Mockingjay:
Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders

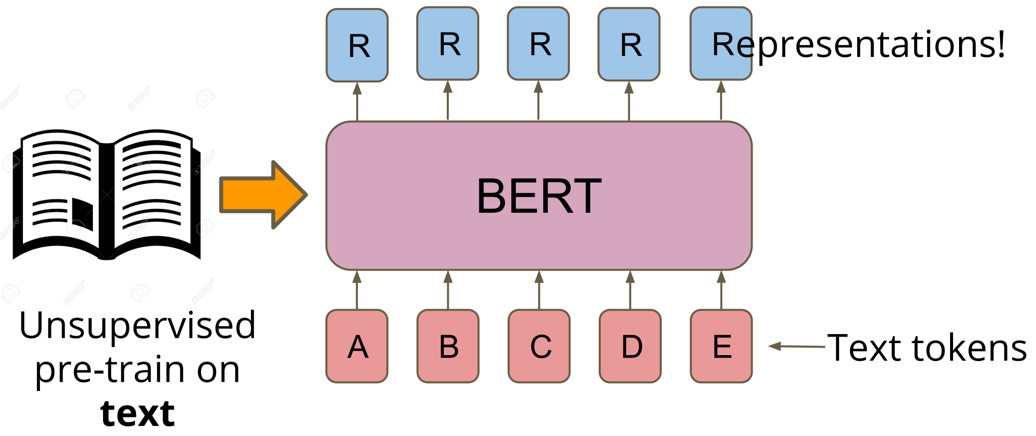
Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, Hung-yi Lee

Speech Lab, National Taiwan University (NTU)

Introduction

NLP BERT:

Language Representation Learning

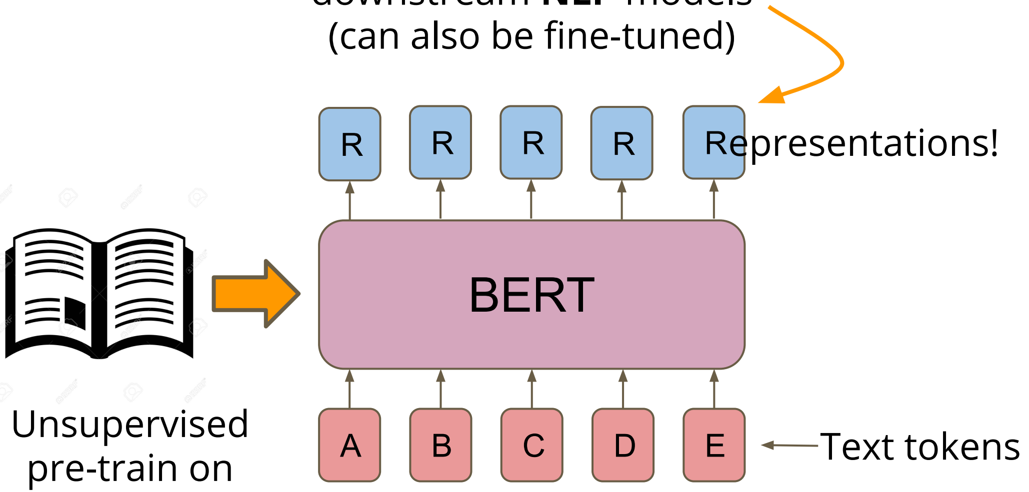


Introduction

NLP BERT:

<u>Language Representation Learning</u>

<u>Usage:</u>
Extracts features for
downstream **NLP** models
(can also be fine-tuned)



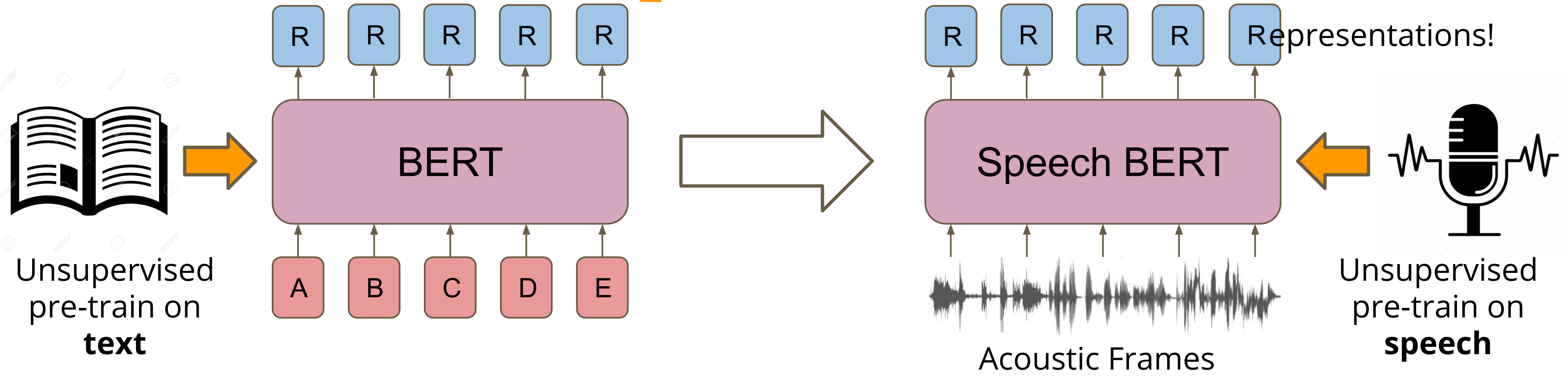
Unsupervised
pre-train on
text

Introduction

NLP BERT:
Language Representation Learning

Speech BERT:
Speech Representation Learning

Usage:
Extracts features for
downstream **NLP** models
(can also be fine-tuned)



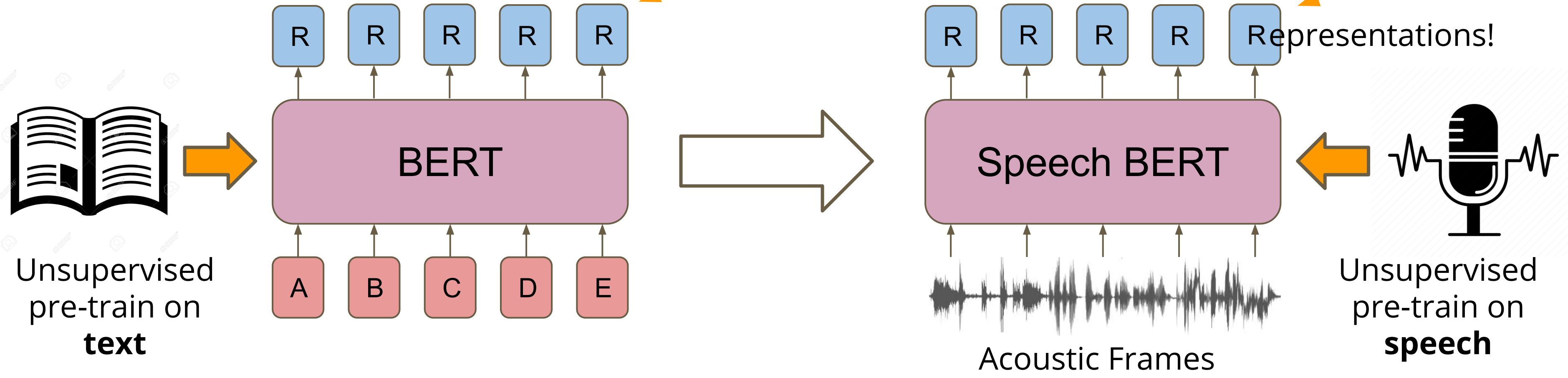
Introduction

NLP BERT:
Language Representation Learning

Usage:
Extracts features for
downstream **NLP** models
(can also be fine-tuned)

Speech BERT:
Speech Representation Learning

Usage:
Extracts features for
downstream **SLP** models
(can also be fine-tuned)



A View of Recent Unsupervised Speech Representation Learning Approaches

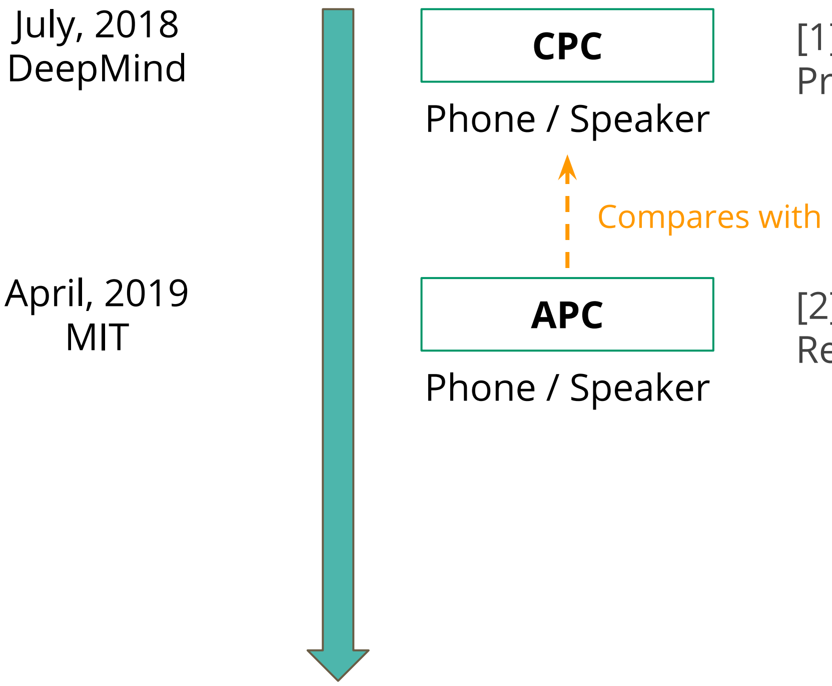
July, 2018
DeepMind

CPC

Phone / Speaker

[1] Representation Learning with Contrastive Predictive Coding

A View of Recent Unsupervised Speech Representation Learning Approaches



[1] Representation Learning with Contrastive Predictive Coding

[2] An Unsupervised Autoregressive Model for Speech Representation Learning

A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

CPC

Phone / Speaker

[1] Representation Learning with Contrastive Predictive Coding

April, 2019
MIT

APC

Phone / Speaker

[2] An Unsupervised Autoregressive Model for Speech Representation Learning

Common heuristic:
They both encode past information and predict information about future frames.

