國立臺灣大學
National Taiwan University

Mockingjay:
Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders

Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, Hung-yi Lee

To appear in ICASSP 2020

2020/02/07

Introduction

NLP BERT: language representation learning:

BERT is a language representation model, it can be fine-tuned with downstream NLP models (QA model, summarization model, etc) to create SOTA results.

Inspired by the state-of-the-art BERT in NLP, we aim the build a speech version BERT.

Speech BERT: speech representation learning:

find a transform from speech that makes high-level information
more accessible to SLP (Speech and Language Processing) downstream tasks.

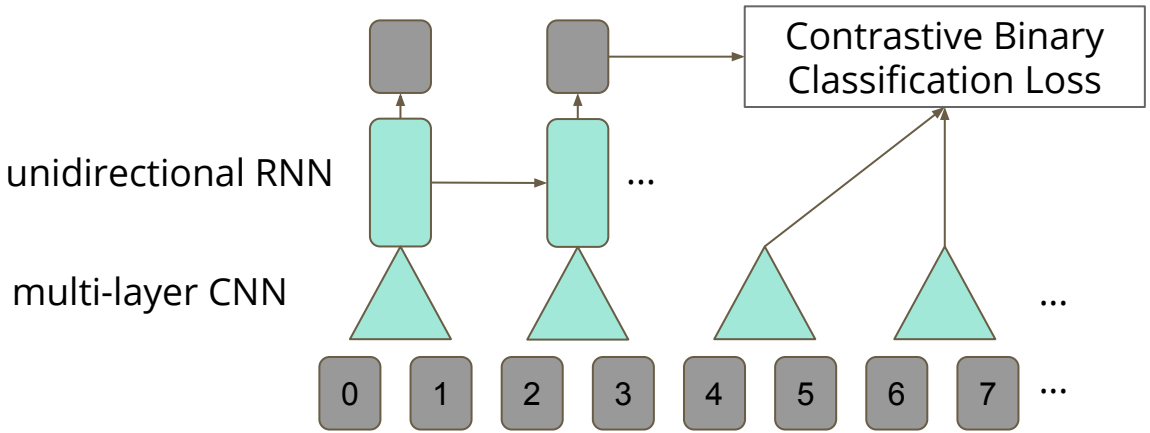
(e.g. phone classification, ASR, VC, speech-translation)

A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

CPC

Representation Learning with Contrastive Predictive Coding



A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

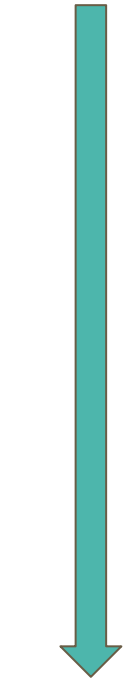
CPC

Representation Learning with Contrastive Predictive Coding

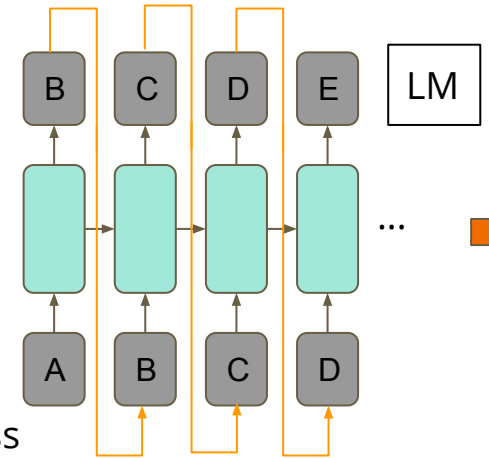
April, 2019
MIT

APC

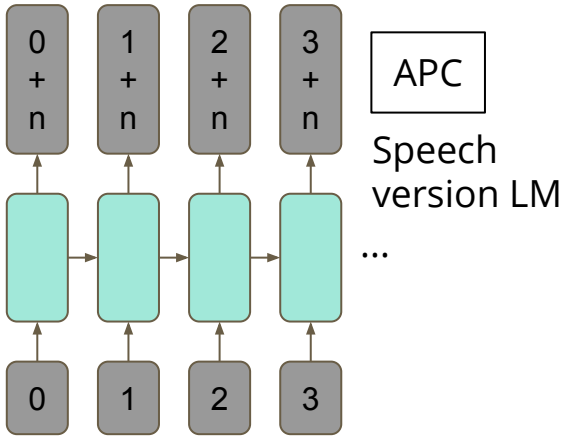
An Unsupervised Autoregressive Model for Speech Representation Learning



Cross-Entropy Loss



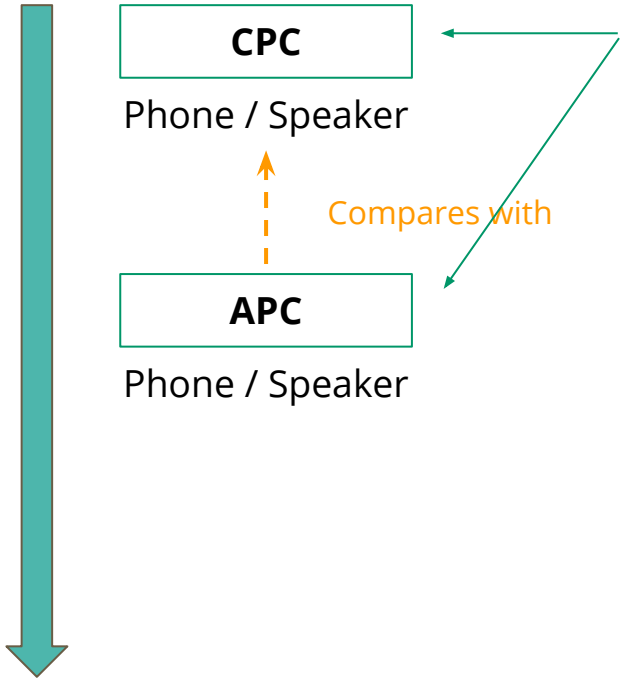
Reconstruction Loss



A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

April, 2019
MIT



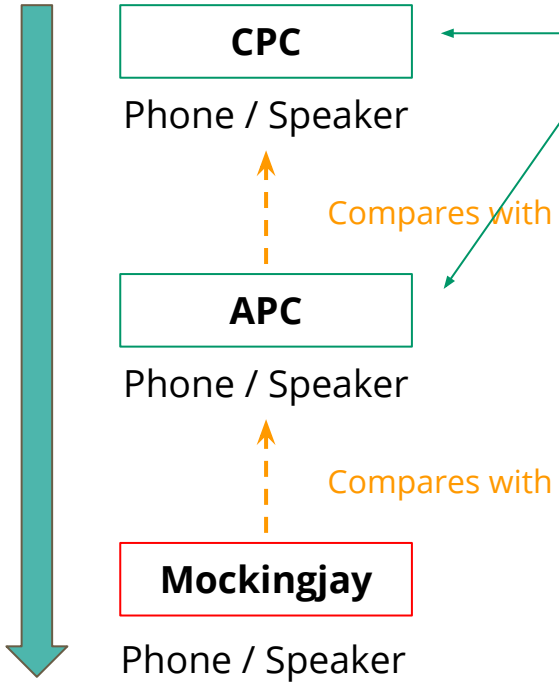
Previous speech representation methods:
encode past information and predict information about future frames.

A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

April, 2019
MIT

October, 2019
Ours



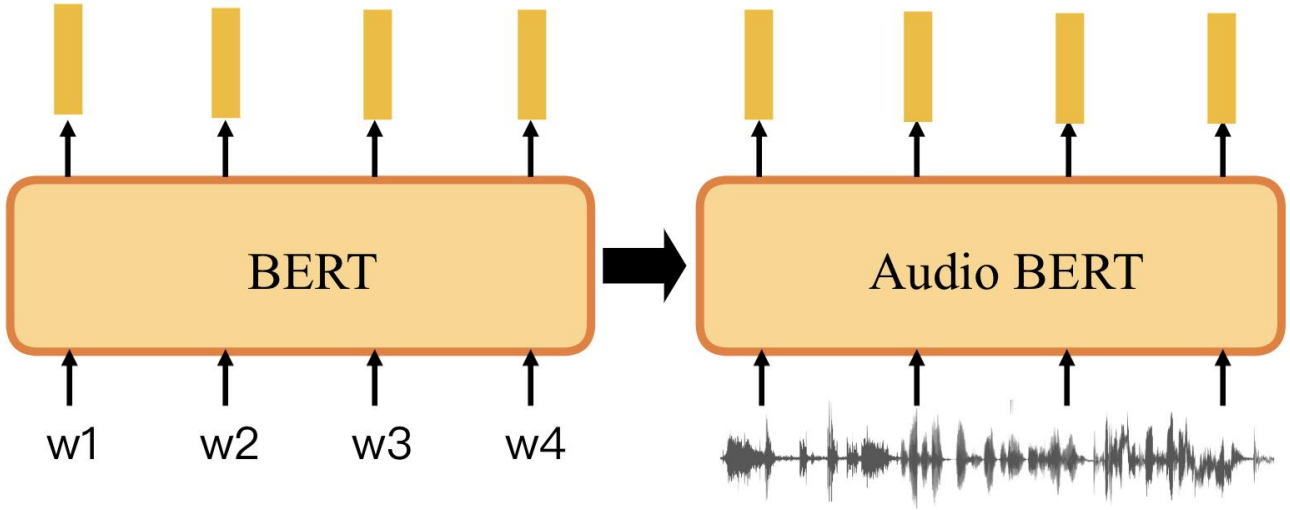
Previous speech representation methods:
encode past information and predicting information about future frames.

Mockingjay:
Predicts the current frame by jointly conditioning on both past and future contexts.

