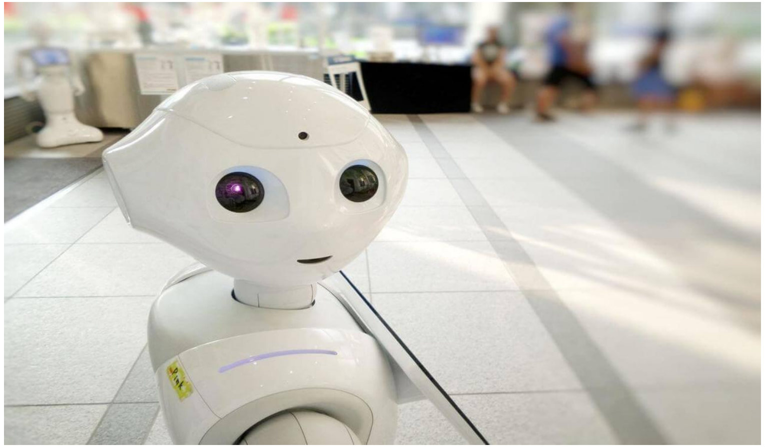


Streaming Audio-Visual Perception

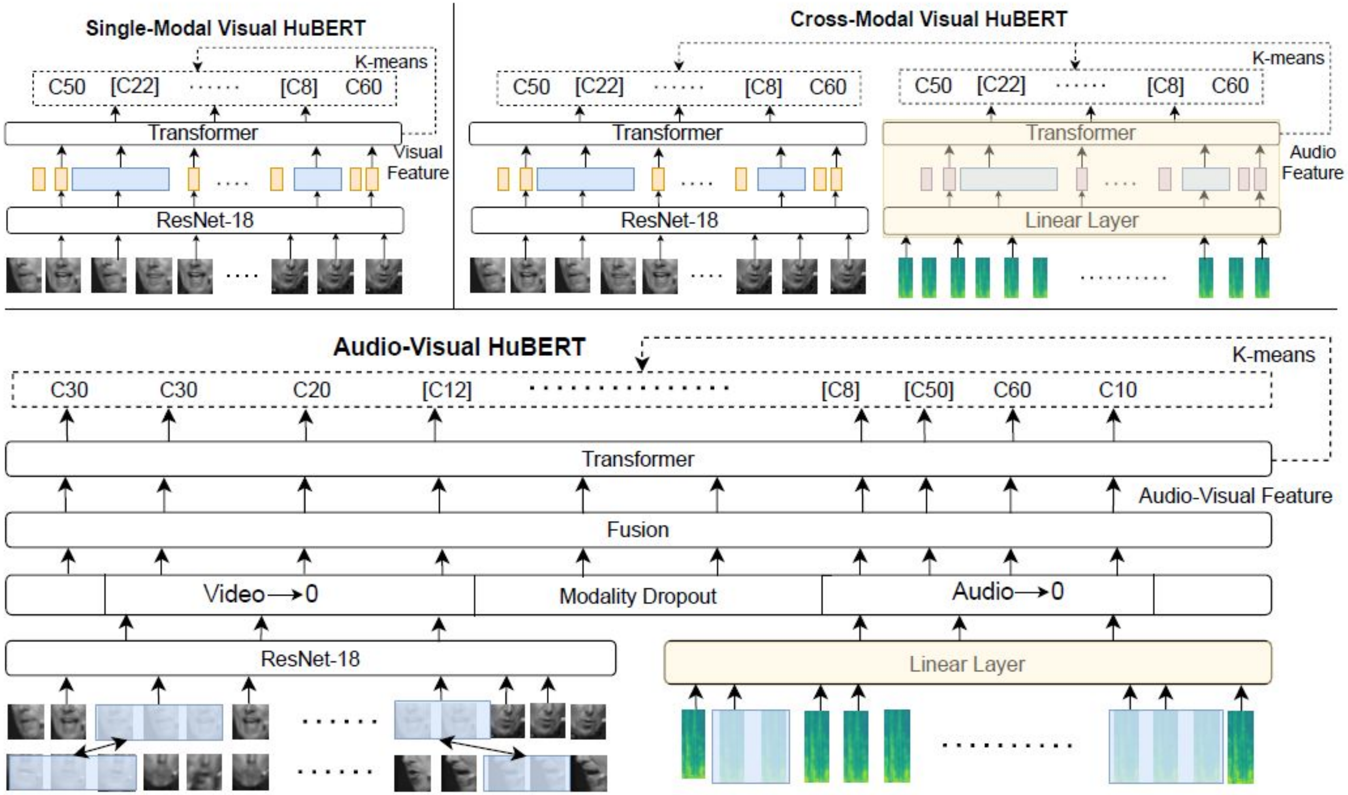
Karthik Ganesan

Why streaming ?

