
Pipeline of Lip Reading



Datasets

Dataset	Source	Split	Dates	# Spk.	# Utt.	Word inst.	Vocab	# hours
GRID [16]	-	-	-	51	33,000	165k	51	27.5
MODALITY [17]	-	-	-	35	5,880	8,085	182	31
LRW [12]	BBC	Train-val	01/2010 - 12/2015	-	514k	514k	500	165
		Test	01/2016 - 09/2016	-	25k	25k	500	8
LRS [11] †	BBC	Train-val	01/2010 - 02/2016	-	106k	705k	17k	68
		Test	03/2016 - 09/2016	-	12k	77k	6,882	7.5
MV-LRS [14] †	BBC	Pre-train	01/2010 - 12/2015	-	430k	5M	30k	730
		Train-val	01/2010 - 12/2015	-	70k	470k	15k	44.4
		Test	01/2016 - 09/2016	-	4,305	30k	4,311	2.8
LRS2-BBC	BBC	Pre-train	01/2010 - 02/2016	-	96k	2M	41k	195
		Train-val	01/2010 - 02/2016	-	47k	337k	18k	29
		Test	03/2016 - 09/2016	-	1,243	6,663	1,693	0.5
		Text-only	01/2016 - 02/2016	-	8M	26M	60k	-
LRS3-TED	TED & TEDx (YouTube)	Pre-train	-	5,543	132k	4.2M	52k	444
		Train-val	-	4,004	32k	358k	17k	30
		Test	-	451	1,452	11k	2,136	1
		Text-only	-	5,543	1.2M	7.2M	57k	-

LRS dataset pipeline



BLEU: a Method for Automatic Evaluation of Machine Translation
http://www.aclweb.org/anthology/P02-1040.pdf

Evaluation Metrics

- Error Rate
  - PER (phoneme error rate)
  - CER (character error rate)
  - WER (word error rate)
- BLEU (BiLingual Evaluation Understudy)
  - a Method for Automatic Evaluation of Machine Translation

$$ER = \frac{S+D+I}{N}$$

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\mathrm{B{\scriptsize LEU}} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

c is the length of the candidate translation
r is the effective reference corpus length

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

Methods

Method	Year	WER on LRS3	Create Dataset	Train on
LipNet	2016			GRID
WAS	2017		LRS	LRS
TM-seq2seq	2018	58.9	LRS2-BBC, LRS3-TED	MV-LRS, LRS2, LRS3
CTC-V2P	2018	55.1	LSVSR	LSVSR
RNN-T	2019	33.6	YT	YT
VTP	2021	30.7	TEDx_ext	LRS2, LRS3, MV-LRS, TEDX_ext
AVHuBERT	2022	26.9		VoxCeleb2, LRS3