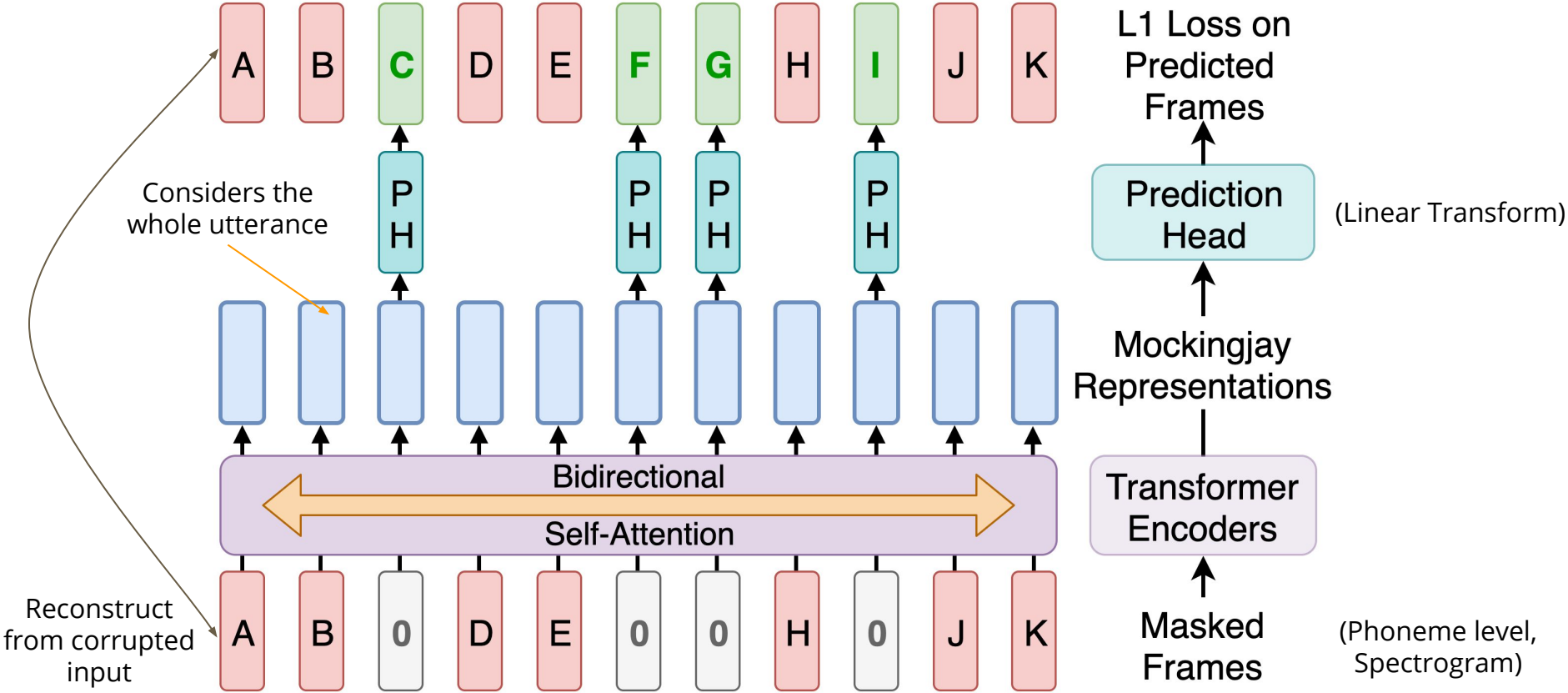
Can be seen as a Speech version of BERT / RoBERTa

The Proposed Method

1. Framework
2. Pre-training task
3. Evaluation：
   a. Phone classification
   b. Speaker Discrimination
   c. ASR（LibriSpeech & TIMIT）



The Proposed Framework

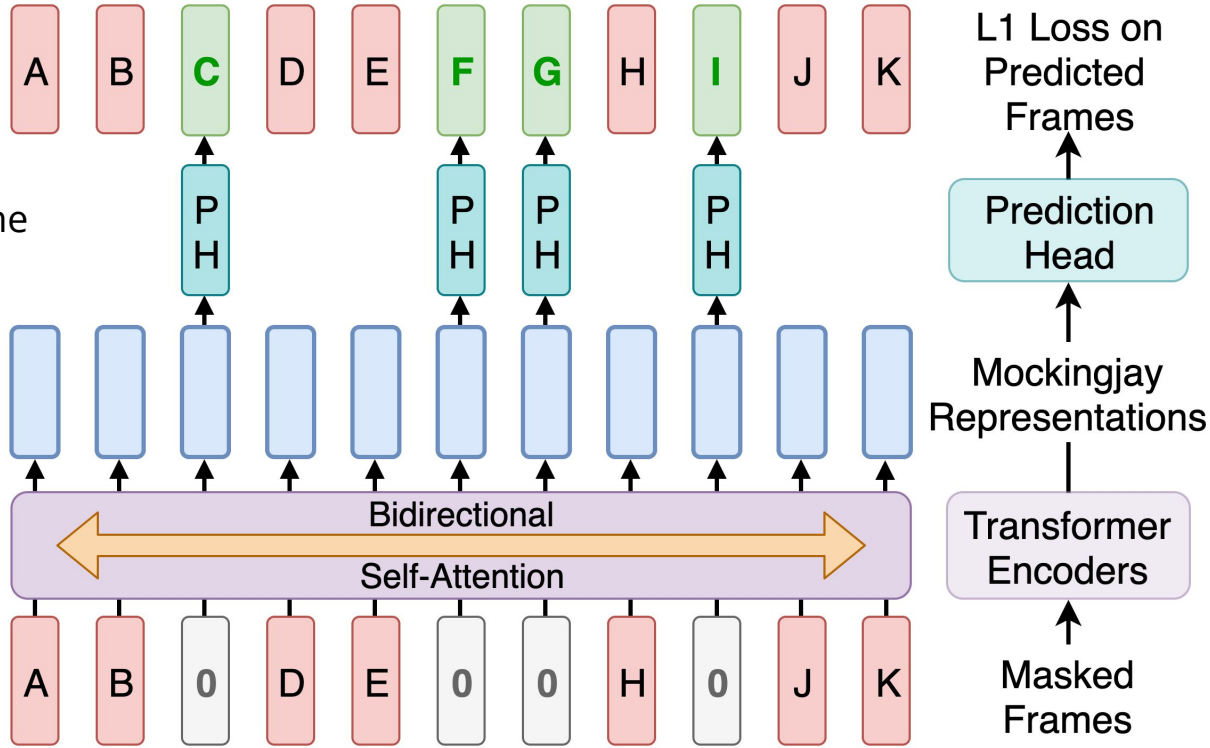


Pre-Training Task: Masked Acoustic Model

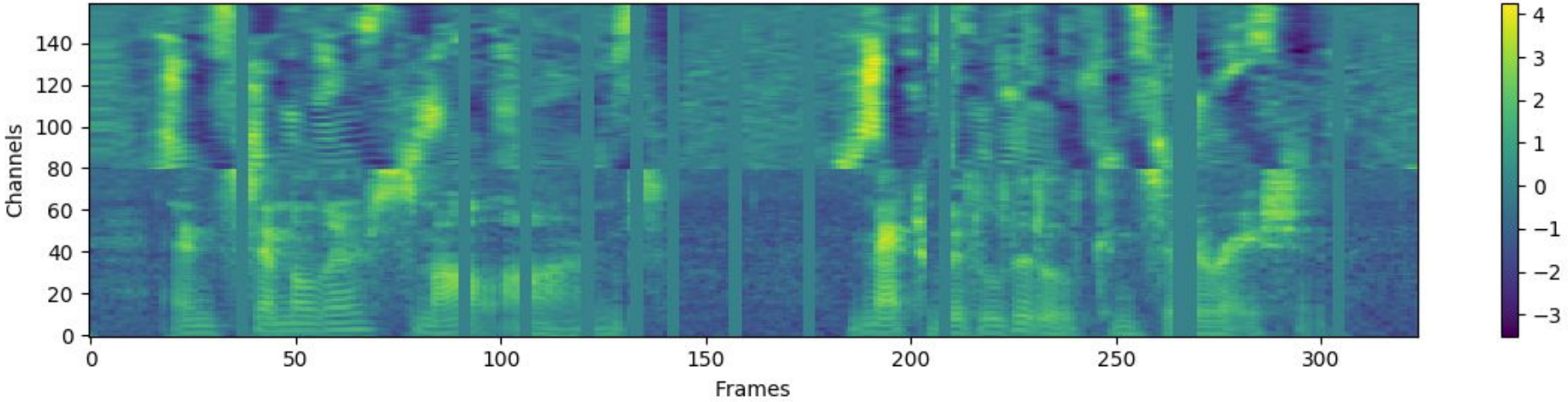
Select **15%** of the frames for prediction.

For all selected frames:

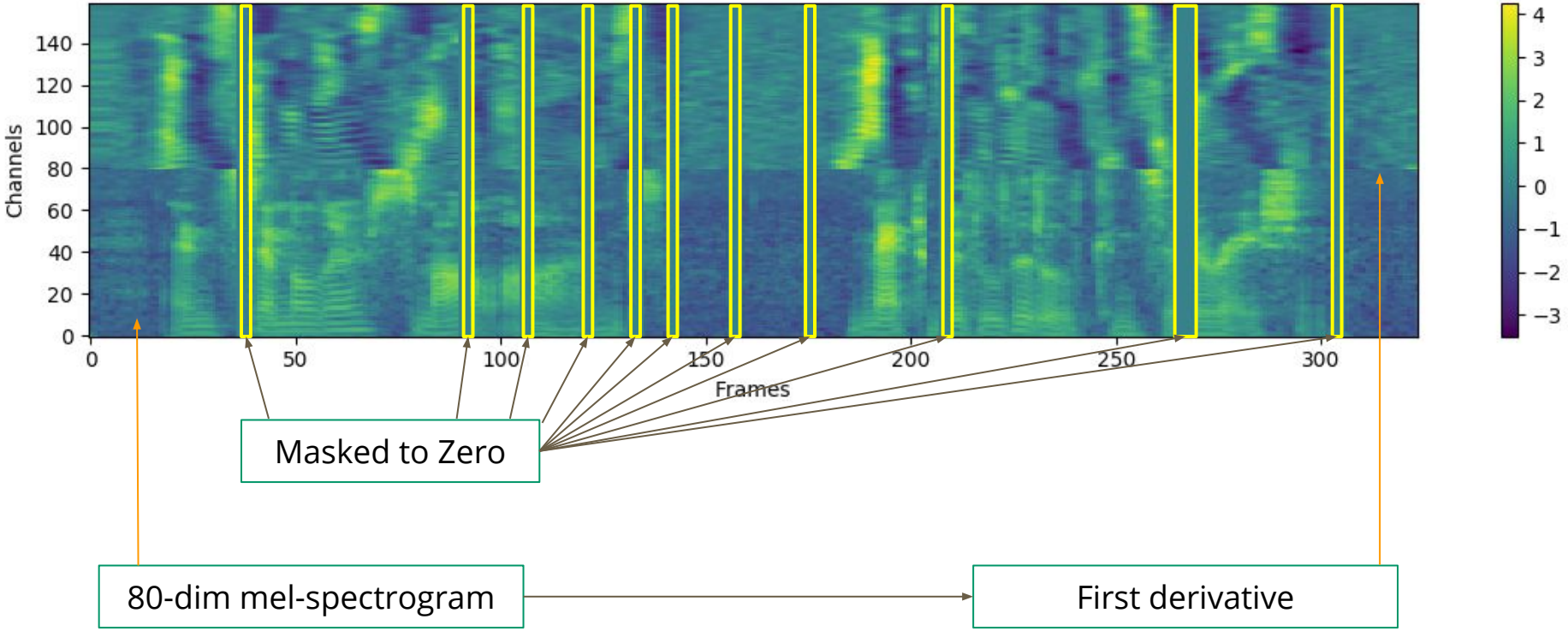
- mask to zero **80%** of the time
- replace randomly **10%** of the time
- leave untouch **10%** of the time