We need embodied agents to understand in real-time

OFF

AV-HuBERT


Single-Modal Visual HuBERT
K-means
C50 [C22] ...... [C8] C60
Transformer
Visual Feature
ResNet-18
Cross-Modal Visual HuBERT
K-means
C50 [C22] ...... [C8] C60
Transformer
ResNet-18
C50 [C22] ...... [C8] C60
Transformer
Audio Feature
Linear Layer
Audio-Visual HuBERT
K-means
C30 C30 C20 [C12] .... [C8] [C50] C60 C10
Transformer
Audio-Visual Feature
Fusion
Video⟶0
Modality Dropout
Audio⟶0
ResNet-18
Linear Layer

Drawbacks of AV-Hubert

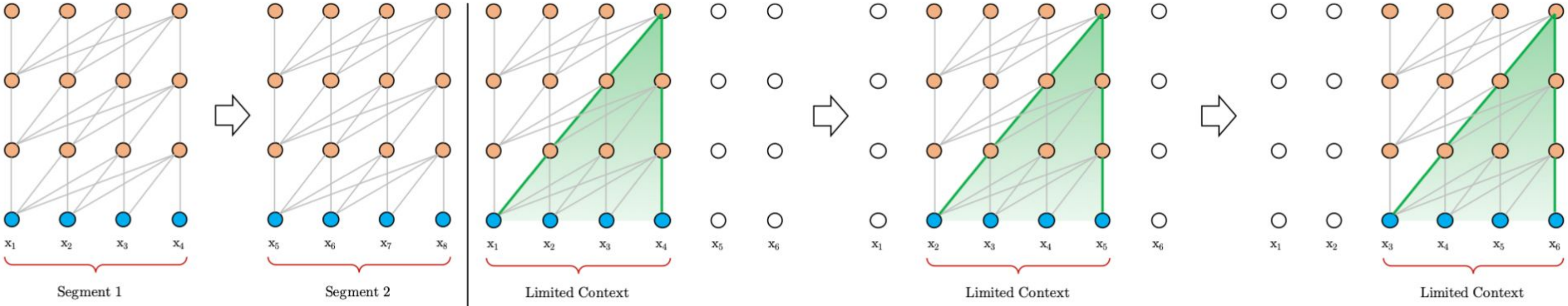
1. We need bidirectional context , thus we need to wait until entire input is provided
2. Chunking is also non-trivial as end-point detection also needs supervision to train

How did we solve this problem in the uni-modal speech recognition ?

STREAMING TRANSFORMER ASR WITH BLOCKWISE SYNCHRONOUS BEAM SEARCH

Emiru Tsunoo[1], *Yosuke Kashiwagi*[1], *Shinji Watanabe*[2]

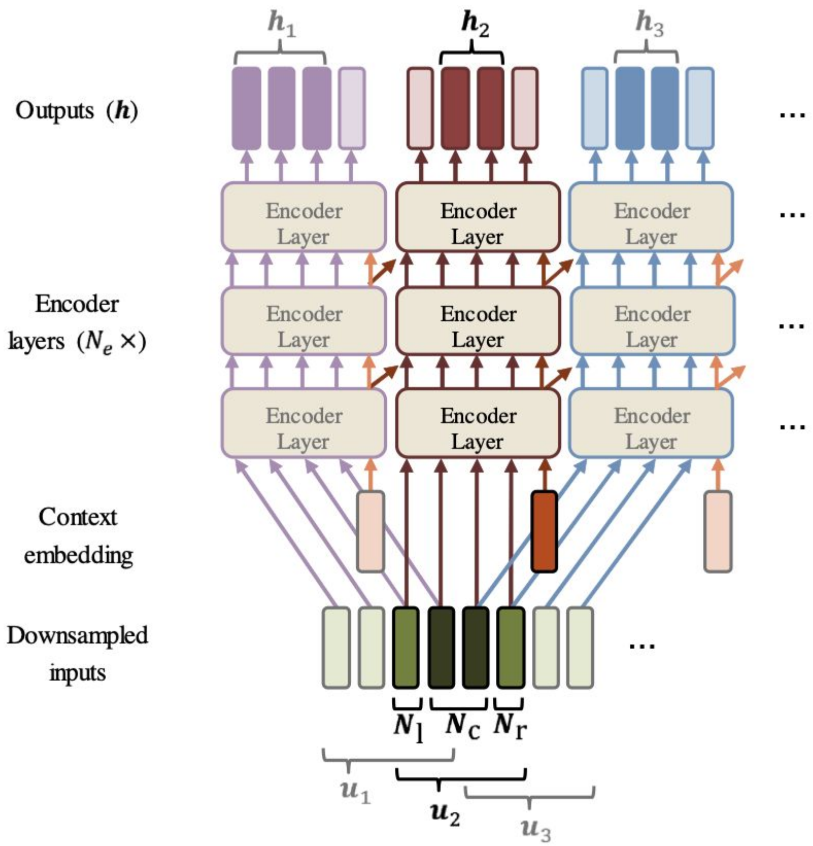
[1]Sony Corporation, Japan
[2]Johns Hopkins University, USA



(a) Train phase.