LipNet

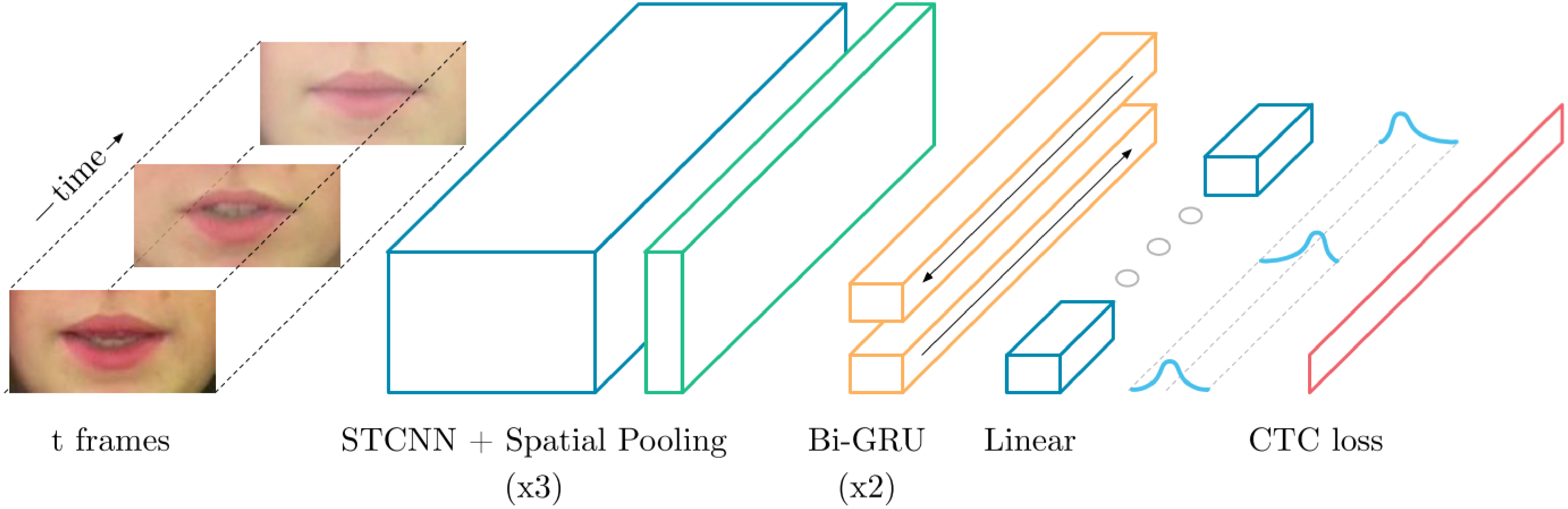


Figure 1: LipNet architecture. A sequence of T frames is used as input, and is processed by 3 layers of STCNN, each followed by a spatial max-pooling layer. The features extracted are processed by 2 Bi-GRUs; each time-step of the GRU output is processed by a linear layer and a softmax. This end-to-end model is trained with CTC.

WLAS

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
R	S	T	U	V	W	X	Y	Z	0	1	2	3	4	5	6	7
8	9	,	.	!	?	:	'	[sos]	[eos]							
[pad]																