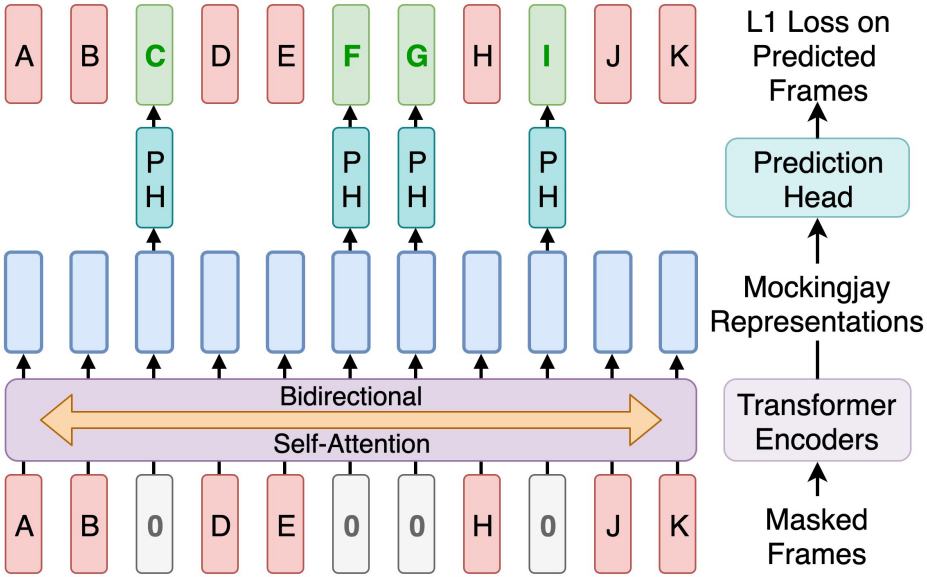


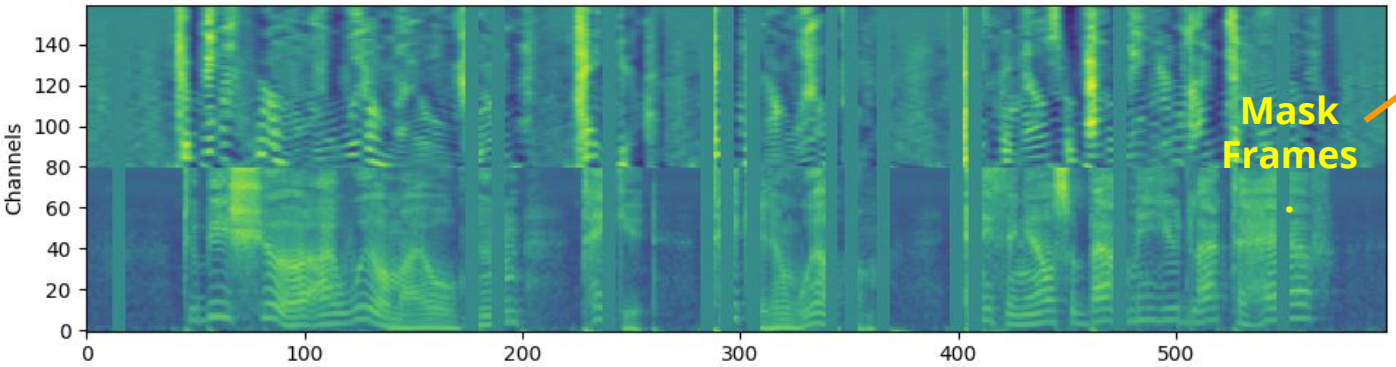
Input Feature: Masked Spectrogram



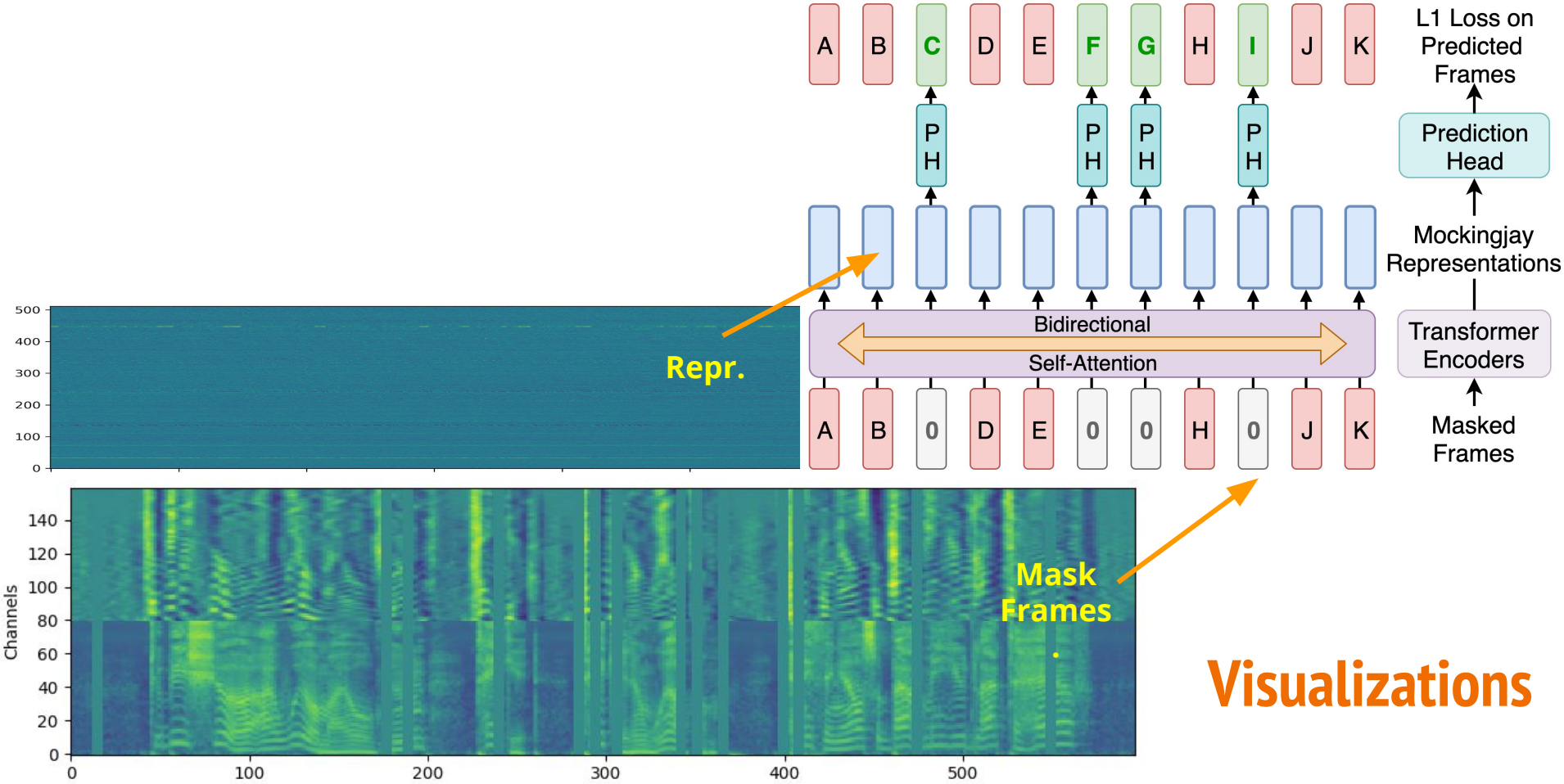
Input Feature: Masked Spectrogram

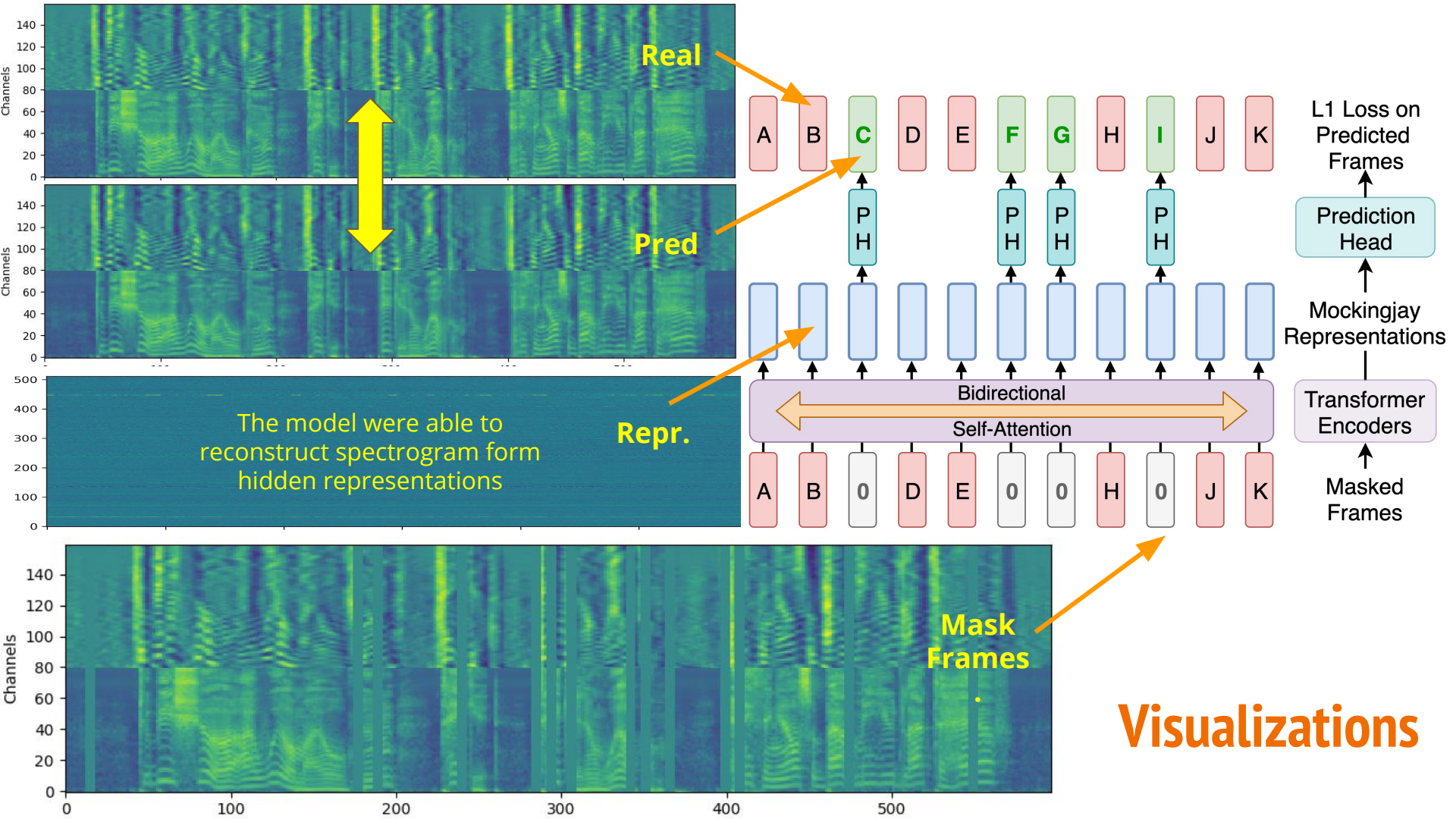






Visualizations

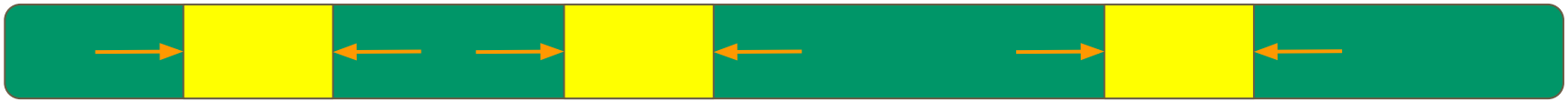

A B C D E F G H I J K
L1 Loss on Predicted Frames
P H
Prediction Head
Mockingjay Representations
Bidirectional Self-Attention
Transformer Encoders
A B 0 D E 0 0 H 0 J K
Masked Frames
Repr.
Mask Frames
Channels
Visualizations