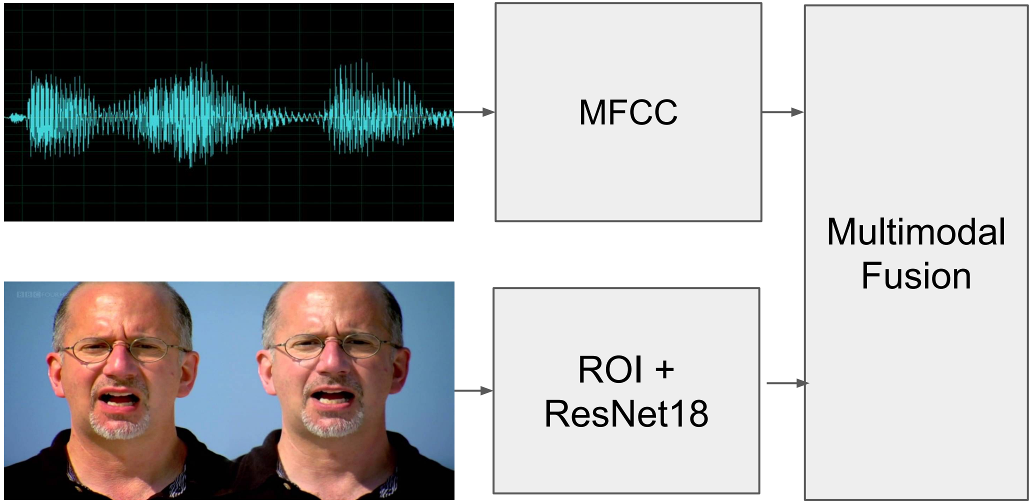
(b) Evaluation phase.

STREAMING TRANSFORMER ASR WITH BLOCK-WISE SYNCHRONOUS BEAM SEARCH



How do we make this approach multimodal?