Table 10. The output characters

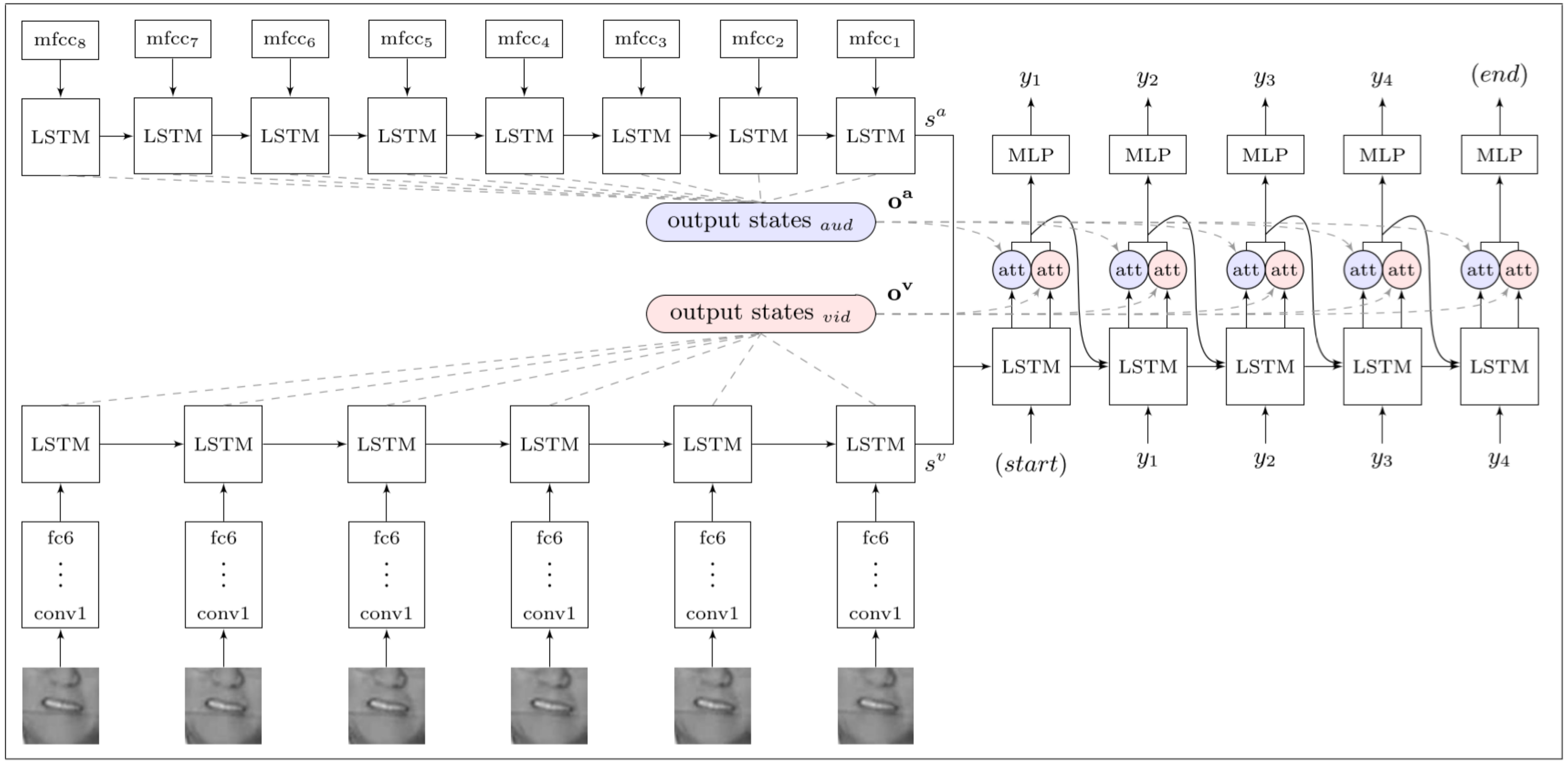


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

$$f_i^v = \mathtt{CNN}(x_i^v) \quad (5)$$

$$h_i^v, o_i^v = \mathtt{LSTM}(f_i^v, h_{i+1}^v) \quad (6)$$

$$s^v = h_1^v \quad (7)$$

$$h_j^a, o_j^a = \mathtt{LSTM}(x_j^a, h_{j+1}^a) \quad (8)$$

$$s^a = h_1^a \quad (9)$$

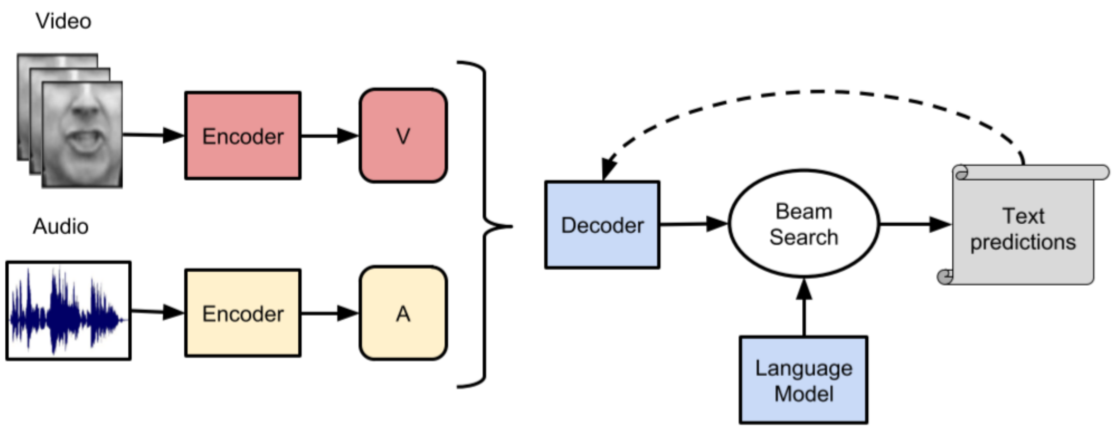
$$h_k^d, o_k^d = \mathtt{LSTM}(h_{k-1}^d, y_{k-1}, c_{k-1}^v, c_{k-1}^a) \quad (10)$$