A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

CPC

Phone / Speaker

[1] Representation Learning with Contrastive Predictive Coding

April, 2019
MIT

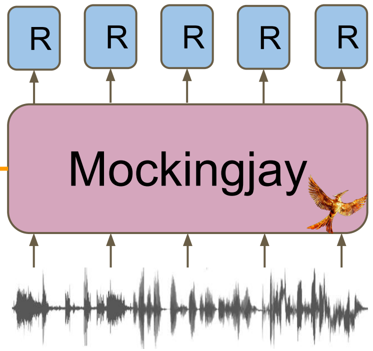
APC

Phone / Speaker

[2] An Unsupervised Autoregressive Model for Speech Representation Learning

October, 2019
Ours

Mockingjay



A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

CPC

Phone / Speaker

[1] Representation Learning with Contrastive Predictive Coding

Compares with

April, 2019
MIT

APC

Phone / Speaker

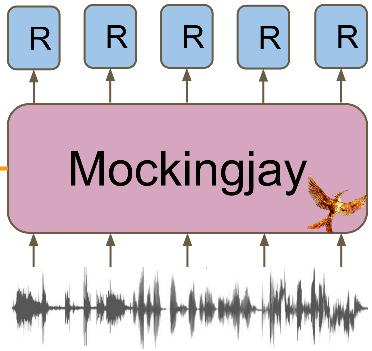
[2] An Unsupervised Autoregressive Model for Speech Representation Learning

Compares with

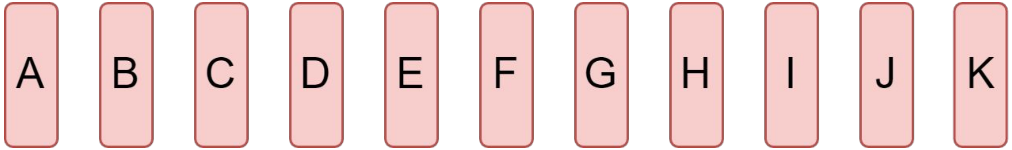
October, 2019
Ours

Mockingjay

Phone / Speaker / Sentiment



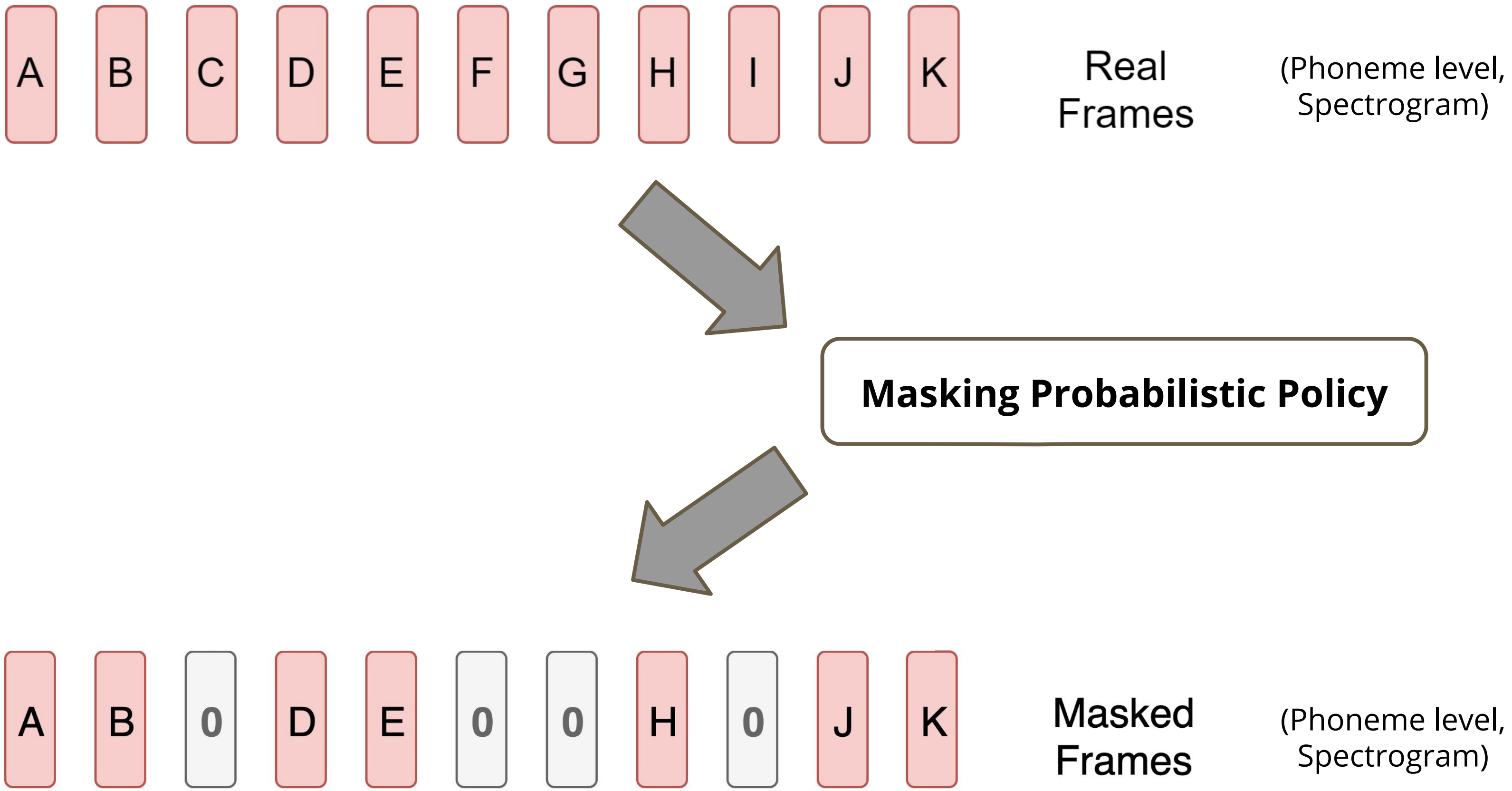
Pre-Training Task: Masked Acoustic Model



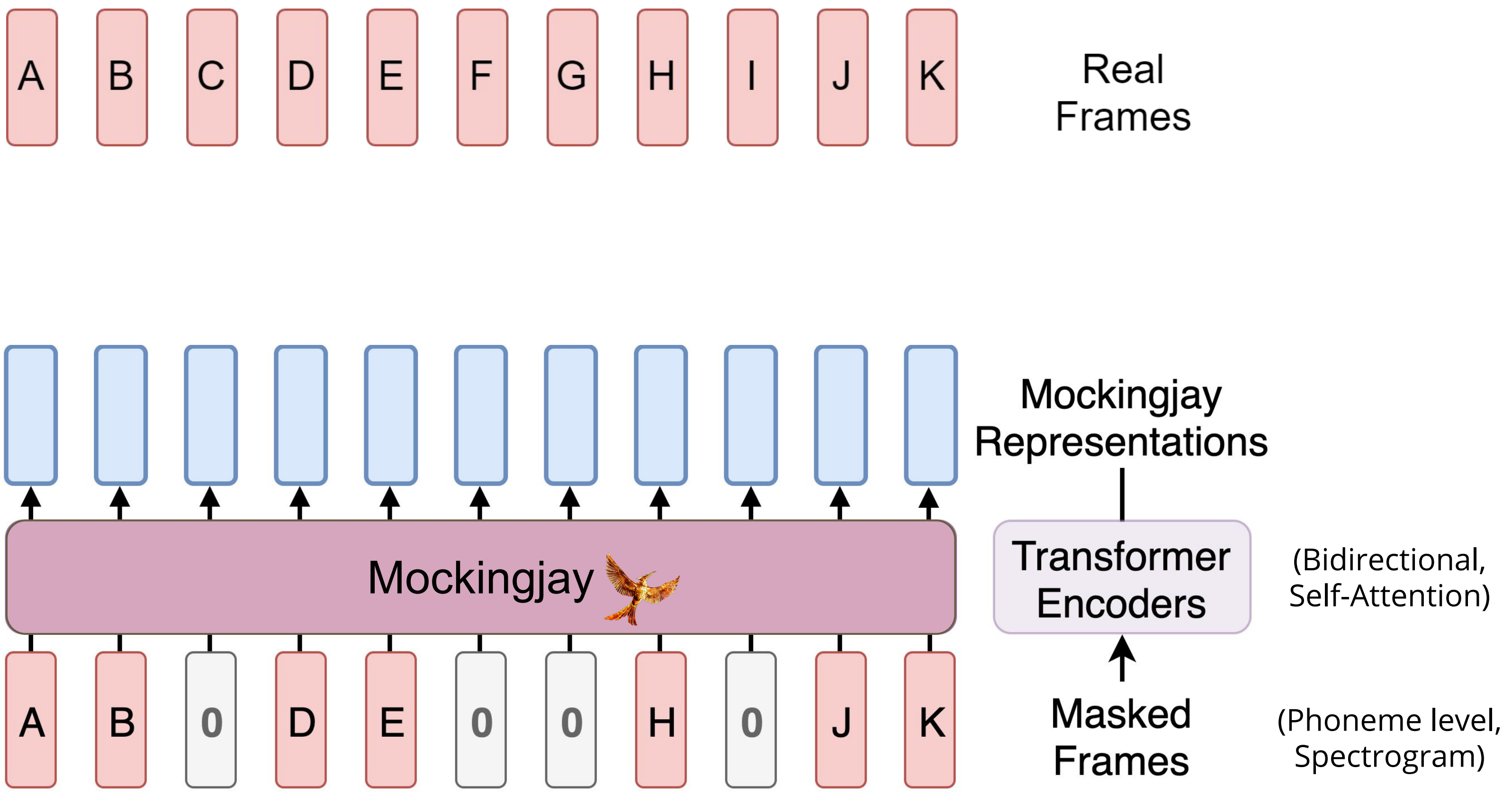
Real Frames

(Phoneme level, Spectrogram)

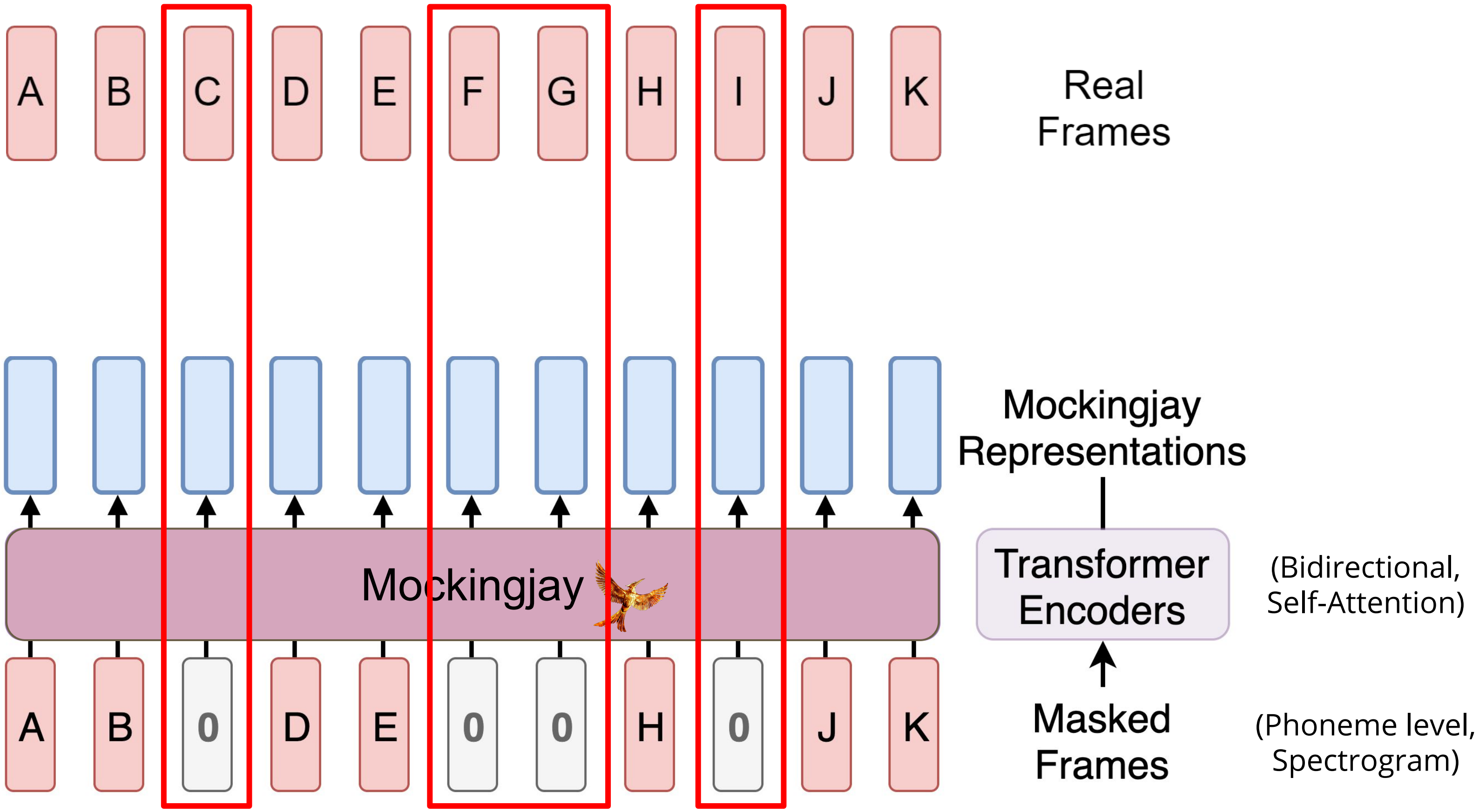
Pre-Training Task: Masked Acoustic Model