Real
Pred
Repr.
The model were able to reconstruct spectrogram form hidden representations
Mask Frames
Channels
A B C D E F G H I J K
P H
Bidirectional Self-Attention
A B 0 D E 0 0 H 0 J K
L1 Loss on Predicted Frames
Prediction Head
Mockingjay Representations
Transformer Encoders
Masked Frames
Visualizations

Differences from BERT

Acoustic Features: long and locally smooth in nature,

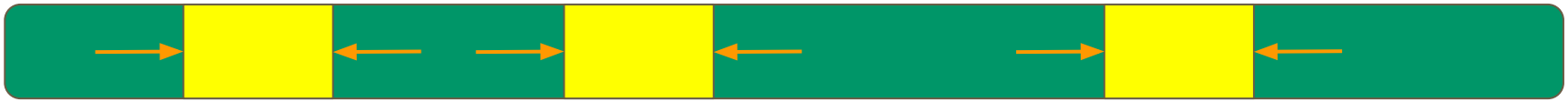
need to shorten the sequence and mask longer portions



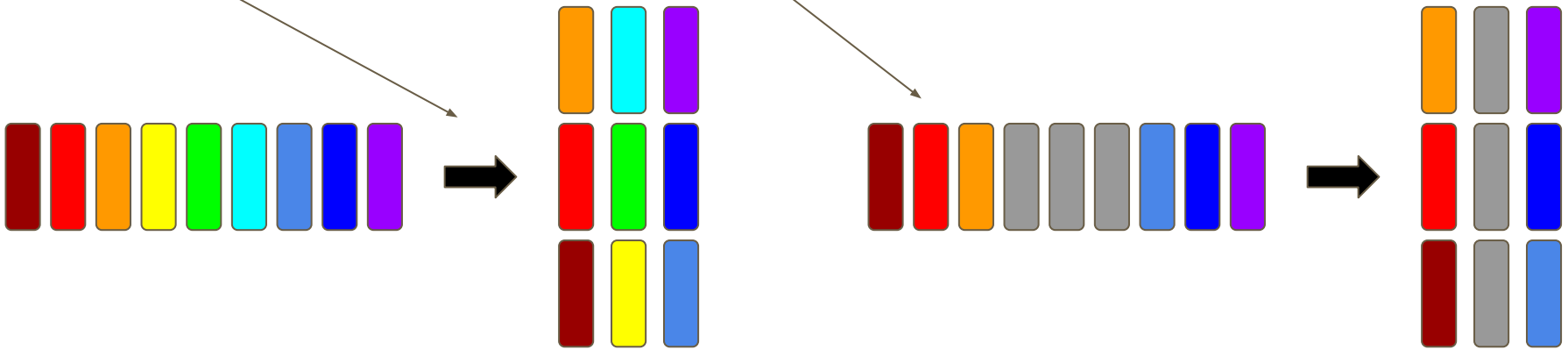
Differences from BERT

Acoustic Features: long and locally smooth in nature,

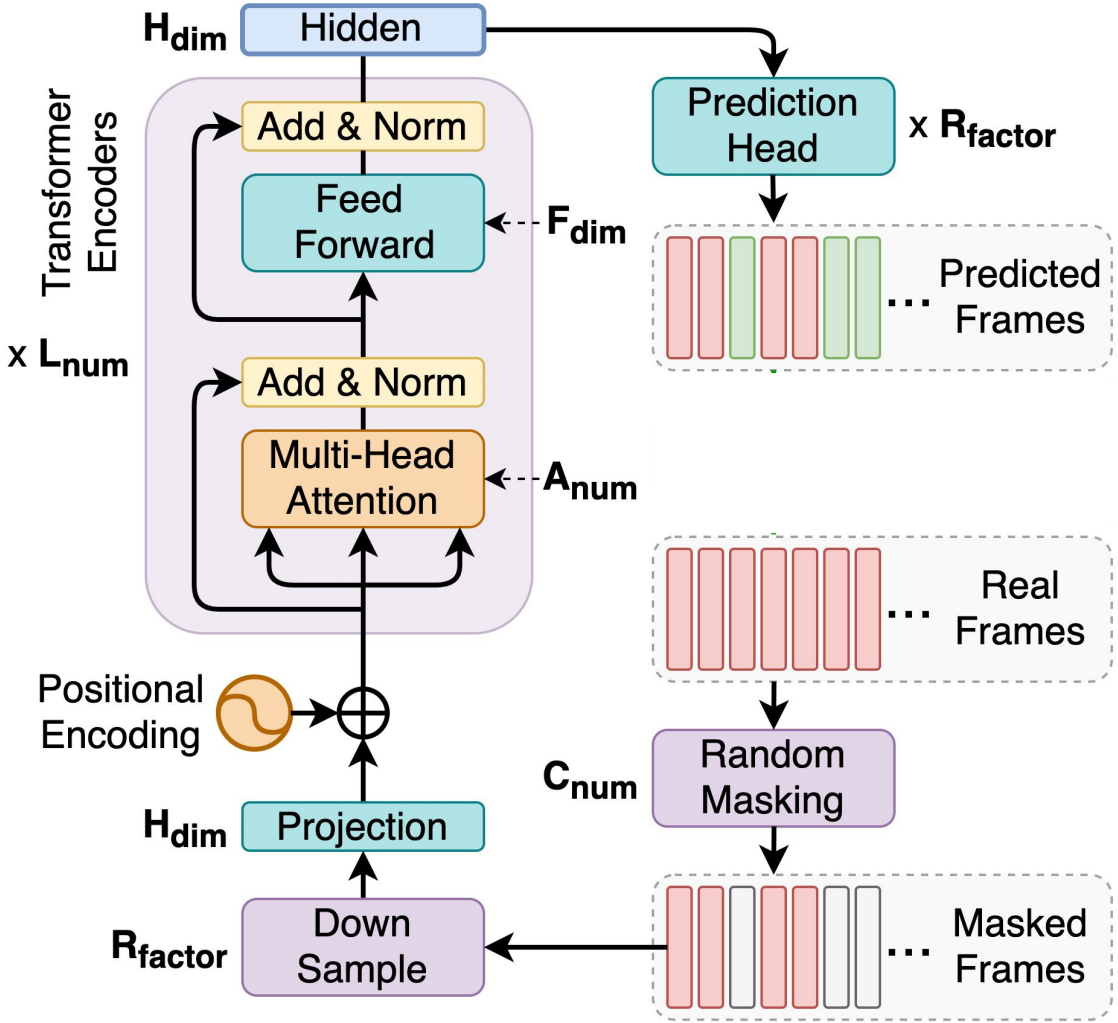
need to shorten the sequence and mask longer portions



Address the long and smooth problem with:
Downsampling, and *consecutive masking*



Model Architecture



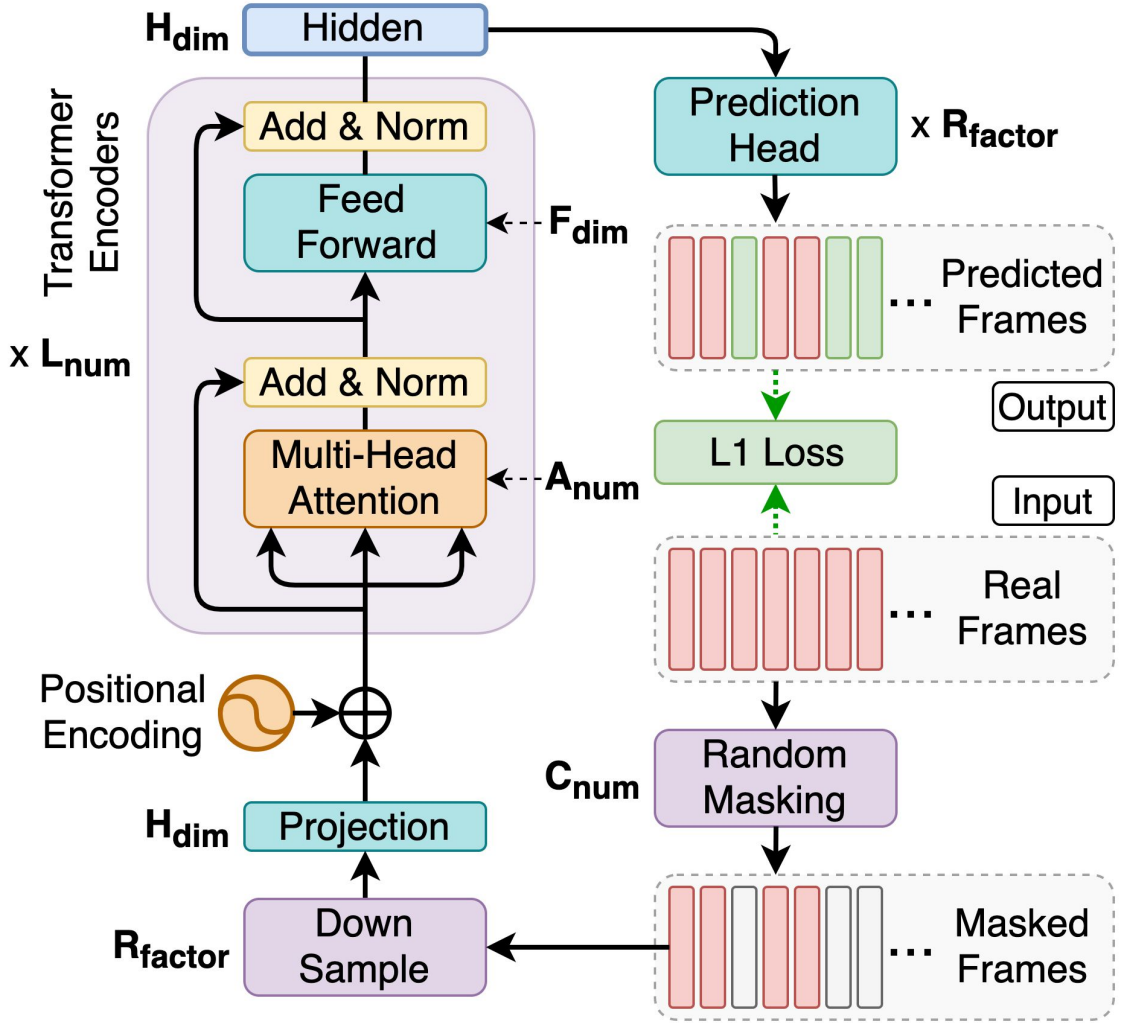
Model Architecture

- H_dim = 768
- F_dim = 3072
- A_num = 12

- Pre-train steps = 500k
- Fine-tune steps = 50k (2-epochs)

Table 1. The proposed BASE and LARGE model

Model	BASE	LARGE
Target	Mel	Linear
L_{num}	3	12
R_{factor}	1	3
C_{num}	7	3
parameters	21.4M	85.4M