Proposed Architecture

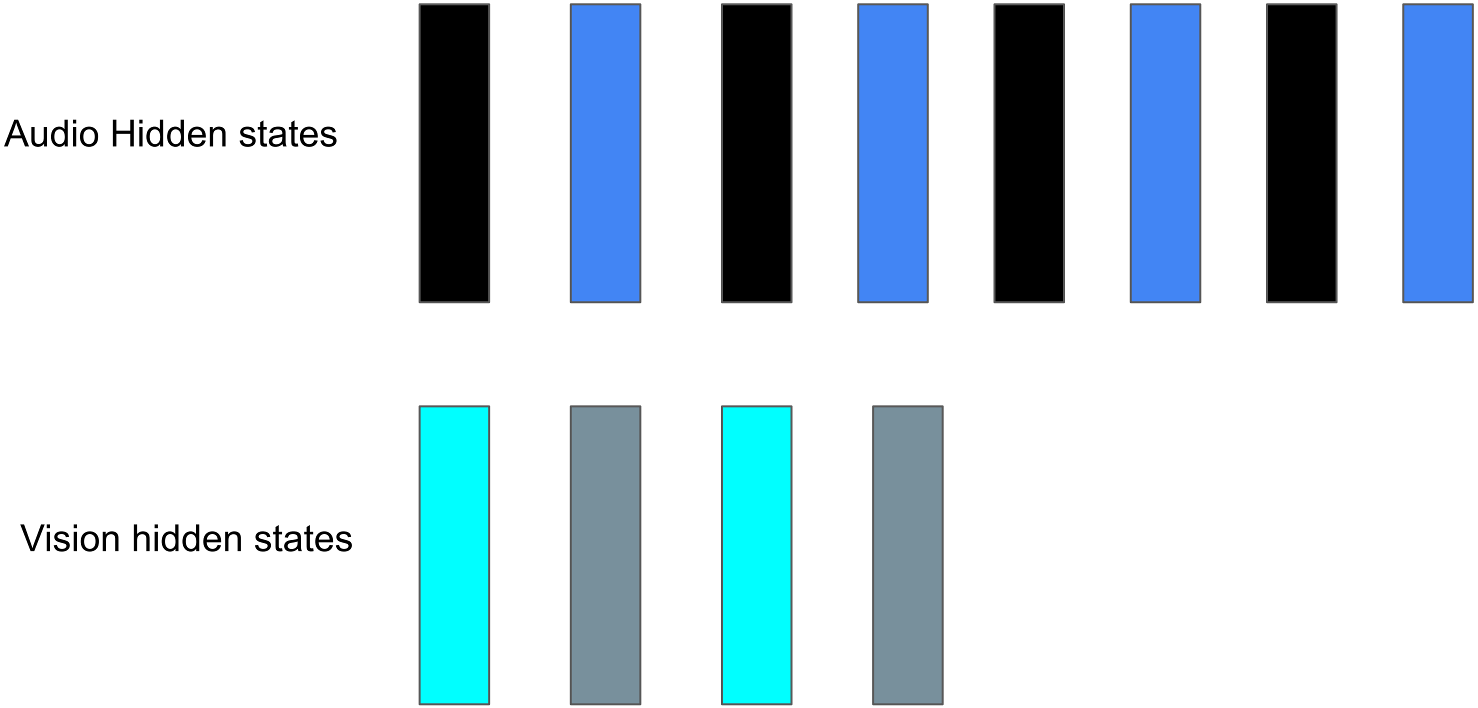


MFCC

ROI +
ResNet18

Multimodal
Fusion

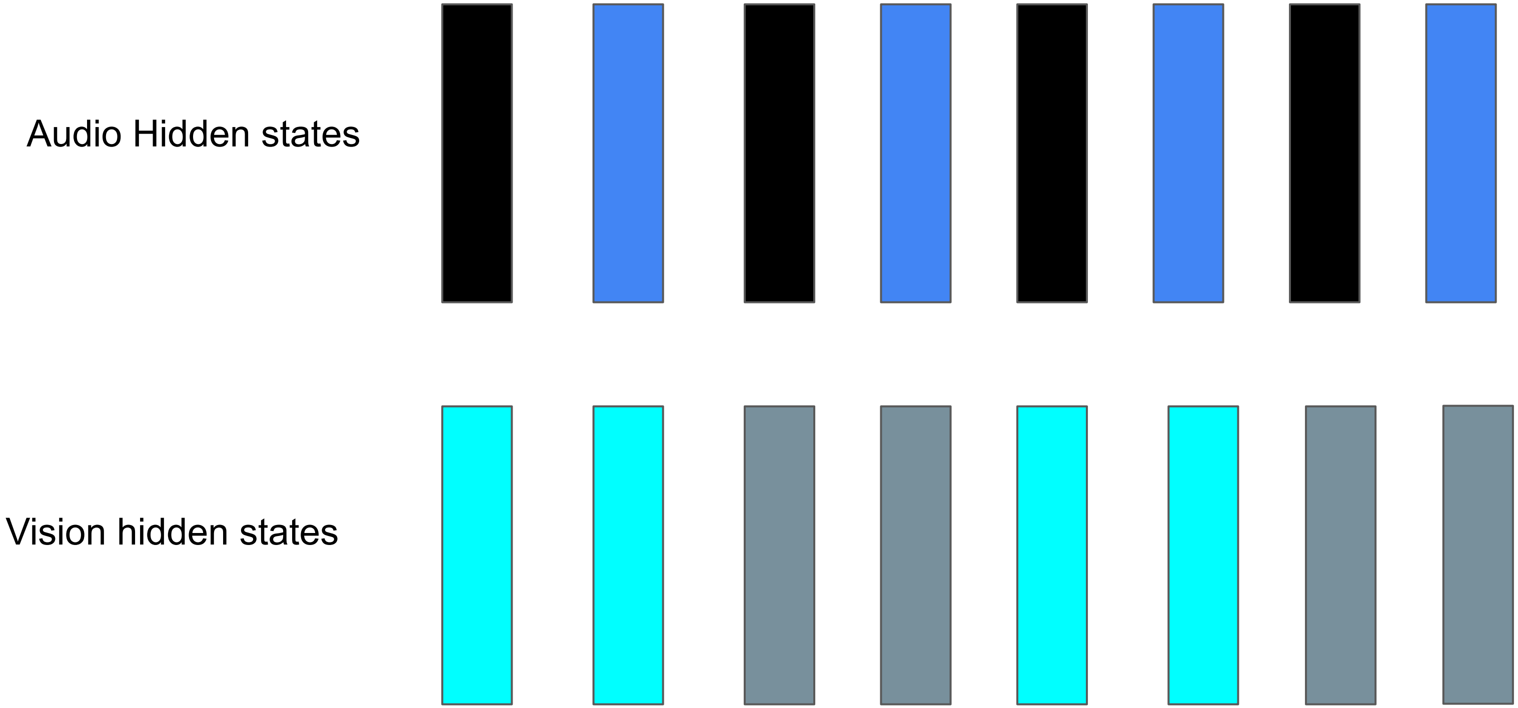
Multimodal fusion