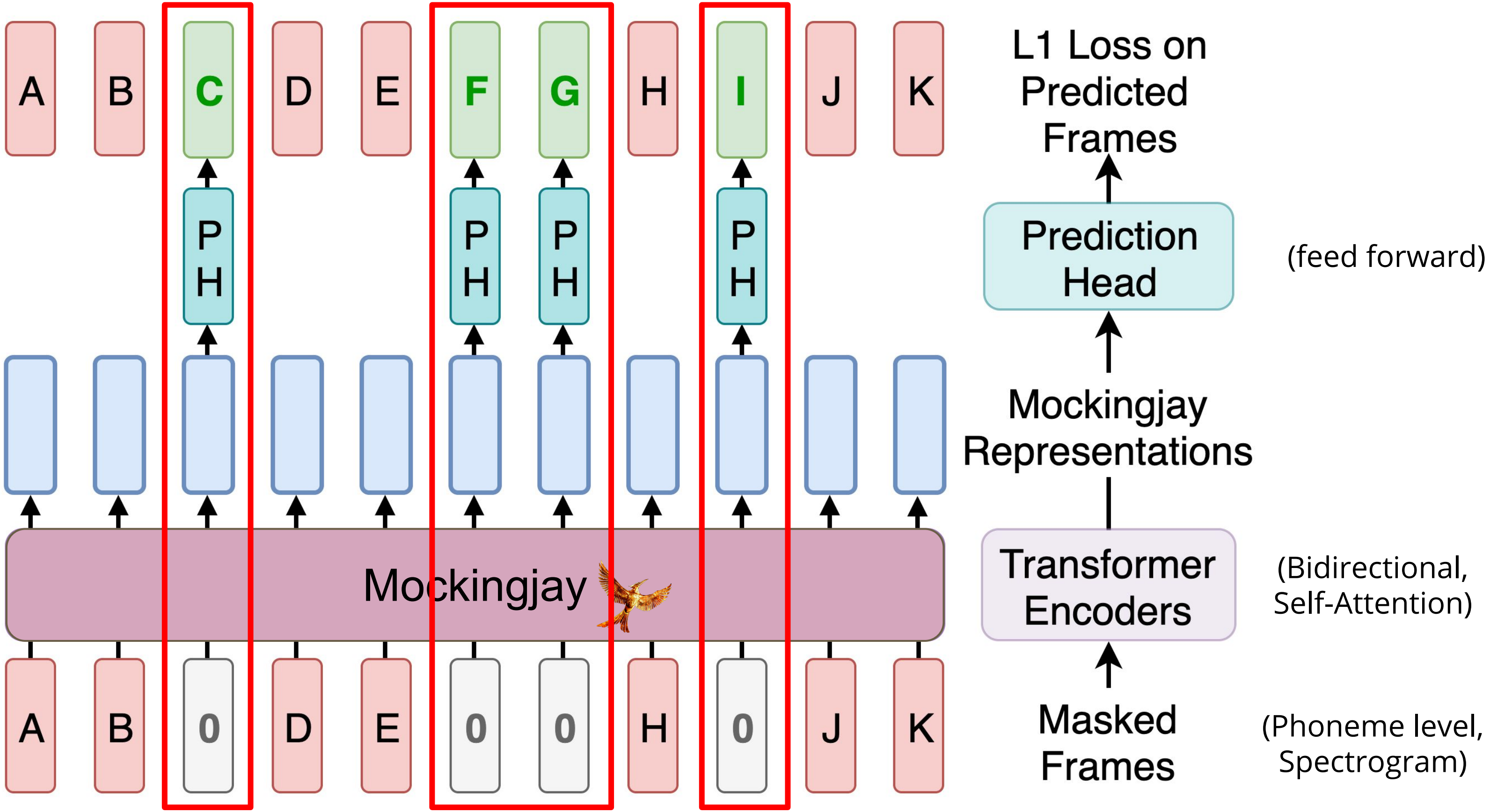
Pre-Training Task: Masked Acoustic Model



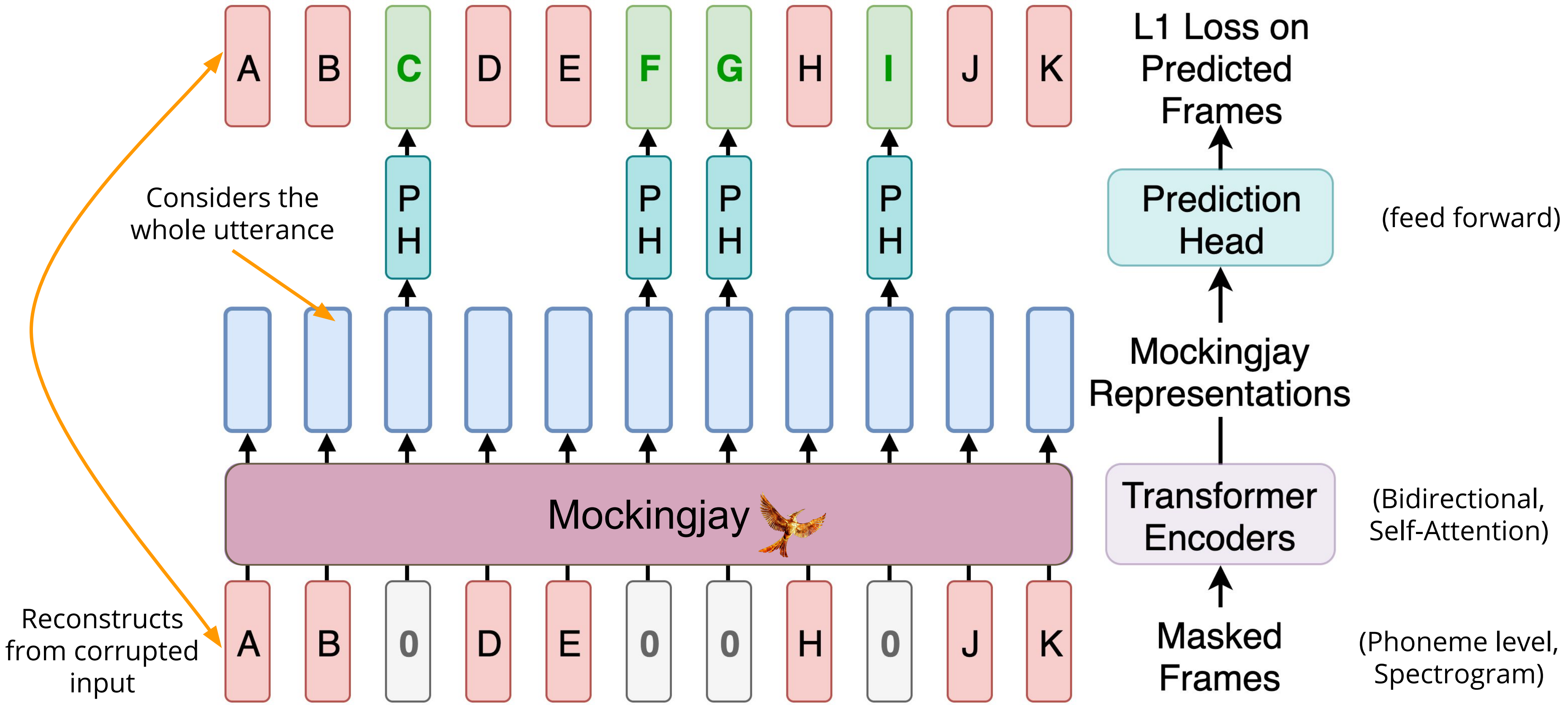
Pre-Training Task: Masked Acoustic Model



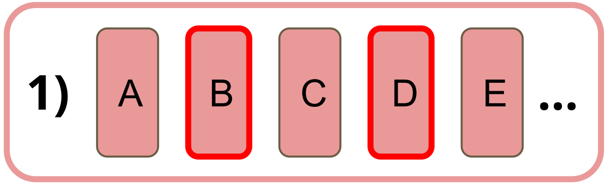
Pre-Training Task: Masked Acoustic Model



Pre-Training Task: Masked Acoustic Model

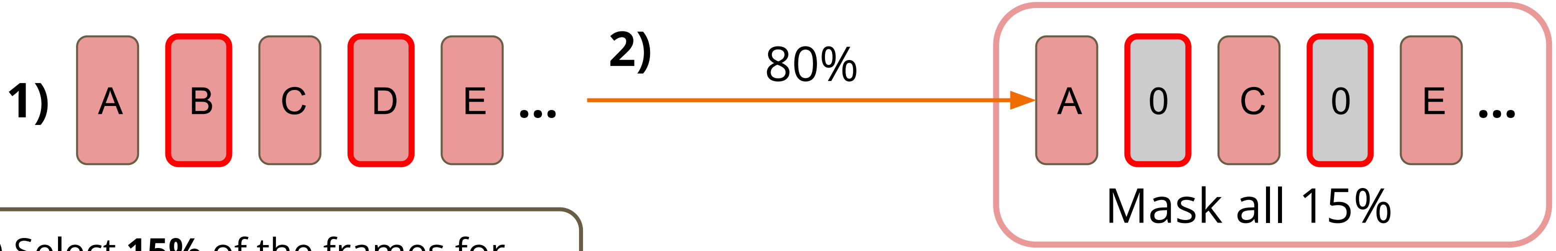


Probabilistic Policy for Masking Frames



1) Select **15%** of the frames for prediction (highlighted in green).