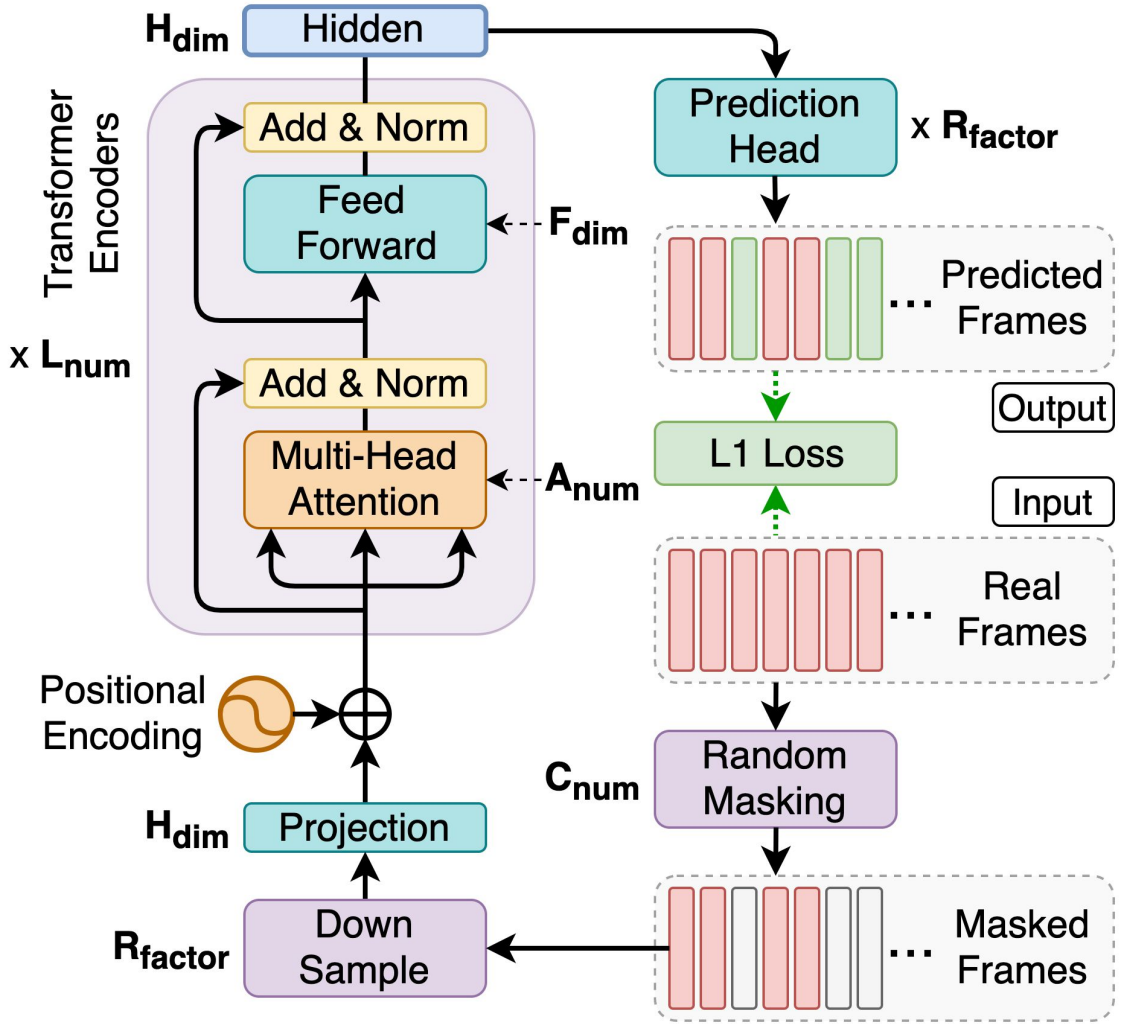
Model Architecture

- H_dim = 768
- F_dim = 3072
- A_num = 12

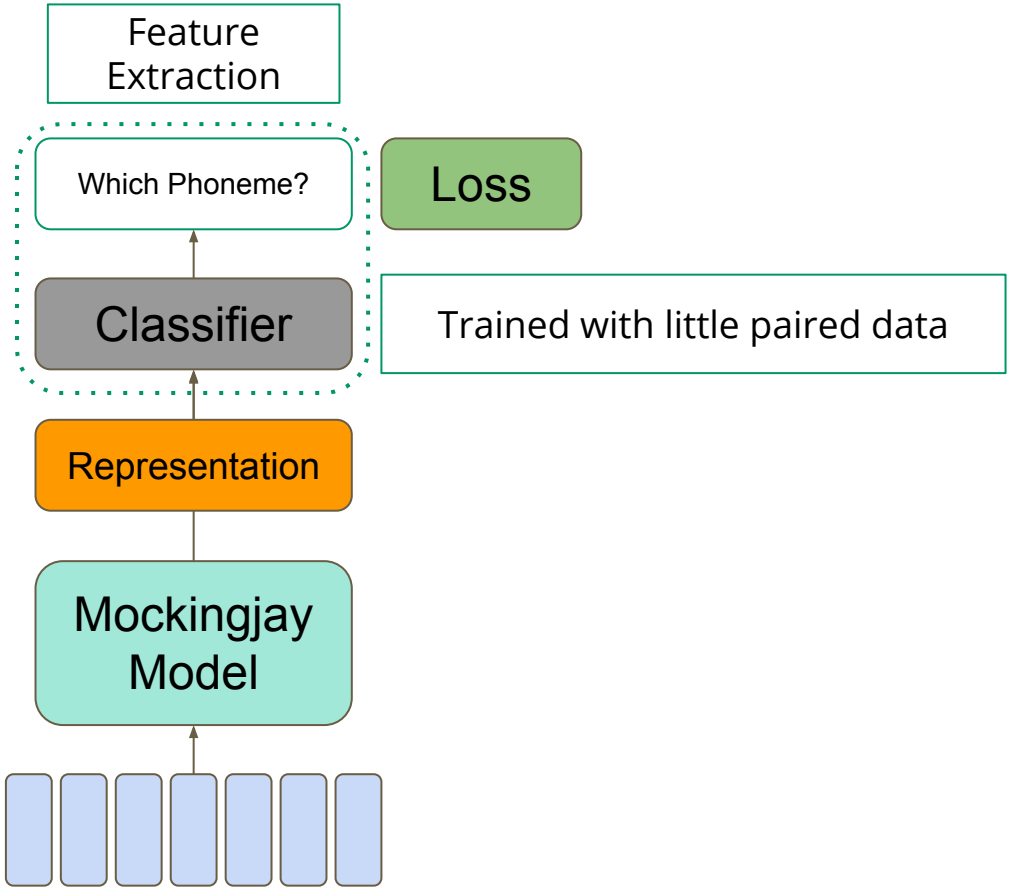
- Pre-train steps = **500k**
- Fine-tune steps = **50k** (2-epochs)

Table 1. The proposed BASE and LARGE model

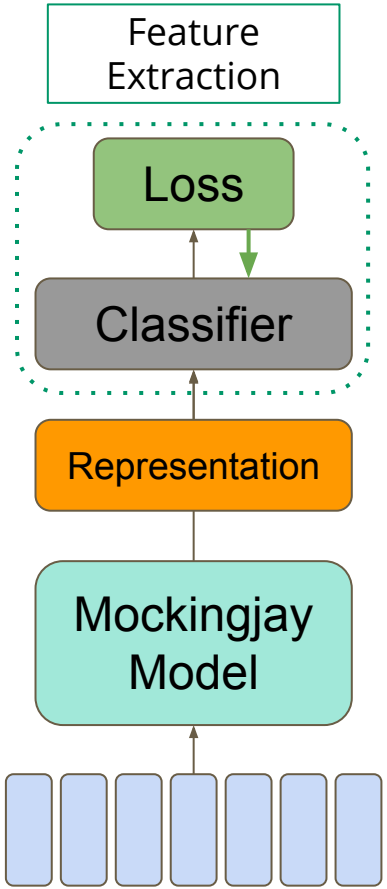
Model	BASE	LARGE
Target	Mel	Linear
L_{num}	3	12
R_{factor}	1	3
C_{num}	7	3
parameters	21.4M	85.4M

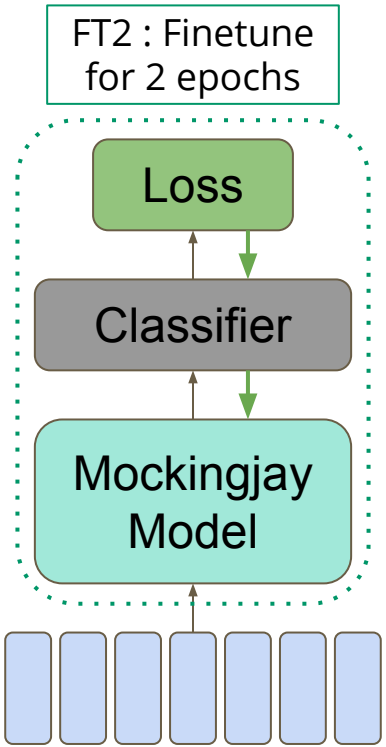


Incorporating with Downstream Tasks

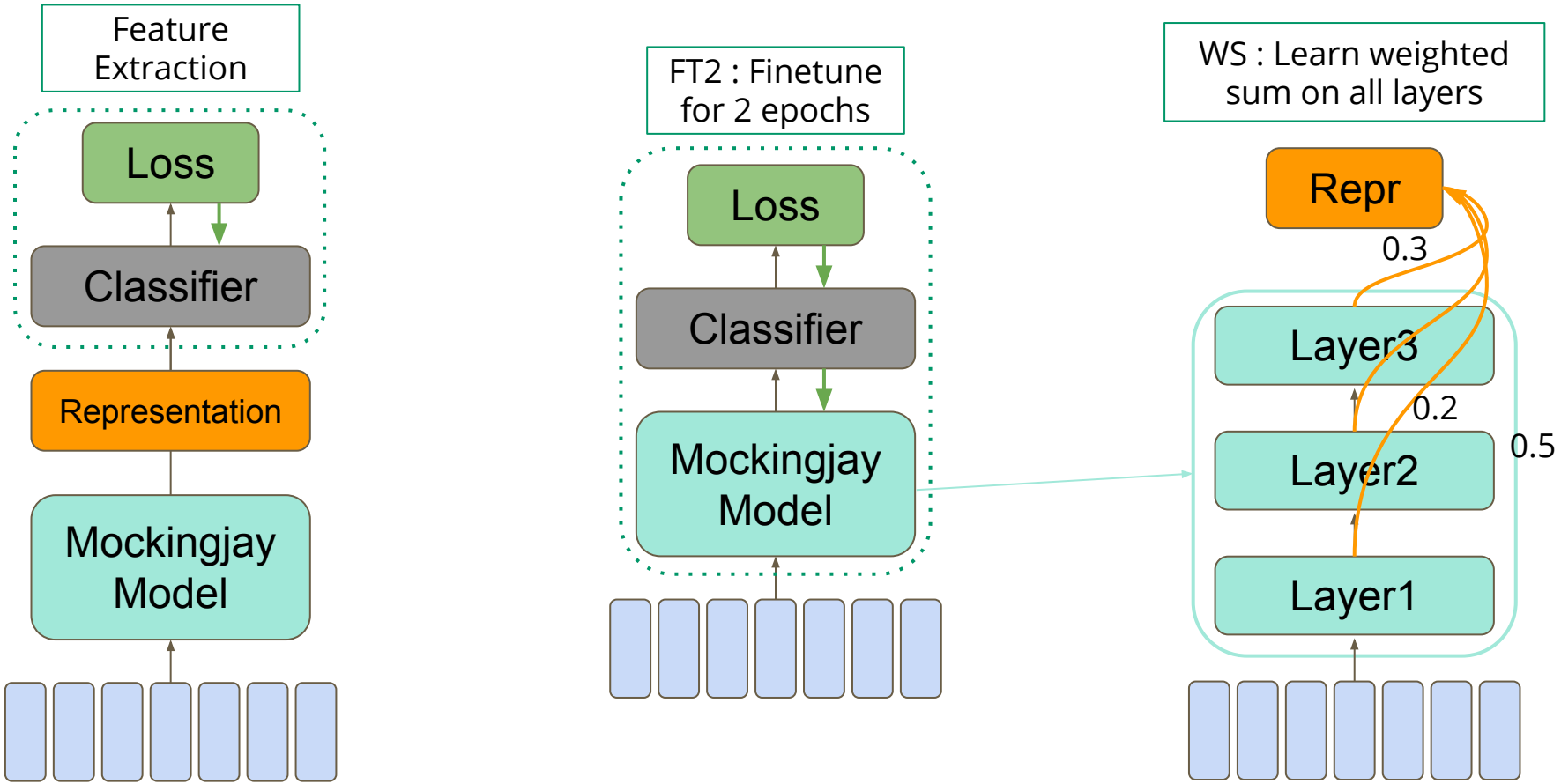


Incorporating with Downstream Tasks





Incorporating with Downstream Tasks



Experiments

We report results from 6 settings:

- Mel-Features
- APC representations
- BASE
- Large
- BASE-FT2
- Large-WS

On 2 different downstream tasks:

- Phoneme Classification
- Speaker Classification

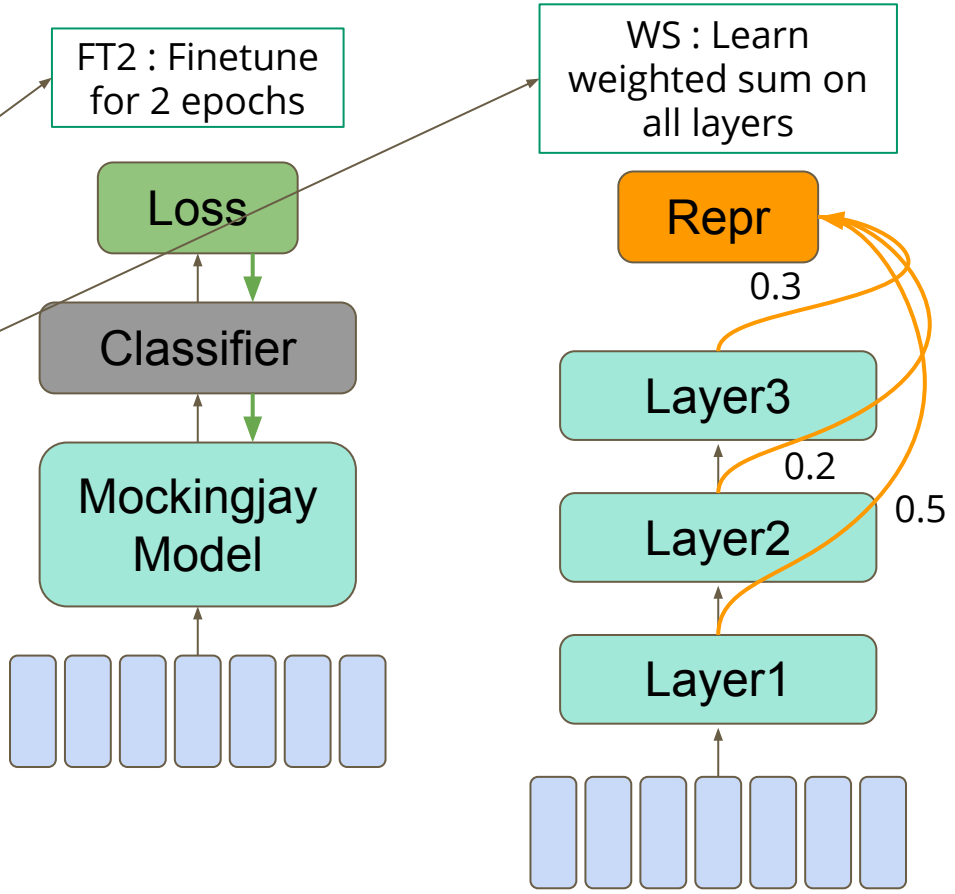
Experiments

We report results from 5 settings:

- Mel-Features
- APC representations
- BASE
- Large
- BASE-FT2
- Large-WS

On 2 different downstream tasks:

- Phoneme Classification
- Speaker Classification



Experiments

Acoustic Features	Phoneme Classification		Speaker Classification
	360 hr of labels (100%)	0.36 hr of labels (0.1%)	63 Speakers
Mel Features	49.1	35.2	70.1
APC representation	74.1	26.6	85.9
BASE	60.9	45.1	94.5
BASE-FT2	**84.3**	**57.9**	**98.1**
LARGE	64.3	46.6	96.3
LARGE-WS	69.9	52.8	96.4

Experiments

Acoustic Features	Phoneme Classification		Speaker Classification
	360 hr of labels (100%)	0.36 hr of labels (0.1%)	63 Speakers
Mel Features	49.1	35.2	70.1
APC representation	74.1	26.6	85.9
BASE	60.9	45.1	94.5
BASE-FT2	**84.3**	**57.9**	**98.1**
LARGE	64.3	46.6	96.3
LARGE-WS	69.9	52.8	96.4

Experiments

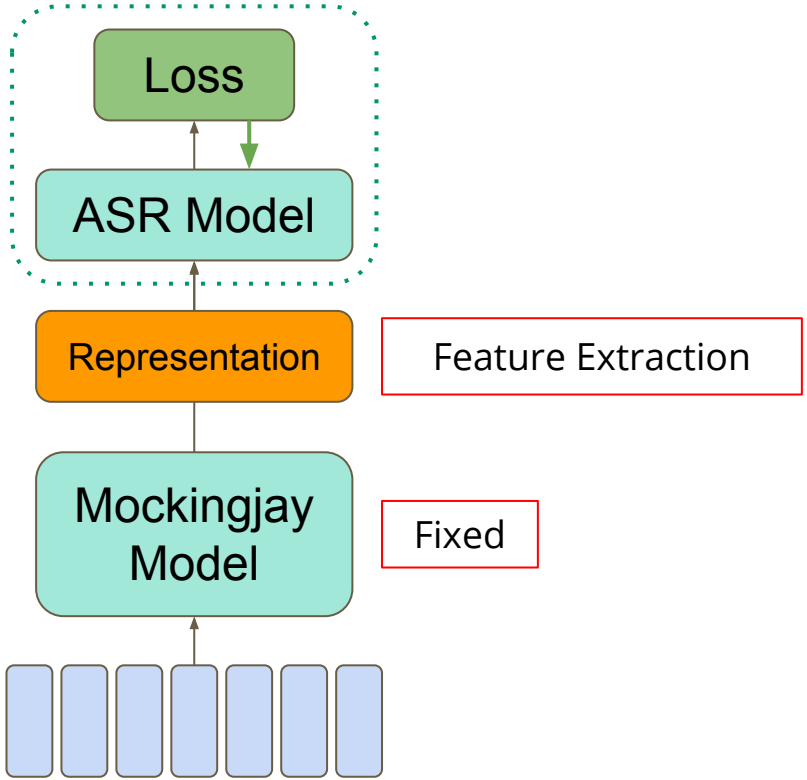
Acoustic Features	Phoneme Classification		Speaker Classification
	360 hr of labels (100%)	0.36 hr of labels (0.1%)	63 Speakers
Mel Features	49.1	35.2	70.1
APC representation	74.1	26.6	85.9
BASE	60.9	45.1	94.5
BASE-FT2	**84.3**	**57.9**	**98.1**
LARGE	64.3	46.6	96.3
LARGE-WS	69.9	52.8	96.4

Experiments

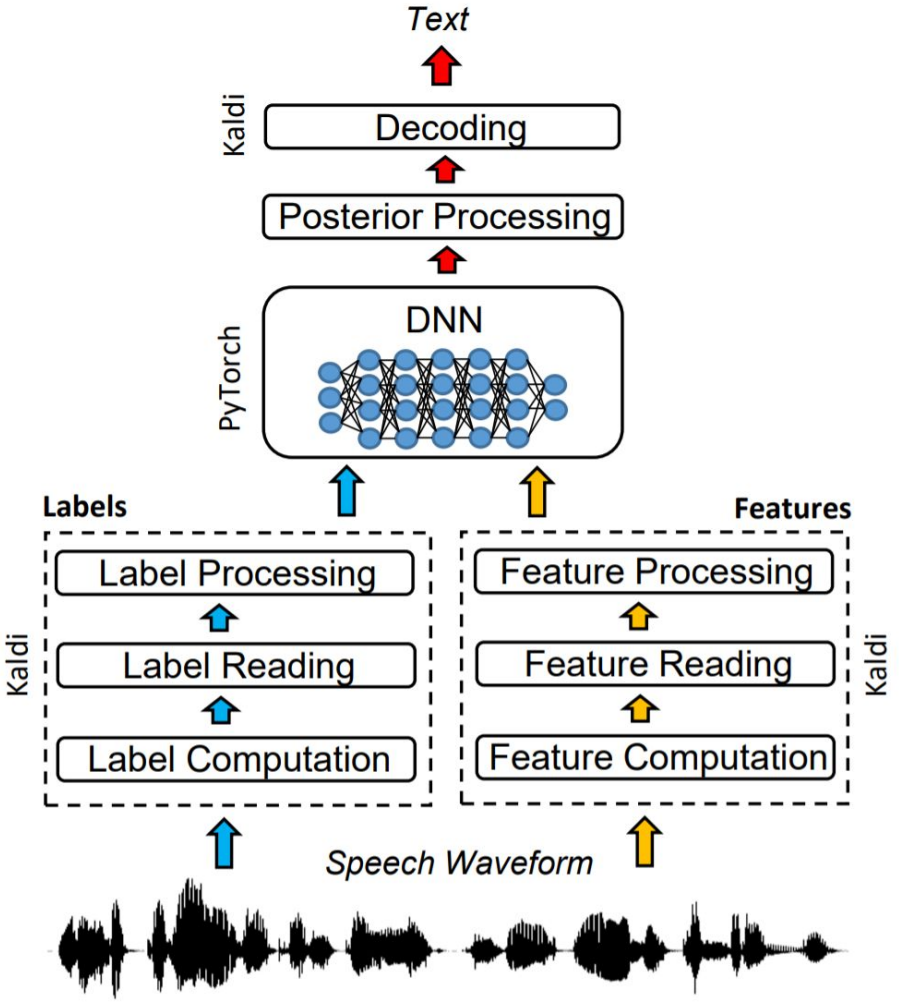
Acoustic Features	Phoneme Classification		Speaker Classification
	360 hr of labels (100%)	0.36 hr of labels (0.1%)	63 Speakers
Mel Features	49.1	35.2	70.1
APC representation	74.1	26.6	85.9
BASE	60.9	45.1	94.5
BASE-FT2	**84.3**	**57.9**	**98.1**
LARGE	64.3	46.6	96.3
LARGE-WS	69.9	52.8	96.4