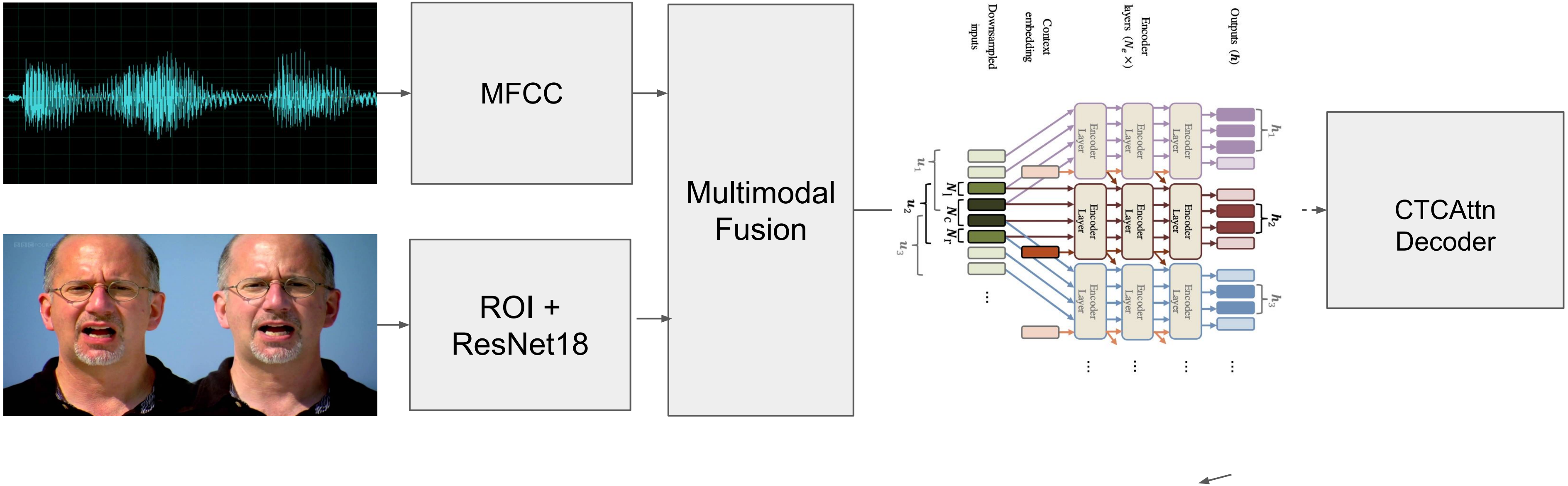
TORCH.REPEAT_INTERLEAVE

```
>>> x = torch.tensor([1, 2, 3])
>>> x.repeat_interleave(2)
tensor([1, 1, 2, 2, 3, 3])
```