Probabilistic Policy for Masking Frames

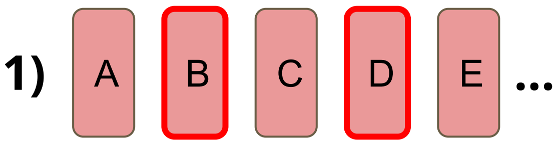


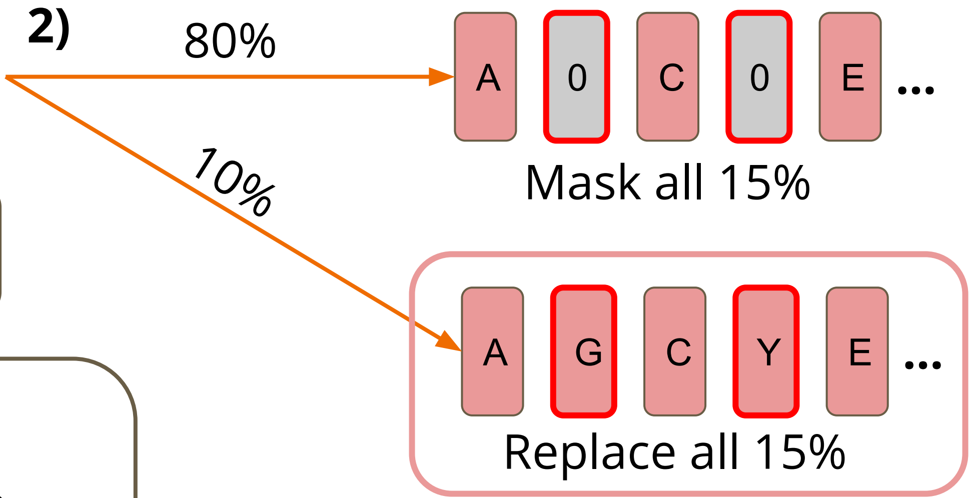
1) Select **15%** of the frames for prediction (highlighted in green).

2) For all selected frames:

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

Probabilistic Policy for Masking Frames



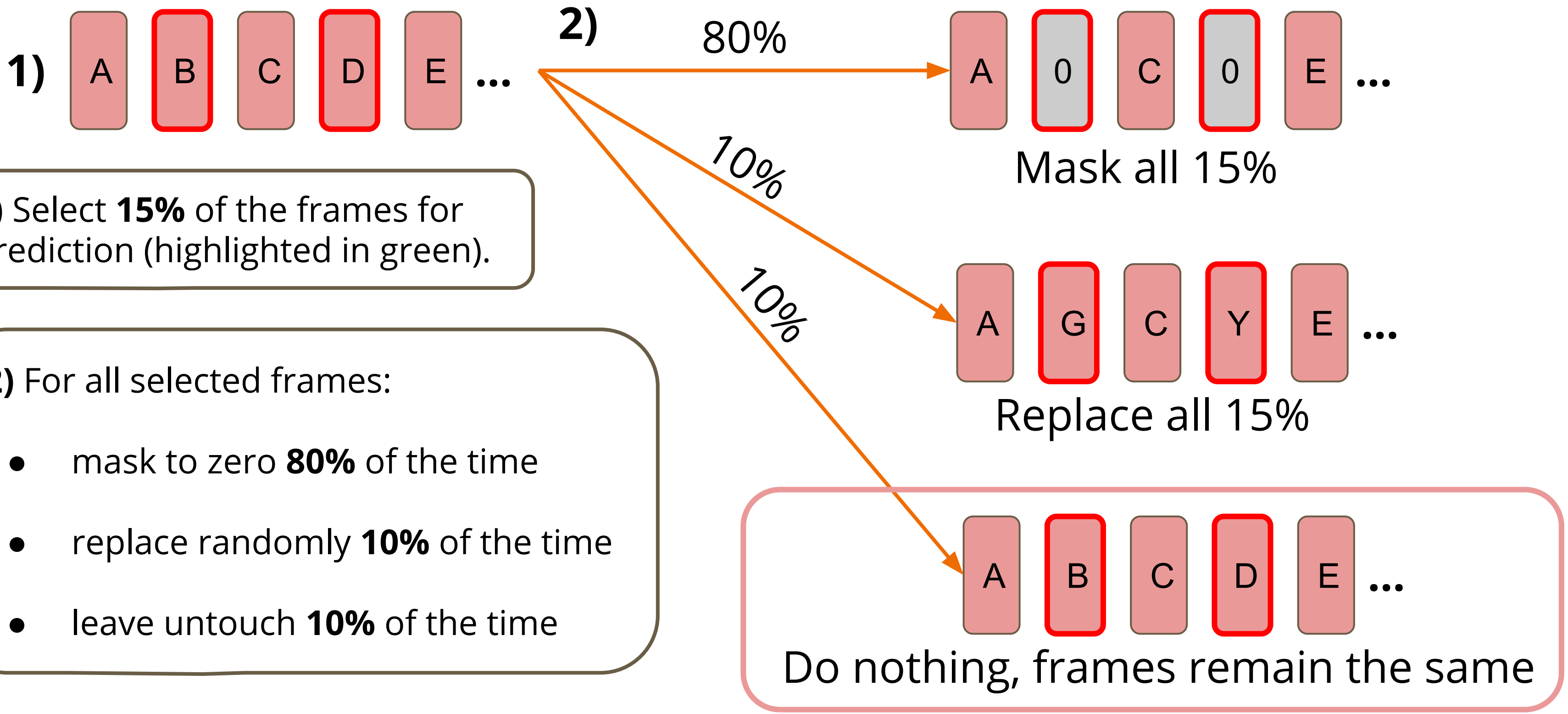


1) Select **15%** of the frames for prediction (highlighted in green).

2) For all selected frames:

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

Probabilistic Policy for Masking Frames

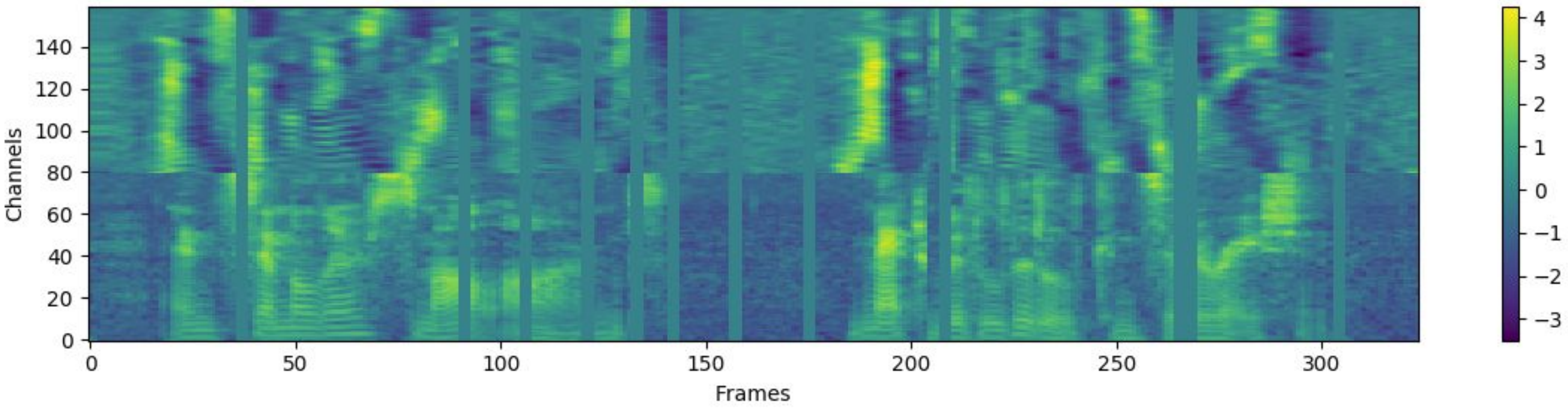


1) Select **15%** of the frames for prediction (highlighted in green).

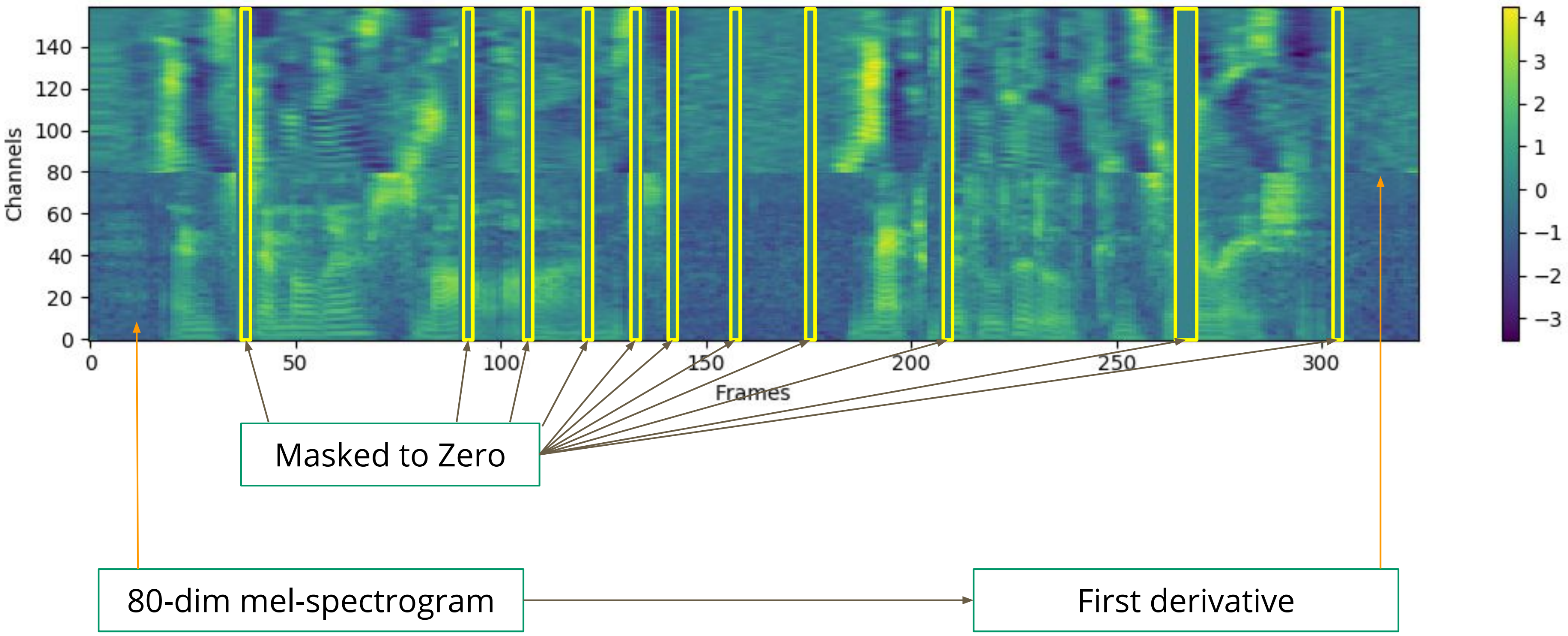
2) For all selected frames:

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

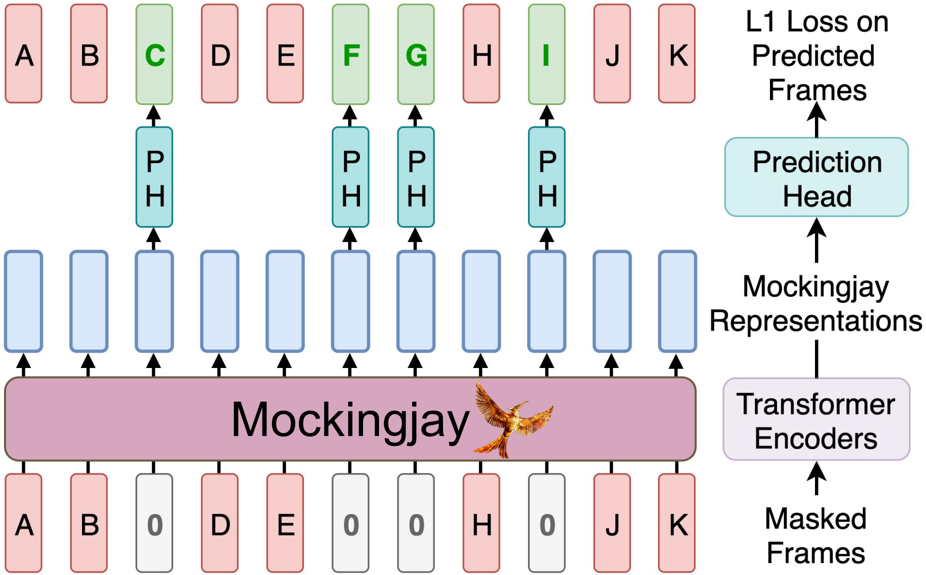
Input Feature: Masked Spectrogram

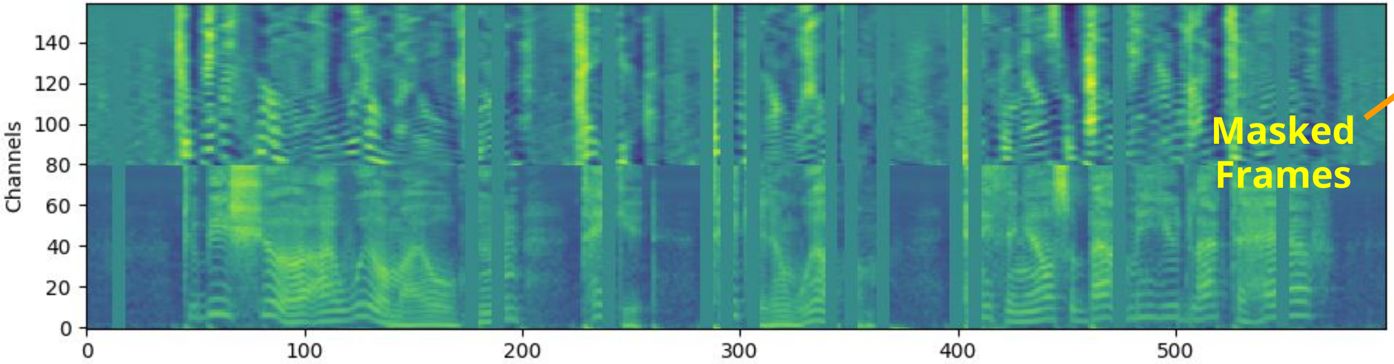


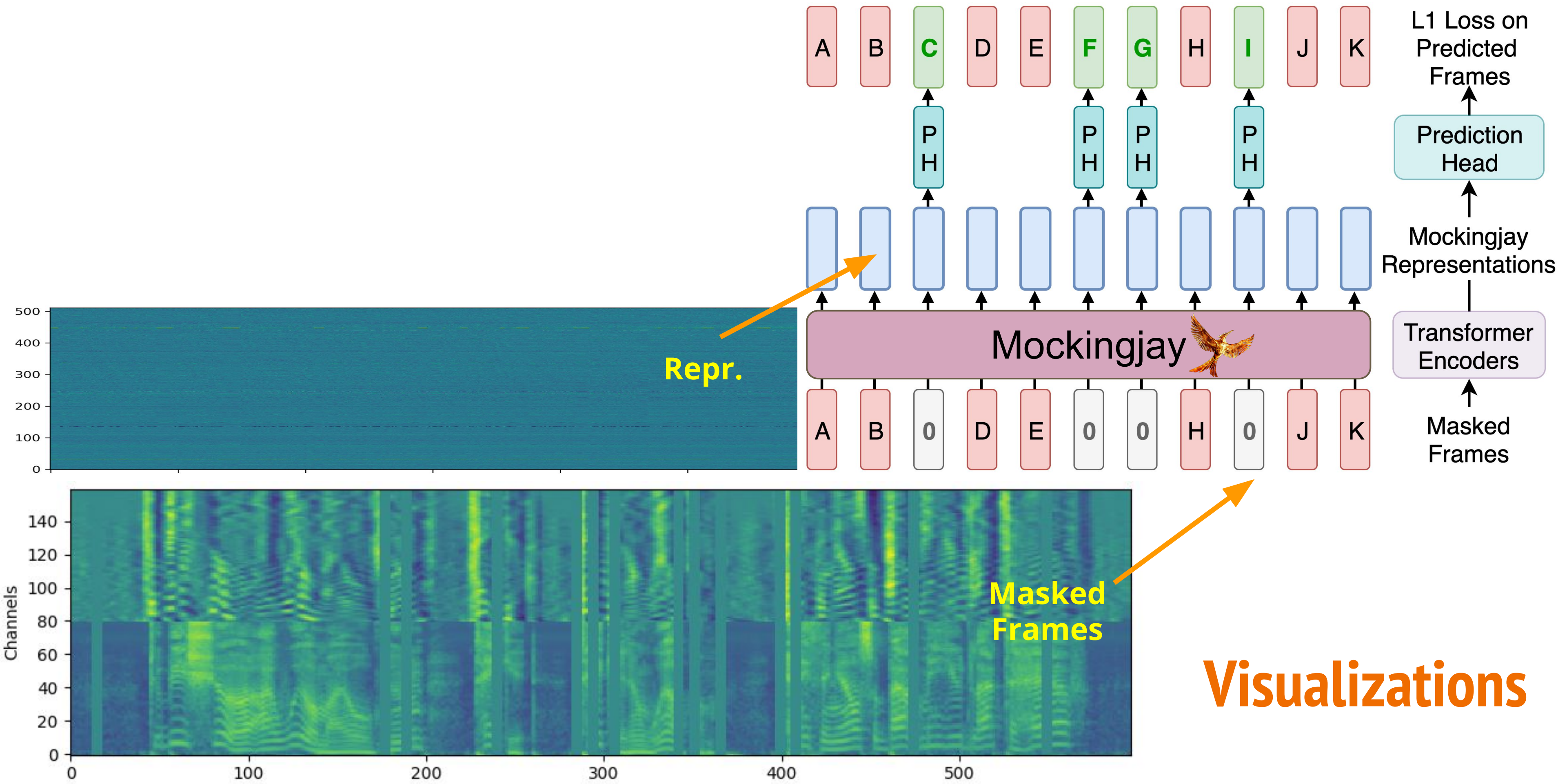
Input Feature: Masked Spectrogram



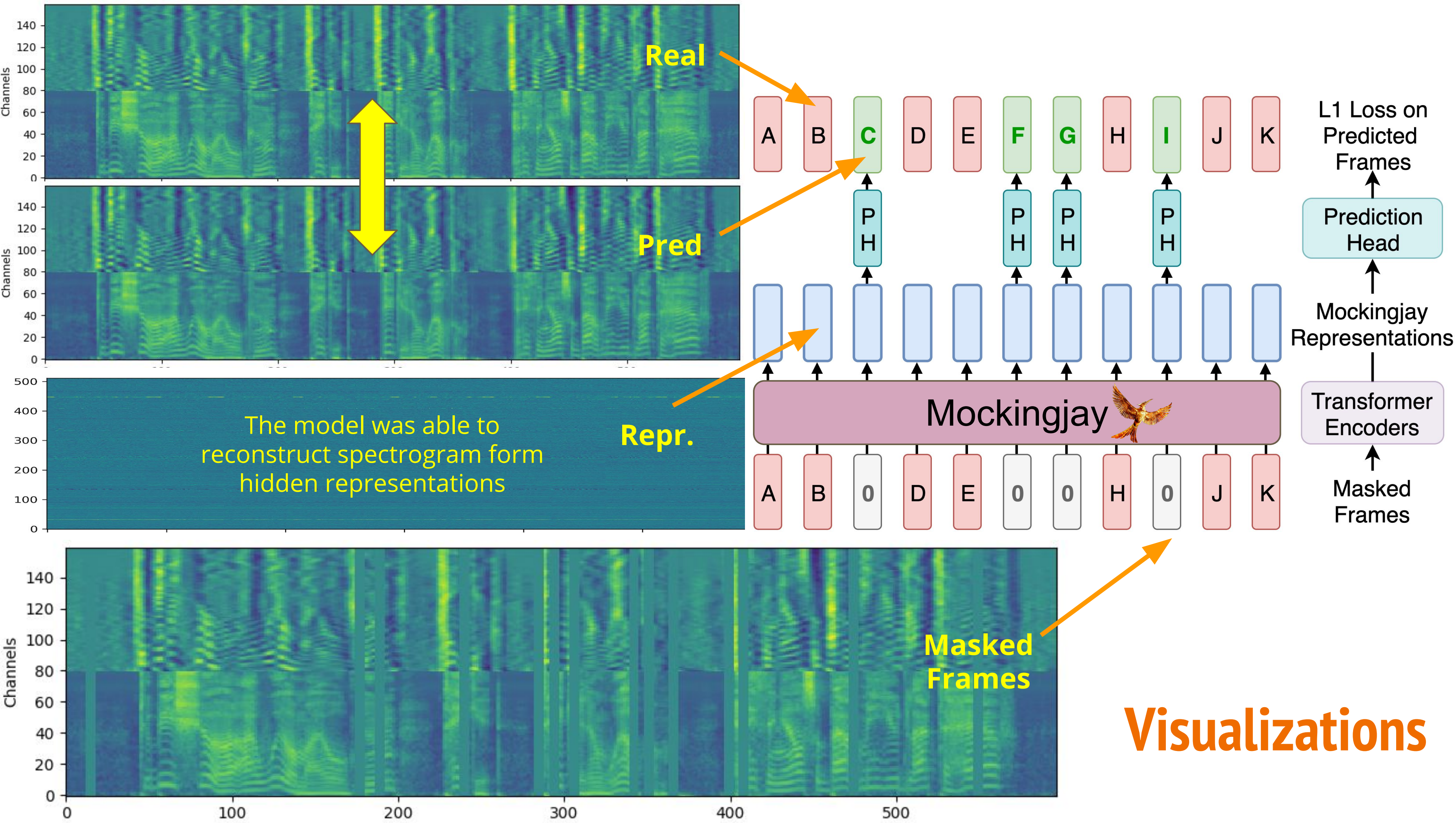
Visualizations






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

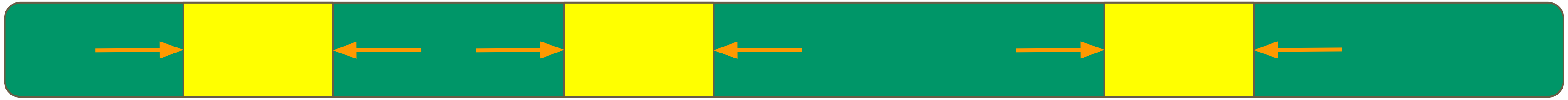
Visualizations



Migrating from text to speech

Acoustic Features: long and locally smooth in nature,

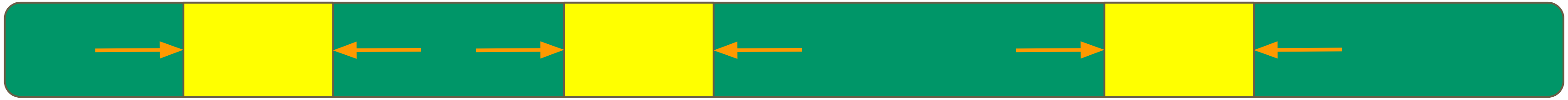
need to 1) shorten the sequence and 2) mask over a longer span