Current work:

Combining with ASR



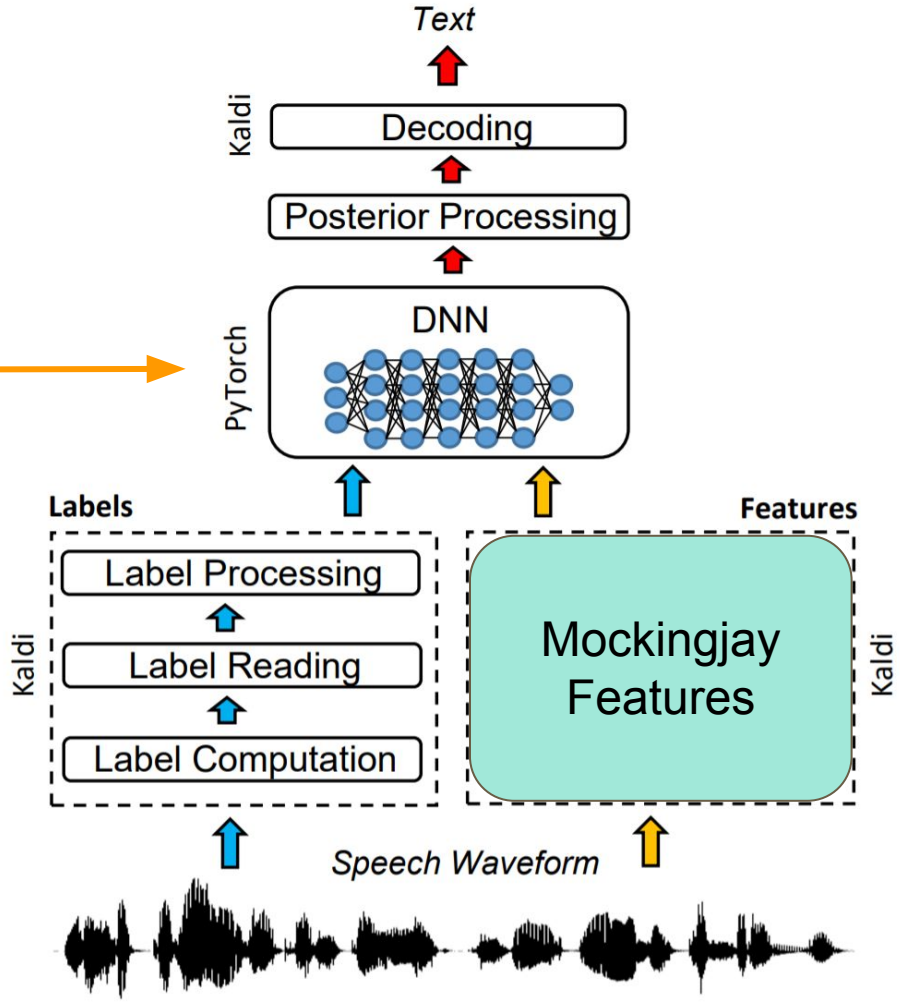
DNN/HMM Hybrid ASR with Pytorch Kaldi



DNN/HMM Hybrid ASR with Pytorch Kaldi

Light-GRU Model

train-clean-100 subset



Preliminary ASR Results - LibriSpeech

Model	Pre-train (hr)	WER (%)
Li-GRU + mfcc	None	8.38
Li-GRU + fMLLR	None	6.2
Bidir CPC [1]	Libri 960	9.41
Bidir CPC [1]	8000	8.70
vq-wav2vec gumbel + Transformer Big [2]	Libri 960	6.2
liGRU + Mockingjay (Ours)	Libri 100	6.32
liGRU + Mockingjay (Ours)	Libri 460	6.15
liGRU + Mockingjay (Ours)	Libri 960	**6.14**

[1] Unsupervised Pre-training of Bidirectional Speech Encoders via Masked Reconstruction https://openreview.net/pdf?id=HJe-blSYvH

[2] vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations https://arxiv.org/abs/1910.05453

Preliminary ASR Results - LibriSpeech

Model	Pre-train (hr)	WER (%)
Li-GRU + mfcc	None	8.38
Li-GRU + fMLLR	None	6.2
Bidir CPC [1]	Libri 960	9.41
Bidir CPC [1]	8000	8.70
vq-wav2vec gumbel + Transformer Big [2]	Libri 960	6.2
liGRU + Mockingjay (Ours)	Libri 100	6.32
liGRU + Mockingjay (Ours)	Libri 460	6.15
liGRU + Mockingjay (Ours)	Libri 960	**6.14**

[1] Unsupervised Pre-training of Bidirectional Speech Encoders via Masked Reconstruction https://openreview.net/pdf?id=HJe-blSYvH

[2] vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations https://arxiv.org/abs/1910.05453

Preliminary ASR Results - LibriSpeech

Model	Pre-train (hr)	WER (%)
Li-GRU + mfcc	None	8.38
Li-GRU + fMLLR	None	6.2
Bidir CPC [1]	Libri 960	9.41
Bidir CPC [1]	8000	8.70
vq-wav2vec gumbel + Transformer Big [2]	Libri 960	6.2
liGRU + Mockingjay (Ours)	Libri 100	6.32
liGRU + Mockingjay (Ours)	Libri 460	6.15
liGRU + Mockingjay (Ours)	Libri 960	**6.14**

[1] Unsupervised Pre-training of Bidirectional Speech Encoders via Masked Reconstruction https://openreview.net/pdf?id=HJe-blSYvH

[2] vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations https://arxiv.org/abs/1910.05453

Preliminary ASR Results - LibriSpeech

Model	Pre-train (hr)	WER (%)
Li-GRU + mfcc	None	8.38
Li-GRU + fMLLR	None	6.2
Bidir CPC [1]	Libri 960	9.41
Bidir CPC [1]	8000	8.70
vq-wav2vec gumbel + Transformer Big [2]	Libri 960	6.2
liGRU + Mockingjay (Ours)	Libri 100	6.32
liGRU + Mockingjay (Ours)	Libri 460	6.15
liGRU + Mockingjay (Ours)	Libri 960	**6.14**

[1] Unsupervised Pre-training of Bidirectional Speech Encoders via Masked Reconstruction https://openreview.net/pdf?id=HJe-blSYvH

[2] vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations https://arxiv.org/abs/1910.05453

Preliminary ASR Results - TIMIT

Model	Pre-train (hr)	PER (%)
CNN + TD-filterbanks	None	18.0
Li-GRU + mfcc	None	16.7
Li-GRU + fMLLR	None	14.9
wav2vec	Libri 80	17.6
wav2vec	Libri 960	15.6
wav2vec	Libri 960 + WSJ 81	14.7
liGRU + Mockingjay (Ours)	Libri 460	**14.4**
liGRU + Mockingjay (Ours)	Libri 960	**14.4**

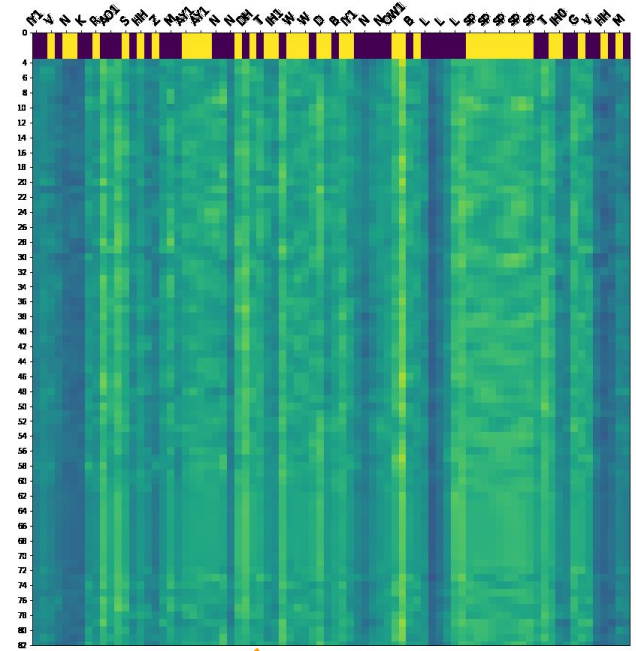
Preliminary ASR Results - TIMIT

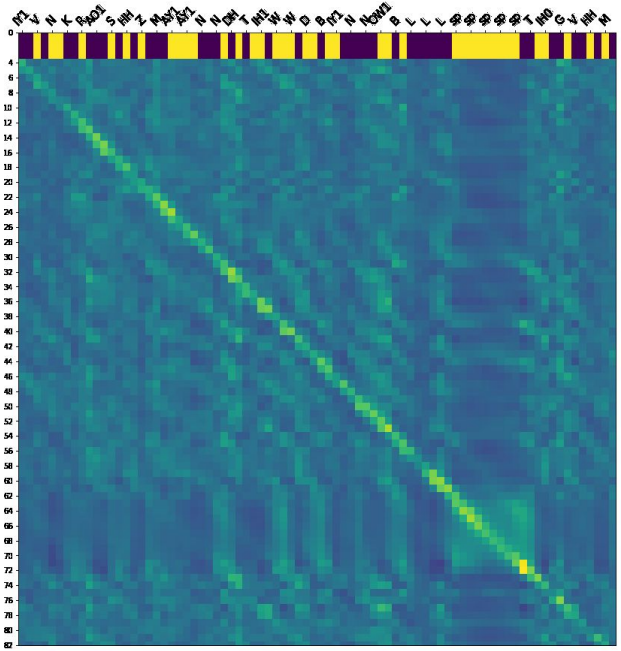
Model	Pre-train (hr)	PER (%)
CNN + TD-filterbanks	None	18.0
Li-GRU + mfcc	None	16.7
Li-GRU + fMLLR	None	14.9
wav2vec	Libri 80	17.6
wav2vec	Libri 960	15.6
wav2vec	Libri 960 + WSJ 81	14.7
liGRU + Mockingjay (Ours)	Libri 460	**14.4**
liGRU + Mockingjay (Ours)	Libri 960	**14.4**

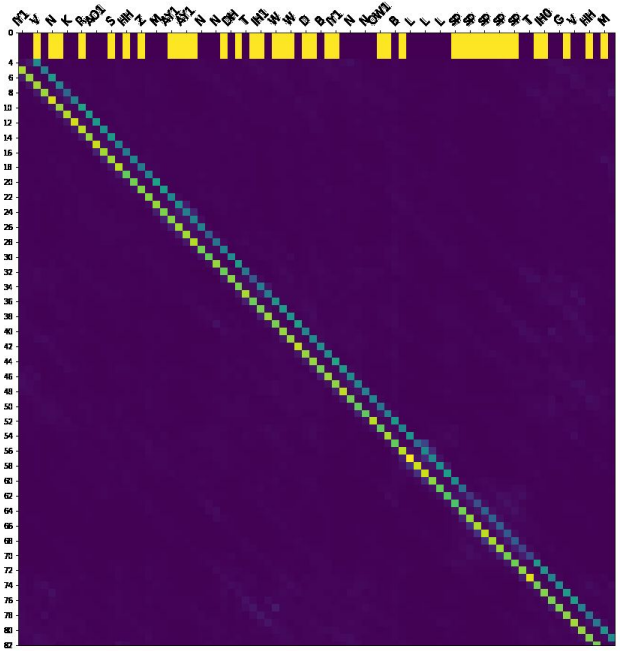
Preliminary ASR Results - TIMIT

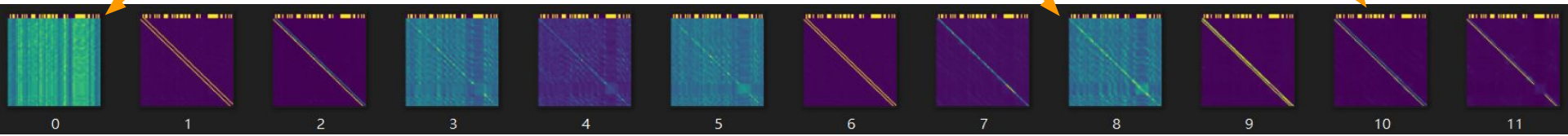
Model	Pre-train (hr)	PER (%)
CNN + TD-filterbanks	None	18.0
Li-GRU + mfcc	None	16.7
Li-GRU + fMLLR	None	14.9
wav2vec	Libri 80	17.6
wav2vec	Libri 960	15.6
wav2vec	Libri 960 + WSJ 81	14.7
liGRU + Mockingjay (Ours)	Libri 460	**14.4**
liGRU + Mockingjay (Ours)	Libri 960	**14.4**

Attention Maps - What Each Layer Does?