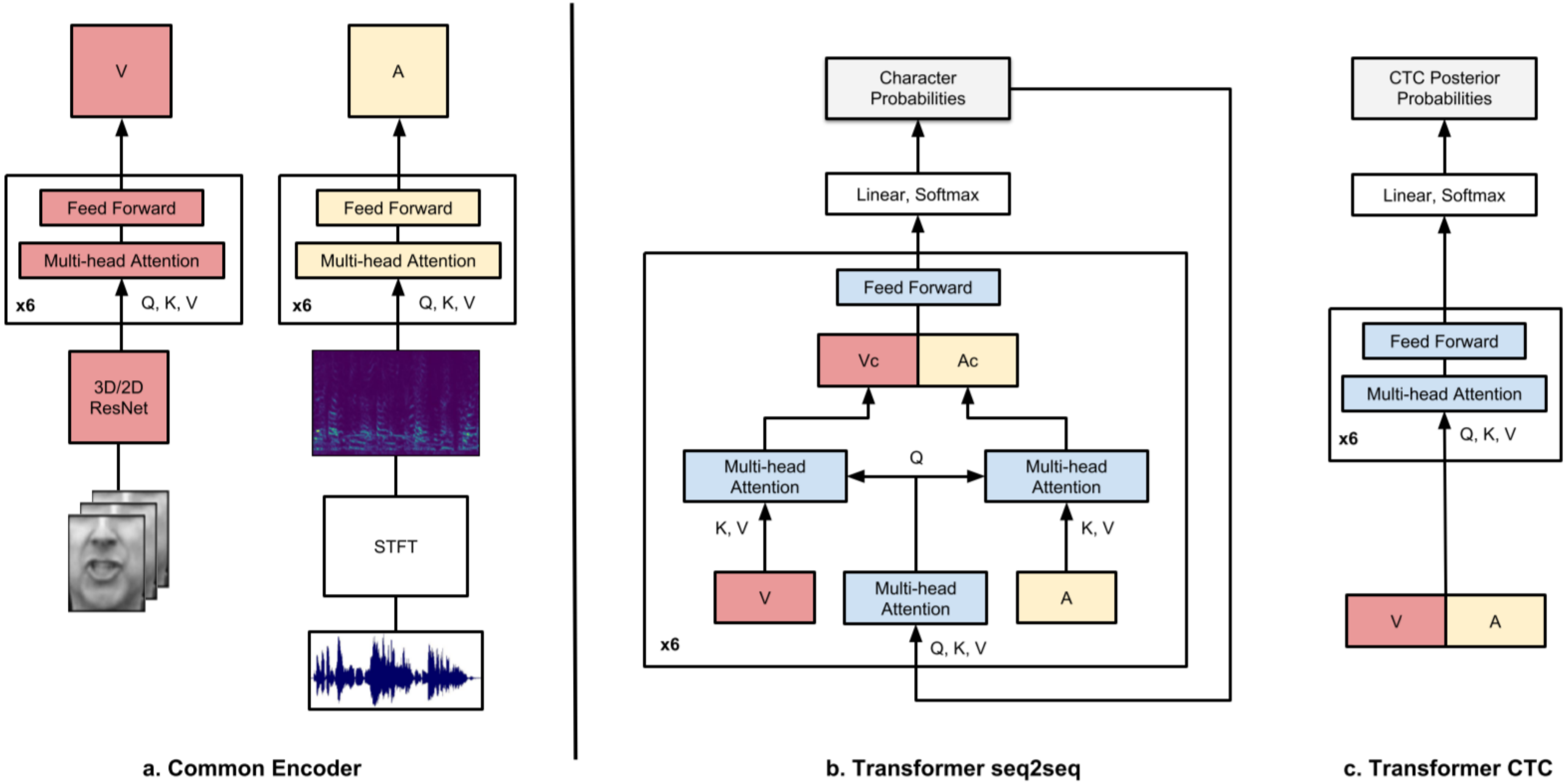
$$c_k^v = \mathbf{o}^v \cdot \mathtt{Attention}^{\mathtt{v}}(h_k^d, \mathbf{o}^v) \quad (11)$$

$$c_k^a = \mathbf{o}^a \cdot \mathtt{Attention}^{\mathtt{a}}(h_k^d, \mathbf{o}^a) \quad (12)$$