Multimodal fusion



Proposed Architecture



MFCC

ROI + ResNet18

Multimodal Fusion

CTCAttn Decoder

Results

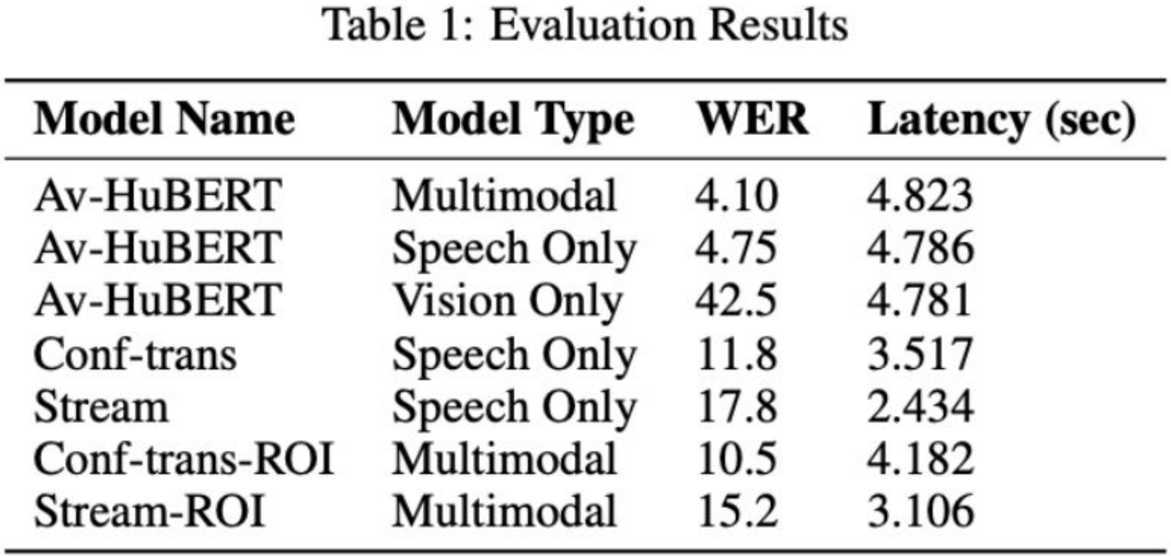


Table 1: Evaluation Results

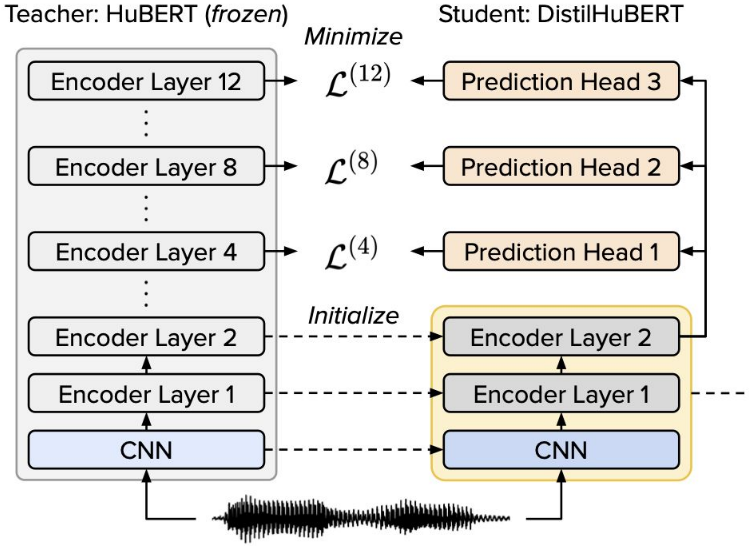
Model Name	Model Type	WER	Latency (sec)
Av-HuBERT	Multimodal	4.10	4.823
Av-HuBERT	Speech Only	4.75	4.786
Av-HuBERT	Vision Only	42.5	4.781
Conf-trans	Speech Only	11.8	3.517
Stream	Speech Only	17.8	2.434
Conf-trans-ROI	Multimodal	10.5	4.182
Stream-ROI	Multimodal	15.2	3.106

In progress research direction

DISTILHUBERT: SPEECH REPRESENTATION LEARNING BY LAYER-WISE DISTILLATION OF HIDDEN-UNIT BERT

Heng-Jui Chang, Shu-wen Yang, Hung-yi Lee

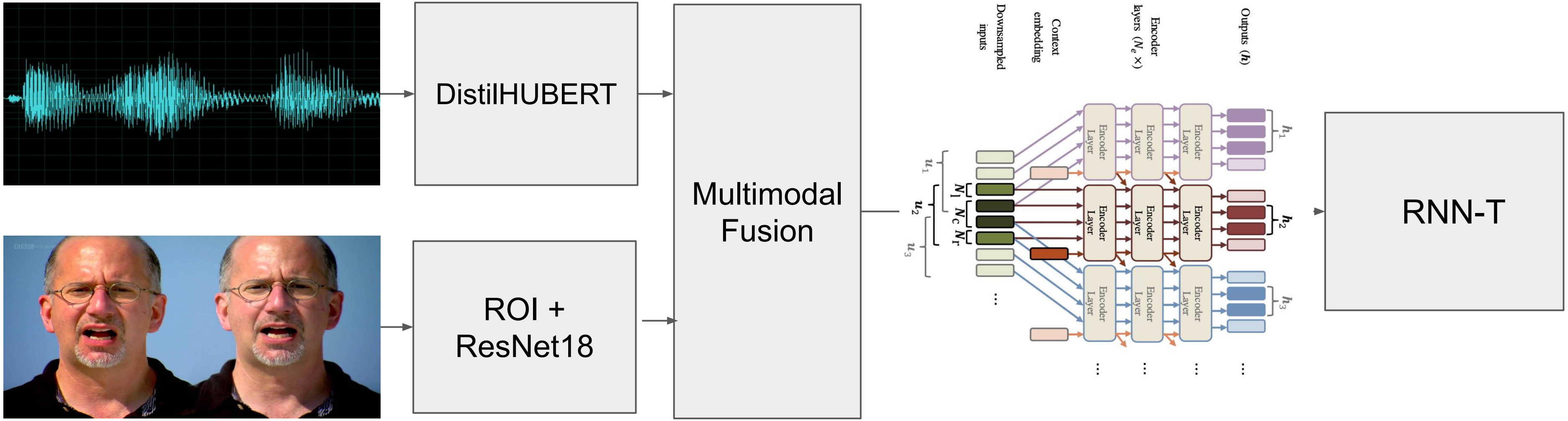
College of Electrical Engineering and Computer Science, National Taiwan University