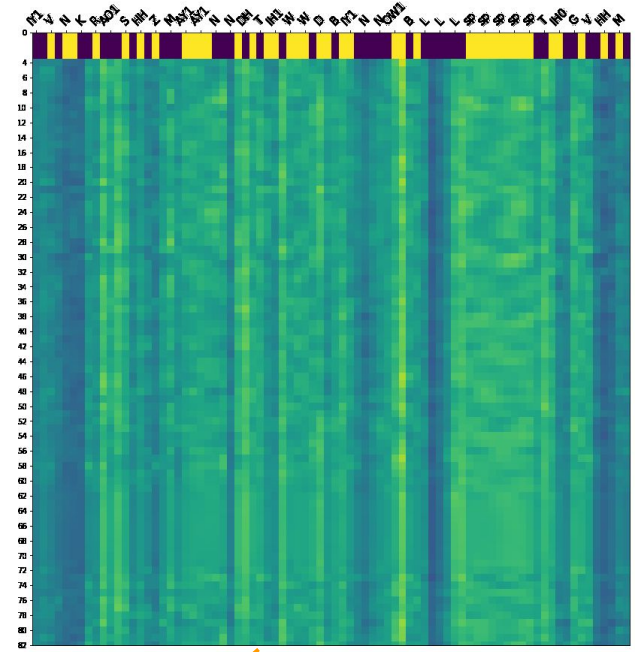


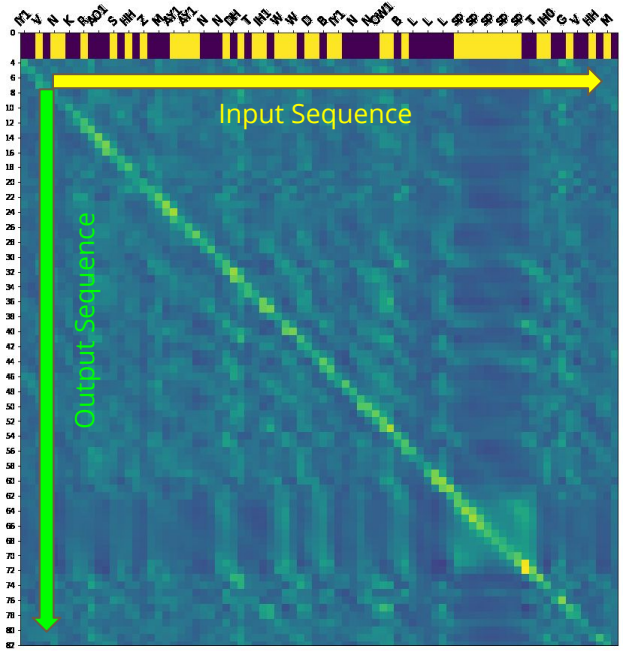


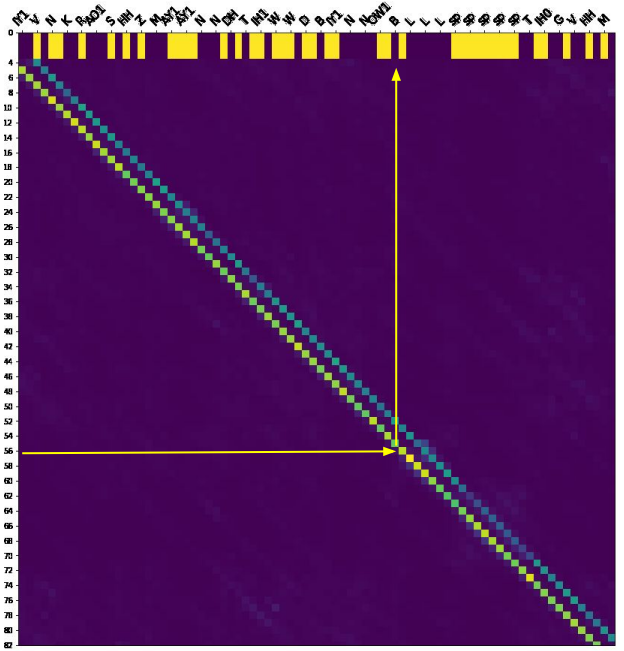


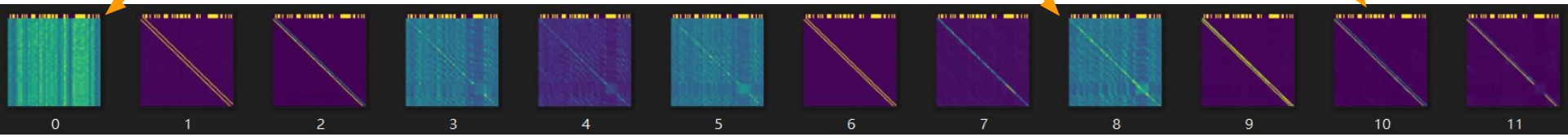


Attention Maps - What Each Layer Does?

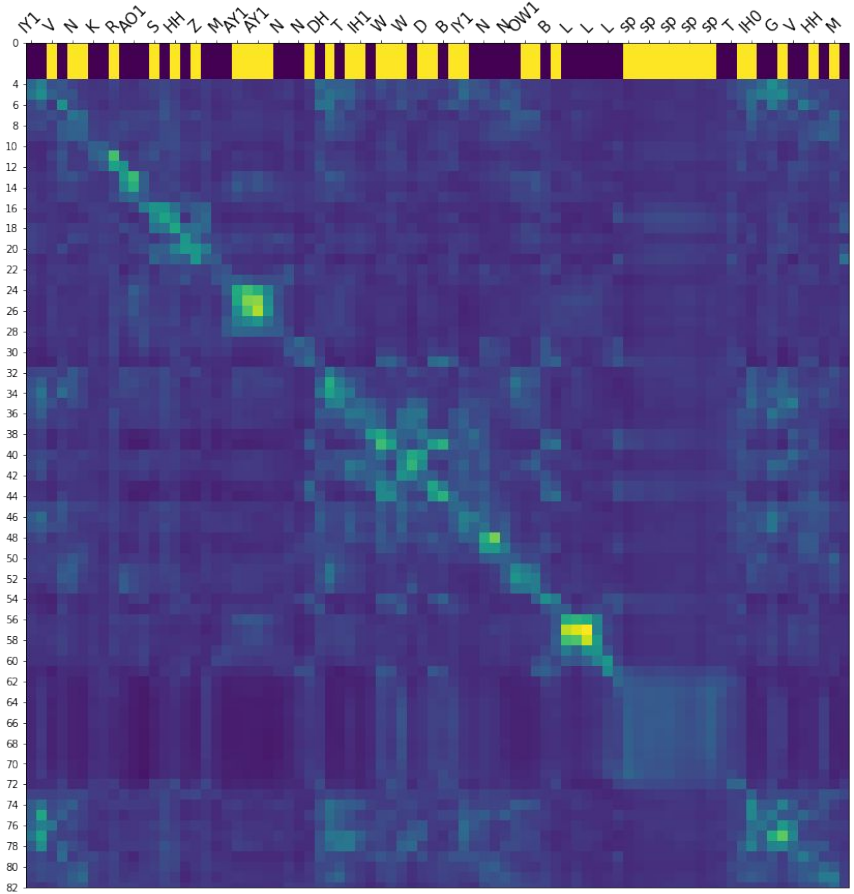




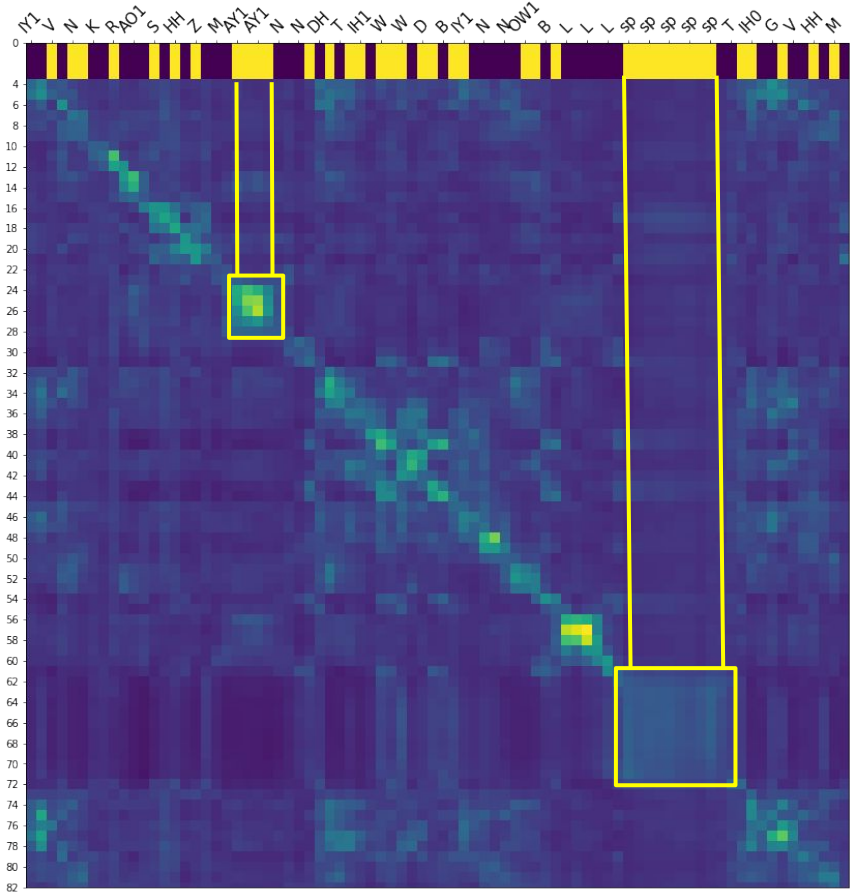




Attention Maps - Observing Phoneme Boundaries



Attention Maps - Observing Phoneme Boundaries



Thank You

Q&A