Migrating from text to speech

Acoustic Features: long and locally smooth in nature,

need to 1) shorten the sequence and 2) mask over a longer span



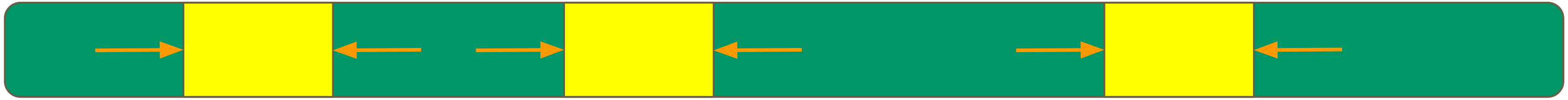
Address the long and smooth problem with:

Downsampling, and *consecutive masking*

Migrating from text to speech

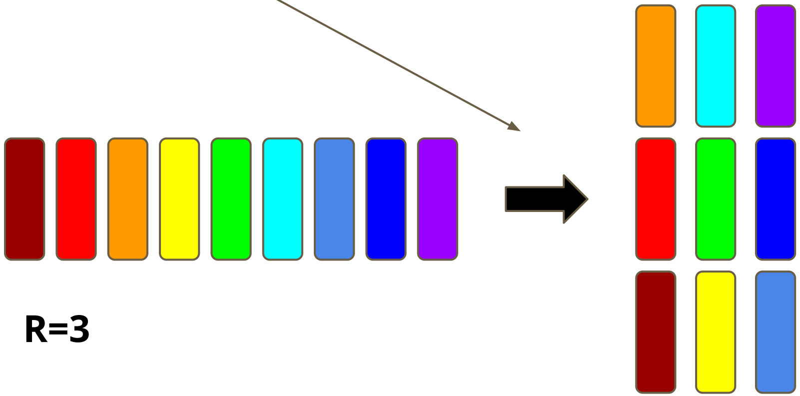
Acoustic Features: long and locally smooth in nature,

need to 1) shorten the sequence and 2) mask over a longer span



Address the long and smooth problem with:

Downsampling, and *consecutive masking*

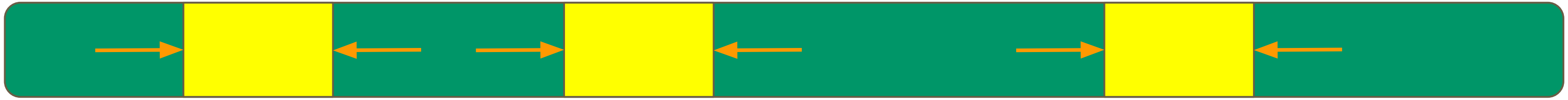


R=3

Migrating from text to speech

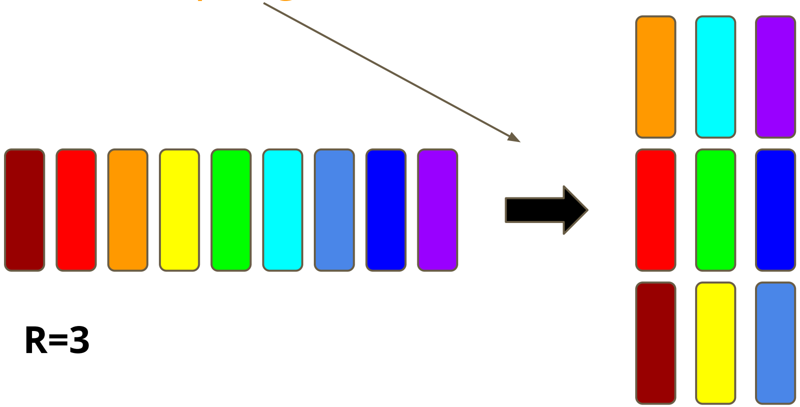
Acoustic Features: long and locally smooth in nature,

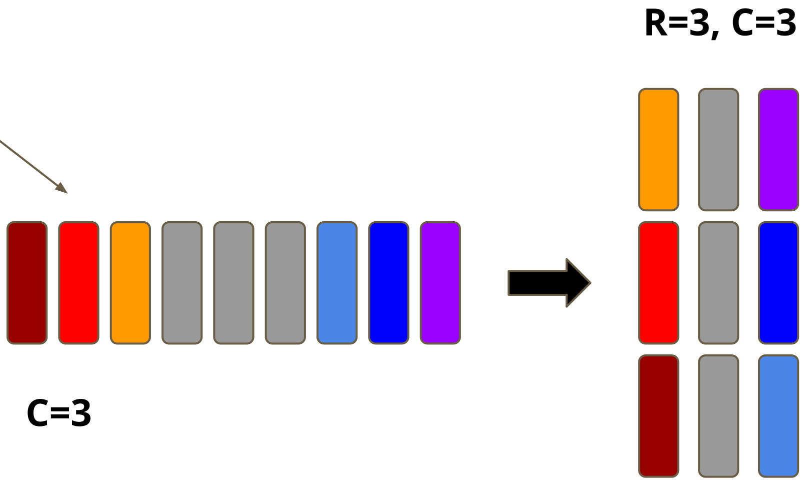
need to 1) shorten the sequence and 2) mask over a longer span



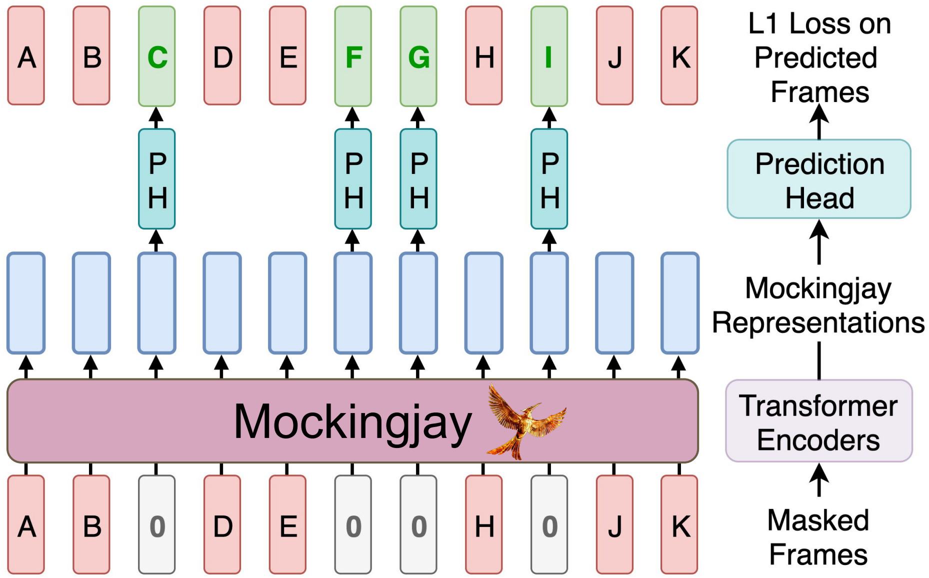
Address the long and smooth problem with:

Downsampling, and *consecutive masking*

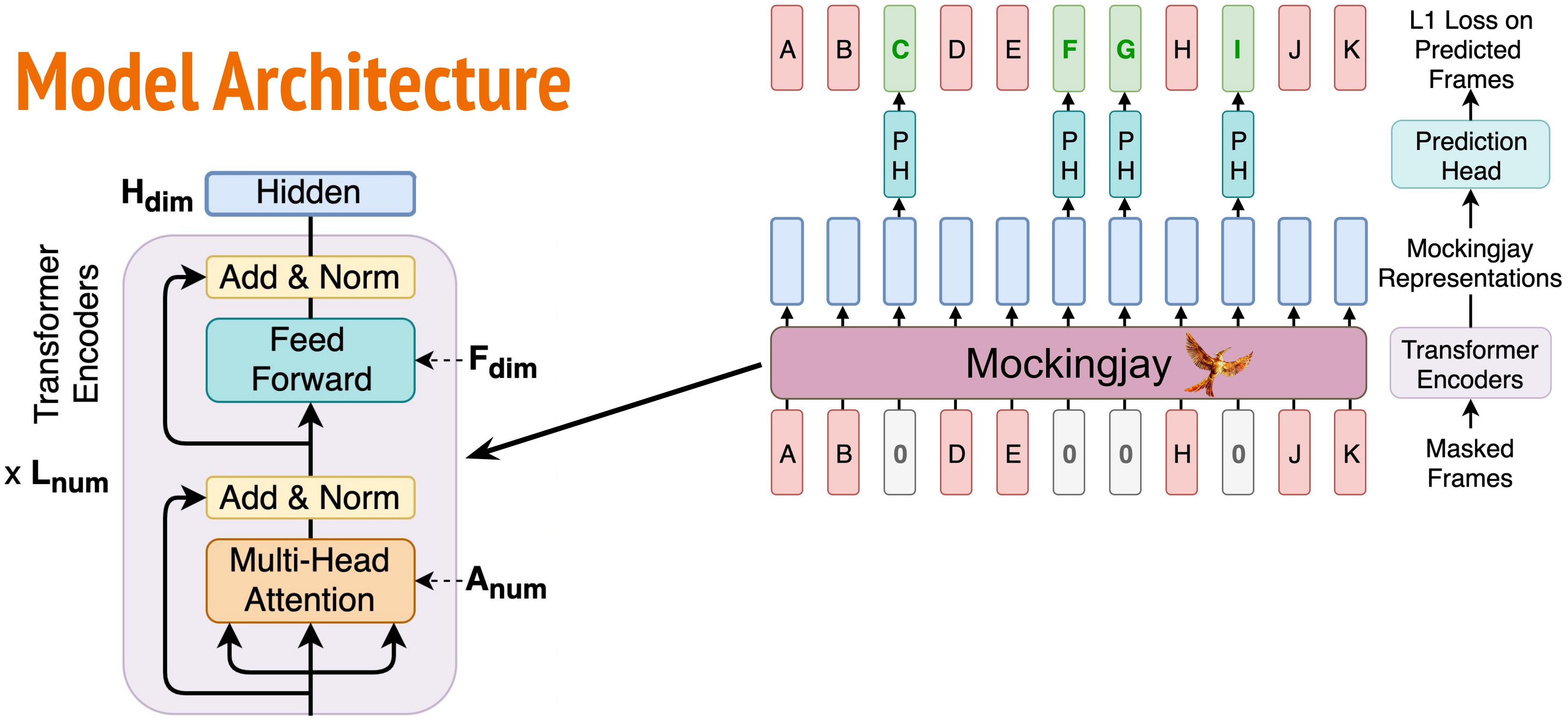




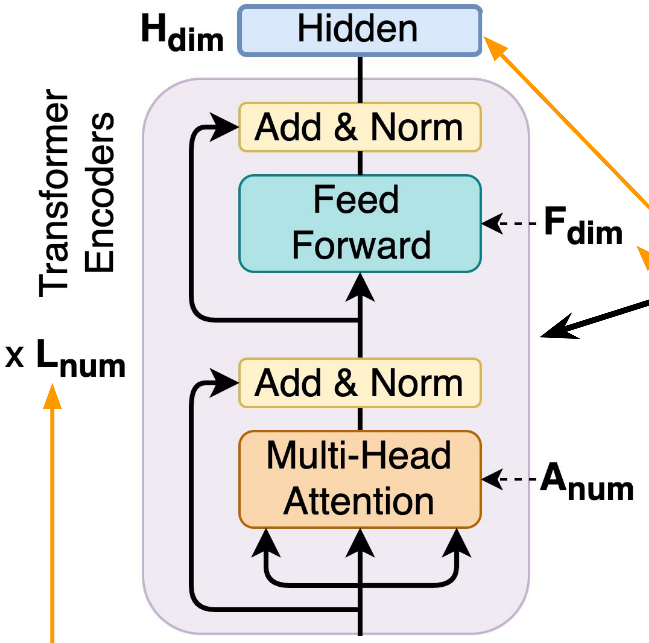
Model Architecture

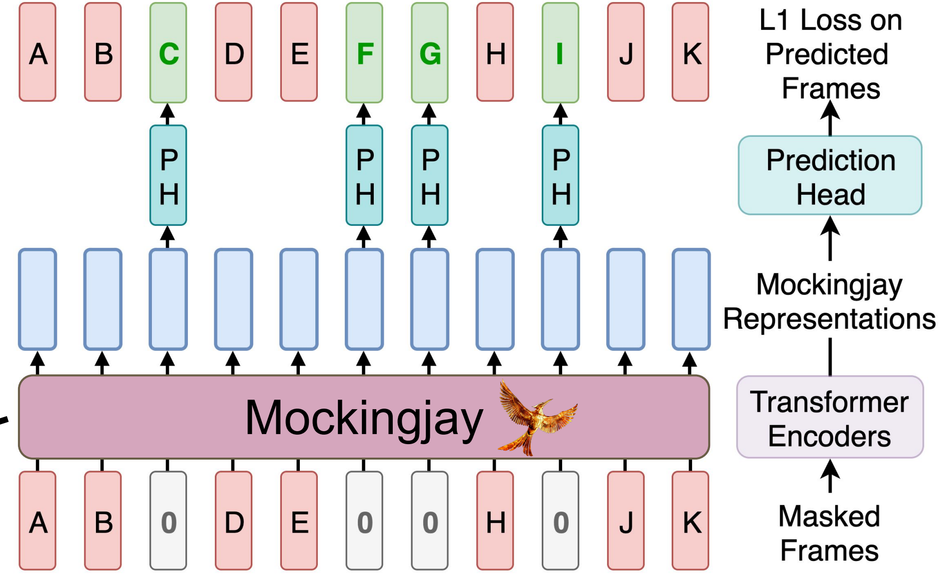


Model Architecture



Model Architecture





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

- BASE (L=3)
- LARGE (L=12)

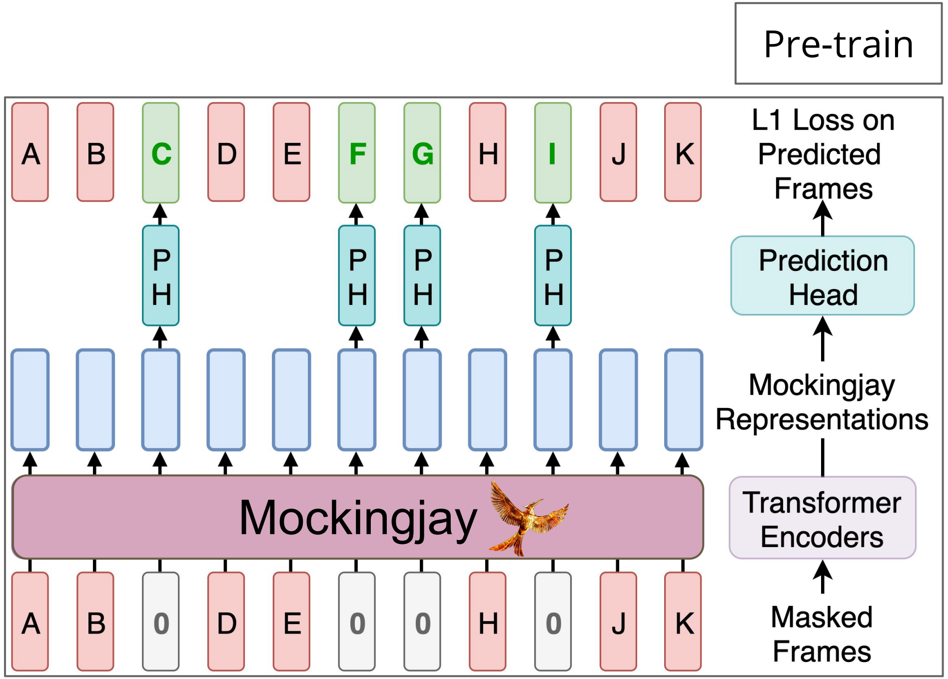
- Train on LibriSpeech **360 hrs**
- Pre-train steps = **500k**
- Fine-tune steps = **50k** (2-epochs)