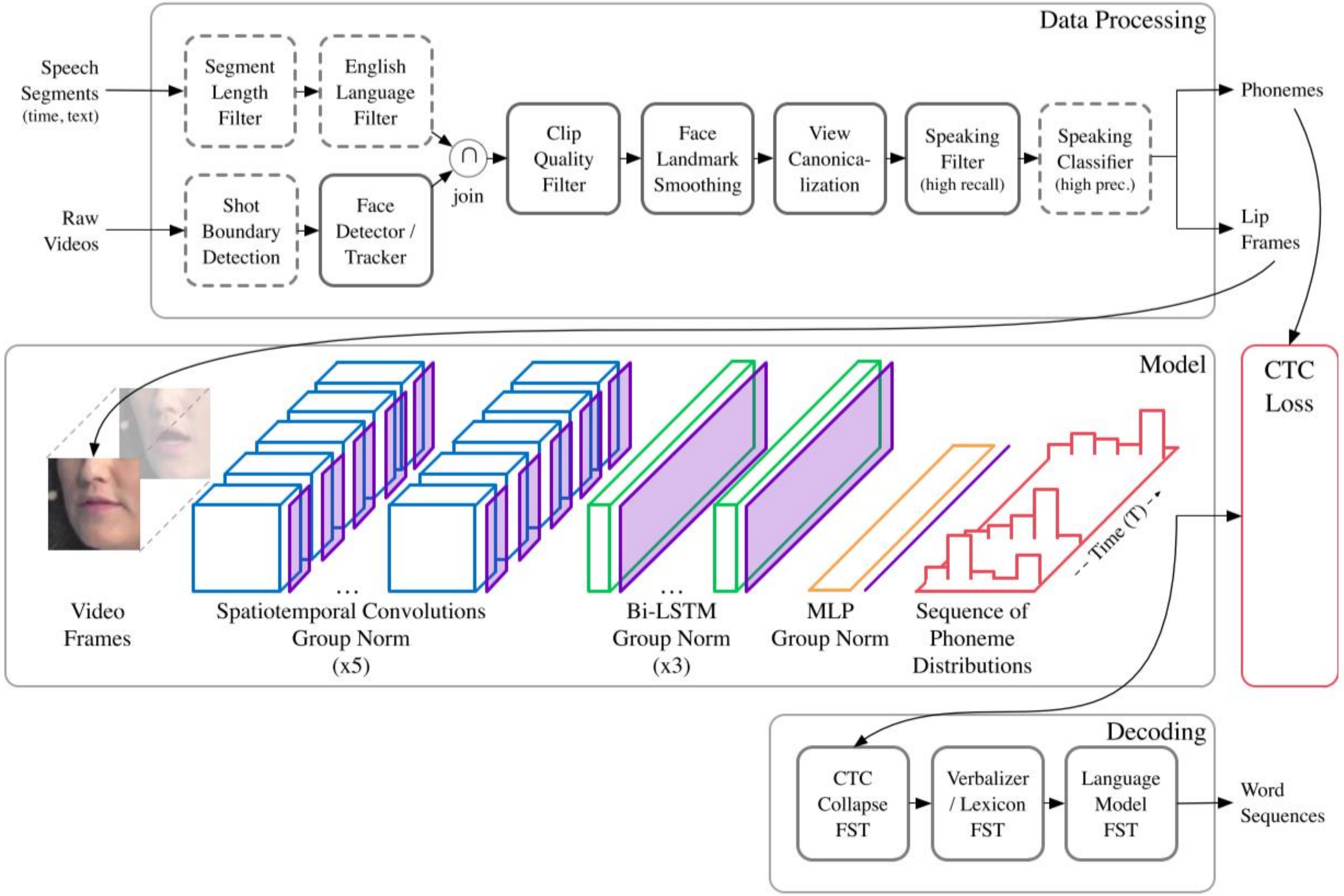
$$P(y_i|\mathbf{x}^v, \mathbf{x}^a, y_{<i}) = \mathtt{softmax}(\mathtt{MLP}(o_k^d, c_k^v, c_k^a)) \quad (13)$$

TM-seq2seq


Video
Encoder
V
Audio
Encoder
A
Decoder
Beam Search
Text predictions
Language Model