(I) Pre-training

$$\begin{aligned}\mathcal{L}^{(l)} &= \mathcal{L}_{\ell 1}^{(l)} + \lambda \mathcal{L}_{\cos}^{(l)} \\ &= \sum_{t=1}^{T}\left[\frac{1}{D}\left\|\boldsymbol{h}_t^{(l)} - \hat{\boldsymbol{h}}_t^{(l)}\right\|_1 - \lambda \log\sigma\left(\cos\left(\boldsymbol{h}_t^{(l)}, \hat{\boldsymbol{h}}_t^{(l)}\right)\right)\right],\end{aligned} \tag{1}$$

Proposed Architecture



DistilHUBERT

ROI + ResNet18

Multimodal Fusion

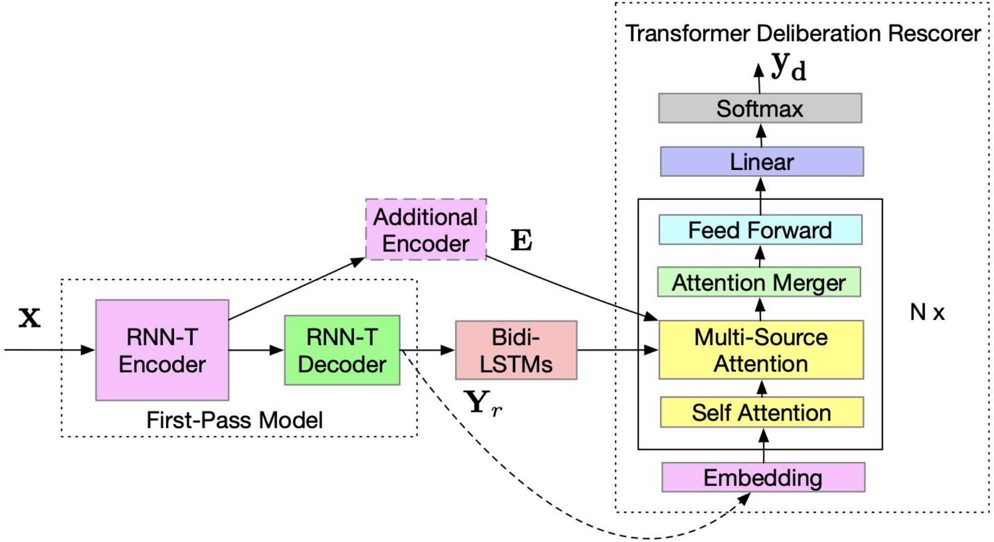
RNN-T

TRANSFORMER BASED DELIBERATION FOR TWO-PASS SPEECH RECOGNITION

Ke Hu, Ruoming Pang, Tara N. Sainath, Trevor Strohman

Google, Inc., USA

{huk,rpang,tsainath,strohman}@google.com

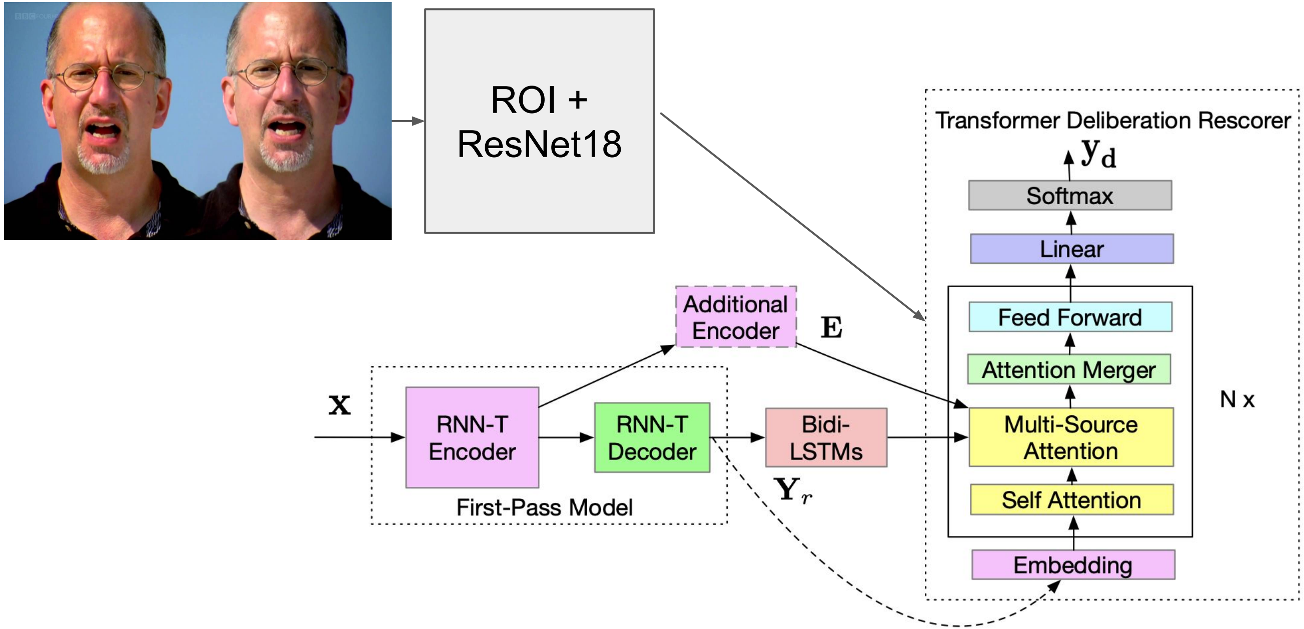


TRANSFORMER BASED DELIBERATION FOR TWO-PASS SPEECH RECOGNITION

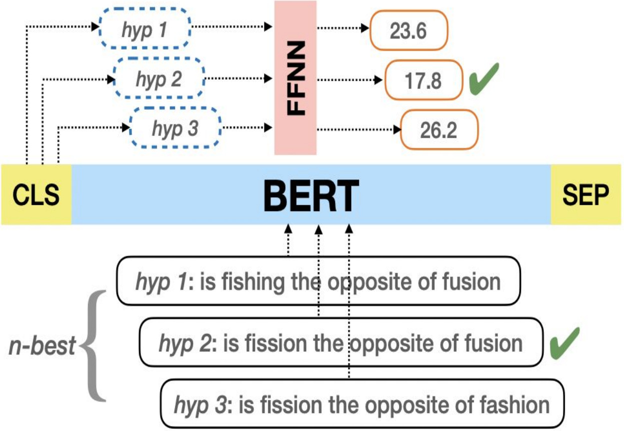
Ke Hu, Ruoming Pang, Tara N. Sainath, Trevor Strohman

Google, Inc., USA

{huk,rpang,tsainath,strohman}@google.com



MWER loss based rescoring


hyp 1
hyp 2
hyp 3
FFNN
23.6
17.8 ✓
26.2
CLS
BERT
SEP
n-best