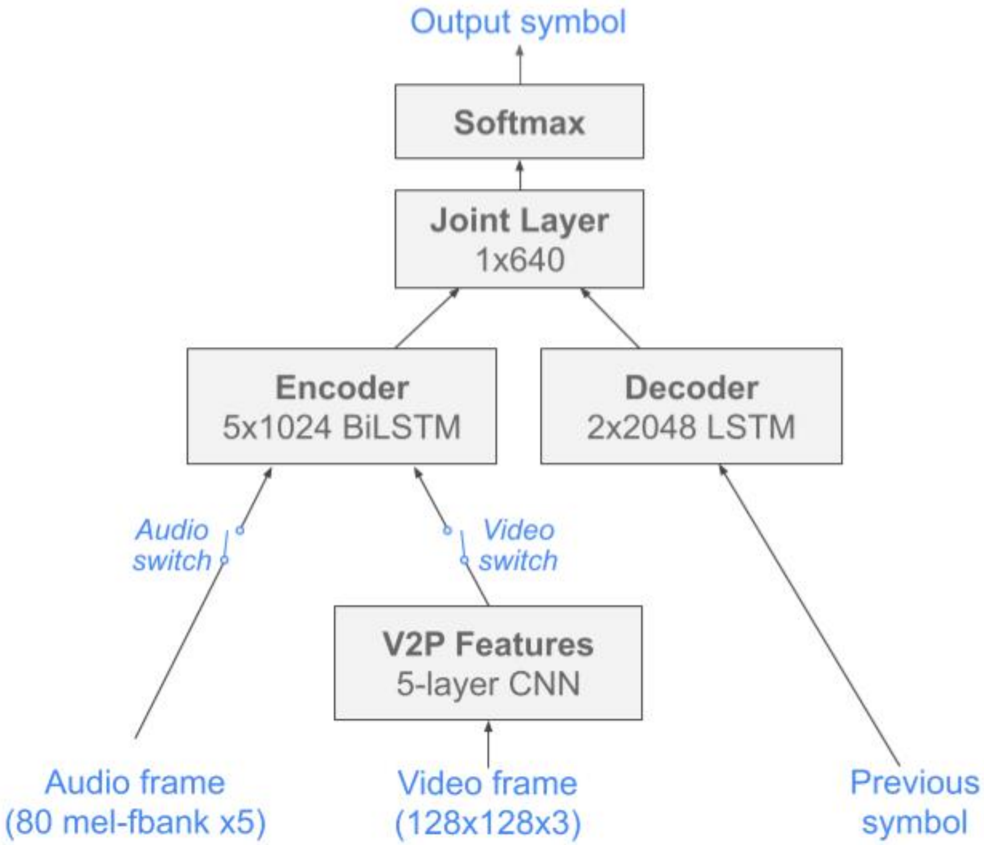

V
Feed Forward
Multi-head Attention
x6
Q, K, V
3D/2D ResNet
A
Feed Forward
Multi-head Attention
x6
Q, K, V
STFT
Character Probabilities
Linear, Softmax
Feed Forward
Vc
Ac
Multi-head Attention
Q
Multi-head Attention
K, V
K, V
V
Multi-head Attention
A
x6
Q, K, V
CTC Posterior Probabilities
Linear, Softmax
Feed Forward
Multi-head Attention
x6
Q, K, V
V
A
a. Common Encoder
b. Transformer seq2seq
c. Transformer CTC