hyp 1: is fishing the opposite of fusion
hyp 2: is fission the opposite of fusion ✓
hyp 3: is fission the opposite of fashion

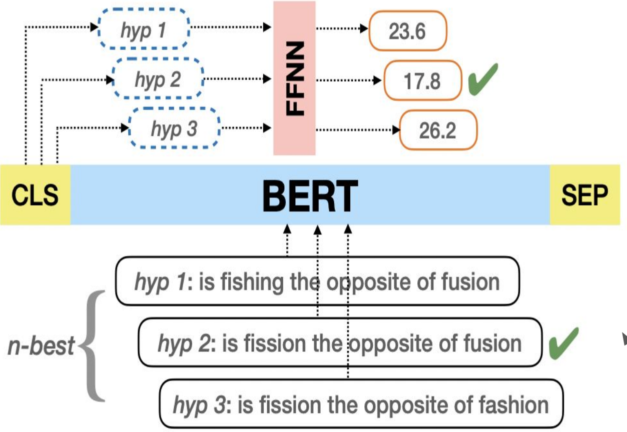
+

Audio cross attention + some secret sauce

$$P_i = \frac{e^{-s_i}}{\sum_{j=1}^{n} e^{-s_j}}$$

$$\bar{\epsilon}_H = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i$$

$$\mathcal{L}_{\text{MWER}} = \sum_{i=1}^{n} P_i \cdot (\epsilon_i - \bar{\epsilon}_H).$$


hyp 1
hyp 2
hyp 3
FFNN
23.6
17.8 ✓
26.2
CLS
BERT
SEP
n-best
hyp 1: is fishing the opposite of fusion
hyp 2: is fission the opposite of fusion ✓
hyp 3: is fission the opposite of fashion

+

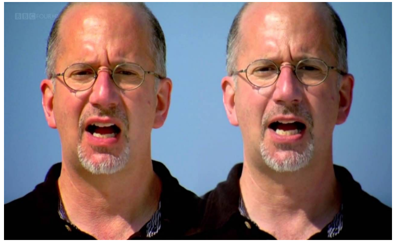
Audio cross attention + some secret sauce

$$P_i = \frac{e^{-s_i}}{\sum_{j=1}^{n} e^{-s_j}}$$

$$\bar{\epsilon}_H = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i$$

$$\mathcal{L}_{\text{MWER}} = \sum_{i=1}^{n} P_i \cdot (\epsilon_i - \bar{\epsilon}_H).$$

ROI + ResNet18



Thank you for your multimodal streaming synchronous attention :)