CTC-V2P



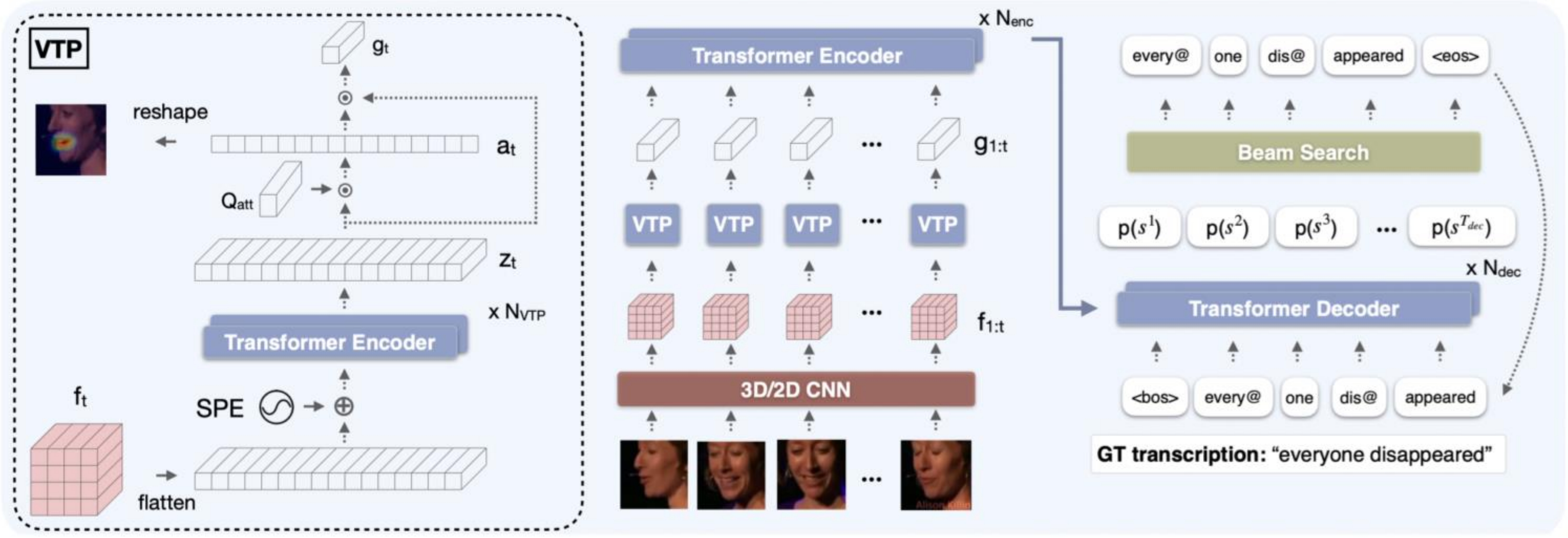
Data Processing
Speech Segments (time, text)
Segment Length Filter
English Language Filter
Raw Videos
Shot Boundary Detection
Face Detector / Tracker
∩
join
Clip Quality Filter
Face Landmark Smoothing
View Canonica-lization
Speaking Filter (high recall)
Speaking Classifier (high prec.)
Phonemes
Lip Frames
Model
Video Frames
Spatiotemporal Convolutions Group Norm (x5)
Bi-LSTM Group Norm (x3)
MLP Group Norm
Sequence of Phoneme Distributions
Time (T)
CTC Loss
Decoding
CTC Collapse FST
Verbalizer / Lexicon FST
Language Model FST
Word Sequences

RNN-T



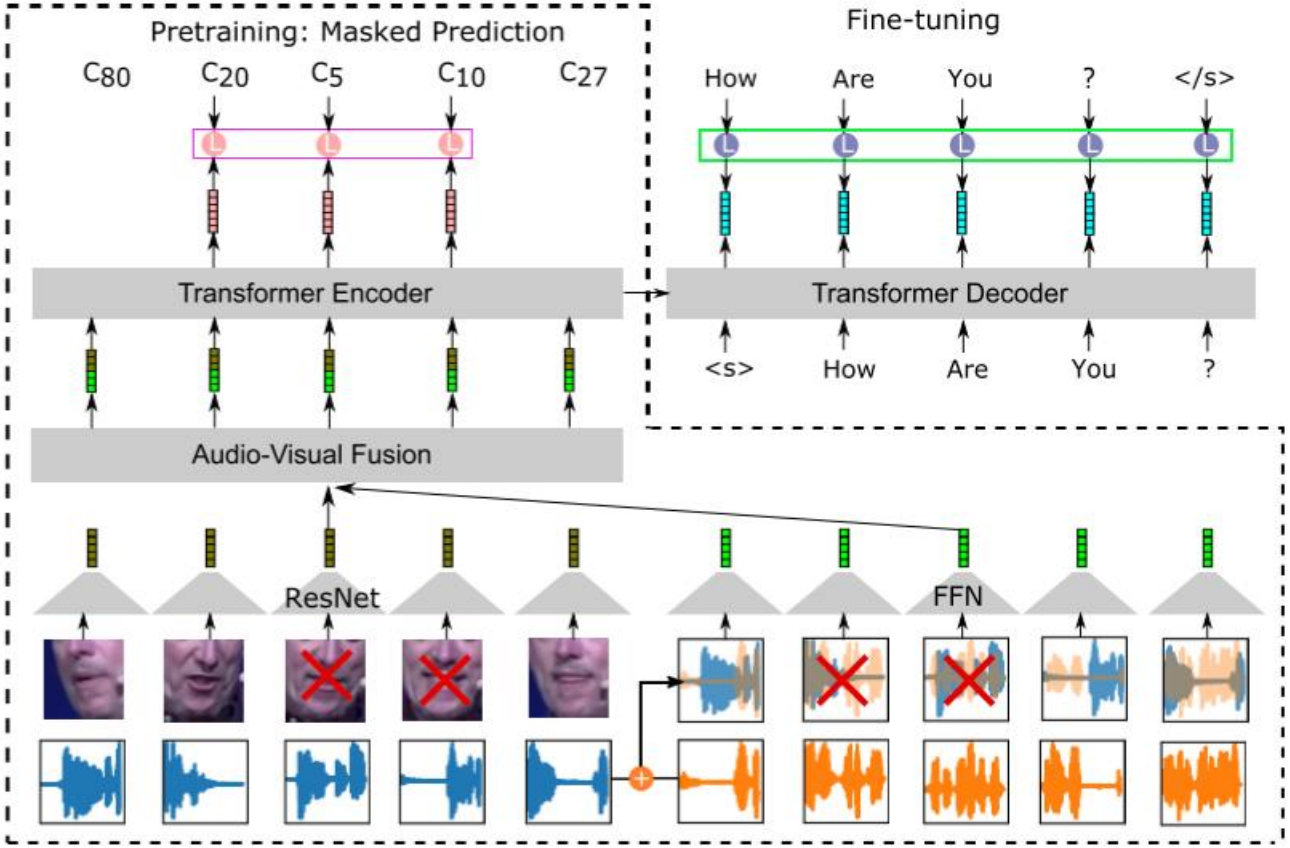
Output symbol
Softmax
Joint Layer
1x640
Encoder
5x1024 BiLSTM
Decoder
2x2048 LSTM
Audio
switch
Video
switch
V2P Features
5-layer CNN
Audio frame
(80 mel-fbank x5)
Video frame
(128x128x3)
Previous
symbol

VTP



VTP
g_t
reshape
a_t
Q_att
z_t
x N_VTP
Transformer Encoder
f_t
SPE
flatten
x N_enc
Transformer Encoder
g_1:t
VTP
f_1:t
3D/2D CNN
every@
one
dis@
appeared
<eos>
Beam Search
p(s^1)
p(s^2)
p(s^3)
p(s^T_dec)
x N_dec
Transformer Decoder
<bos>
GT transcription: "everyone disappeared"

AV-HuBERT


Pretraining: Masked Prediction
Fine-tuning
C80
C20
C5
C10
C27
How
Are
You
?
</s>
Transformer Encoder
Transformer Decoder
<s>
How
Are
You
?
Audio-Visual Fusion
ResNet
FFN

Conclusion

- It is critical to develop a better way to capture both spatial and temporal information
- It is (used to be) important to build a dataset
  - a pipeline for pre-process is needed
  - language variation, vocabulary variation, ...
  - different video quality
  - bigger dataset
- Training protocol is important
  - curriculum learning
  - pre-training
  - avoid single modality dominates
  - ...

Reference

Method	Year	Paper Title	PDF
LipNet	2016	LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING	https://arxiv.org/abs/1611.05358v2
WAS	2017	Lip Reading Sentences in the Wild	https://arxiv.org/abs/1611.05358v2
TM-seq2seq	2018	Deep Audio-visual Speech Recognition	https://arxiv.org/abs/1809.02108
CTC-V2P	2018	LARGE-SCALE VISUAL SPEECH RECOGNITION	https://arxiv.org/abs/1807.05162
RNN-T	2019	RECURRENT NEURAL NETWORK TRANSDUCER FOR AUDIO-VISUAL SPEECH RECOGNITION	https://arxiv.org/abs/1911.04890
VTP	2021	Sub-word Level Lip Reading With Visual Attention	https://arxiv.org/abs/2110.07603v2
AVHuBERT	2022	Robust Self-Supervised Audio-Visual Speech Recognition	https://arxiv.org/abs/2201.01763