Incorporating with Downstream Tasks

1) Feature Extraction

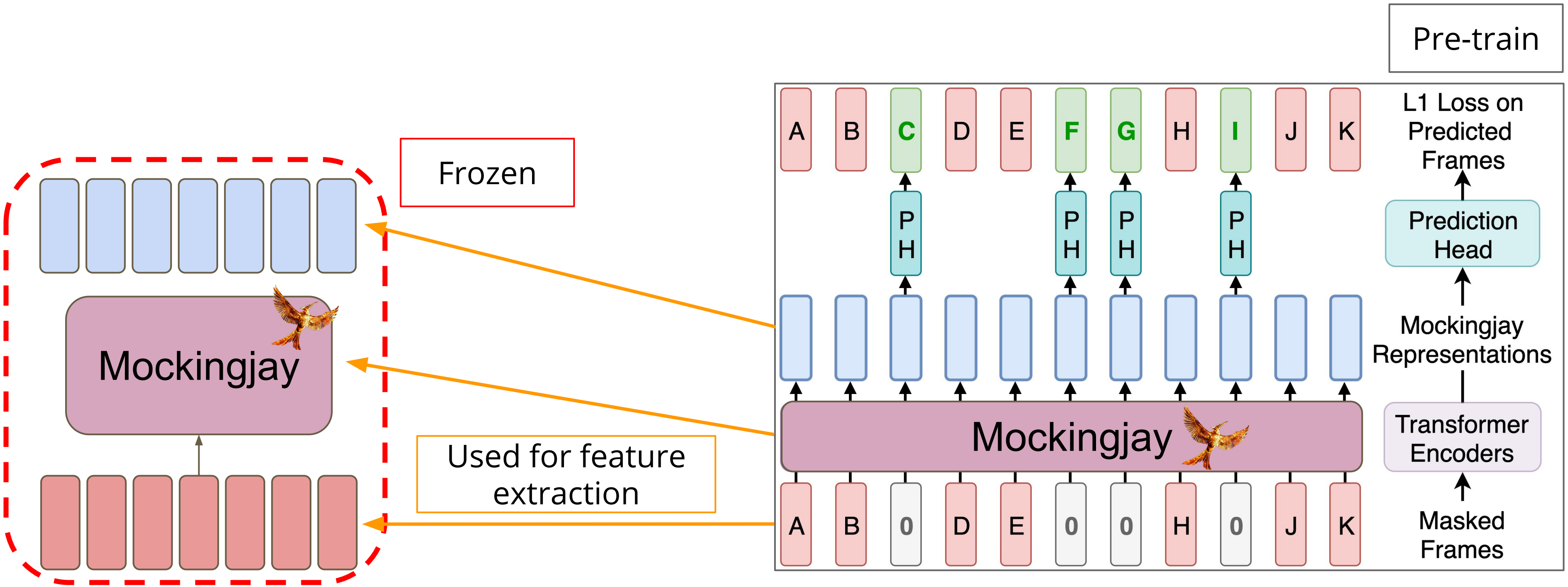
Incorporating with Downstream Tasks

1) Feature Extraction



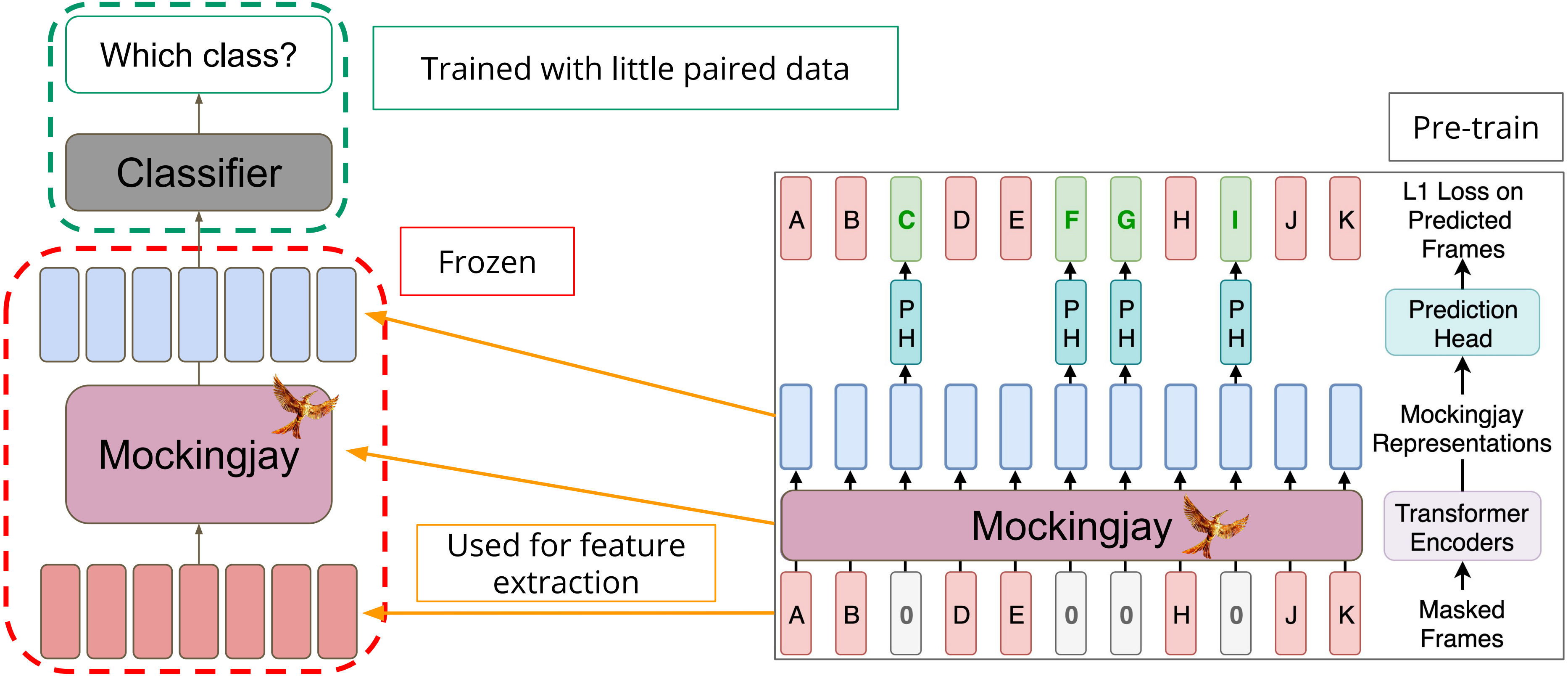
Incorporating with Downstream Tasks

1) Feature Extraction



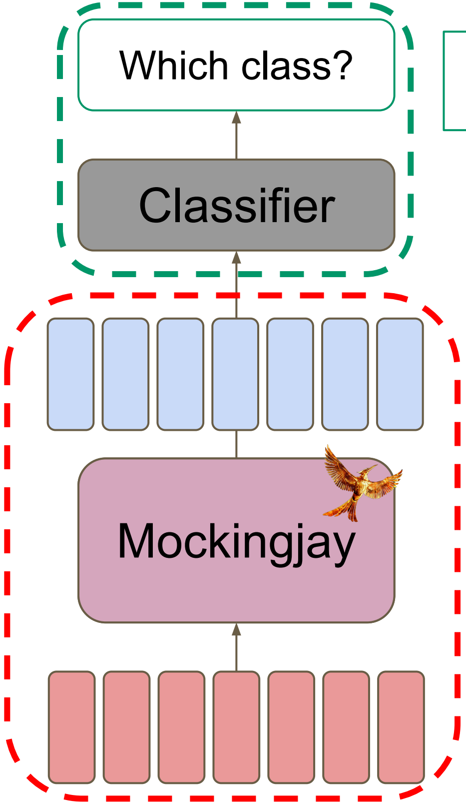
Incorporating with Downstream Tasks

1) Feature Extraction



Incorporating with Downstream Tasks

2) Weighted Sum from All Layers (WS)

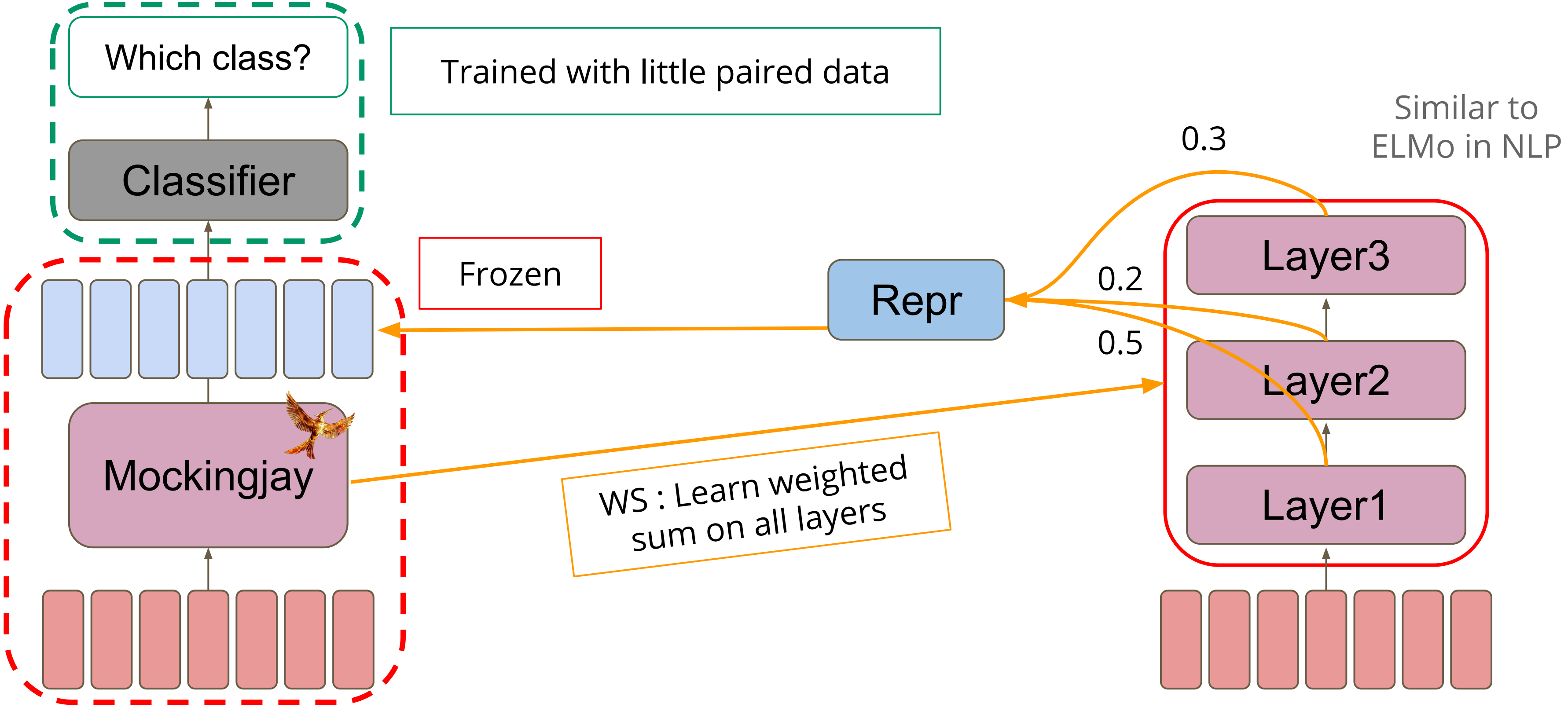


Trained with little paired data

Frozen

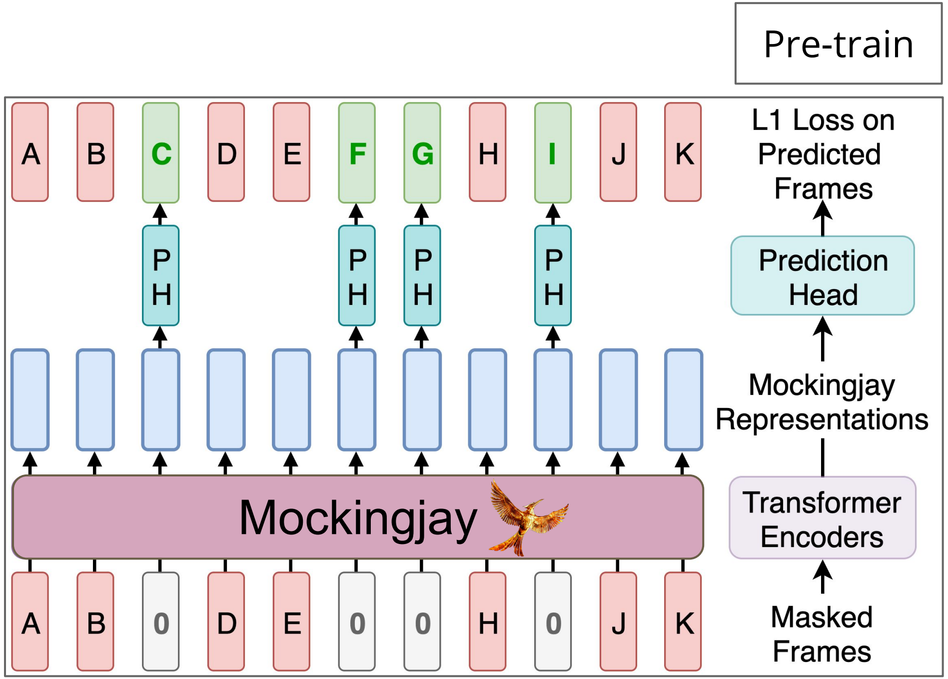
Incorporating with Downstream Tasks

2) Weighted Sum from All Layers (WS)



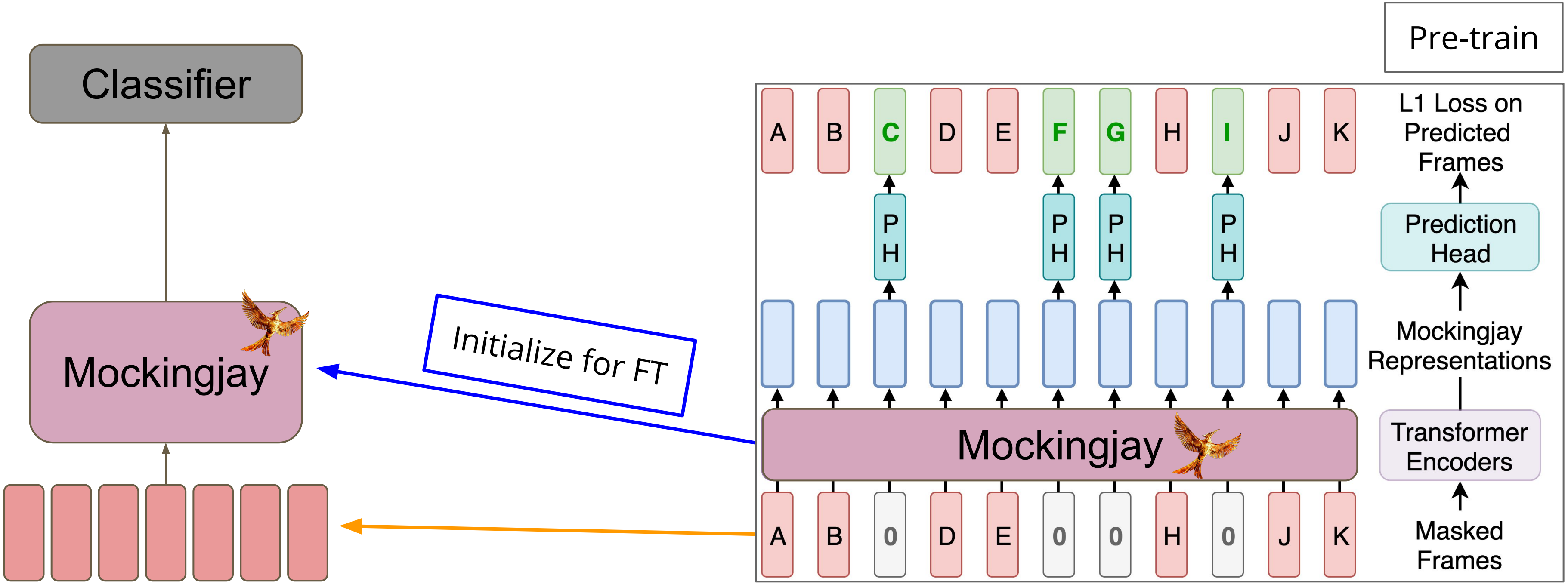
Incorporating with Downstream Tasks

3) Fine-tune (FT2)



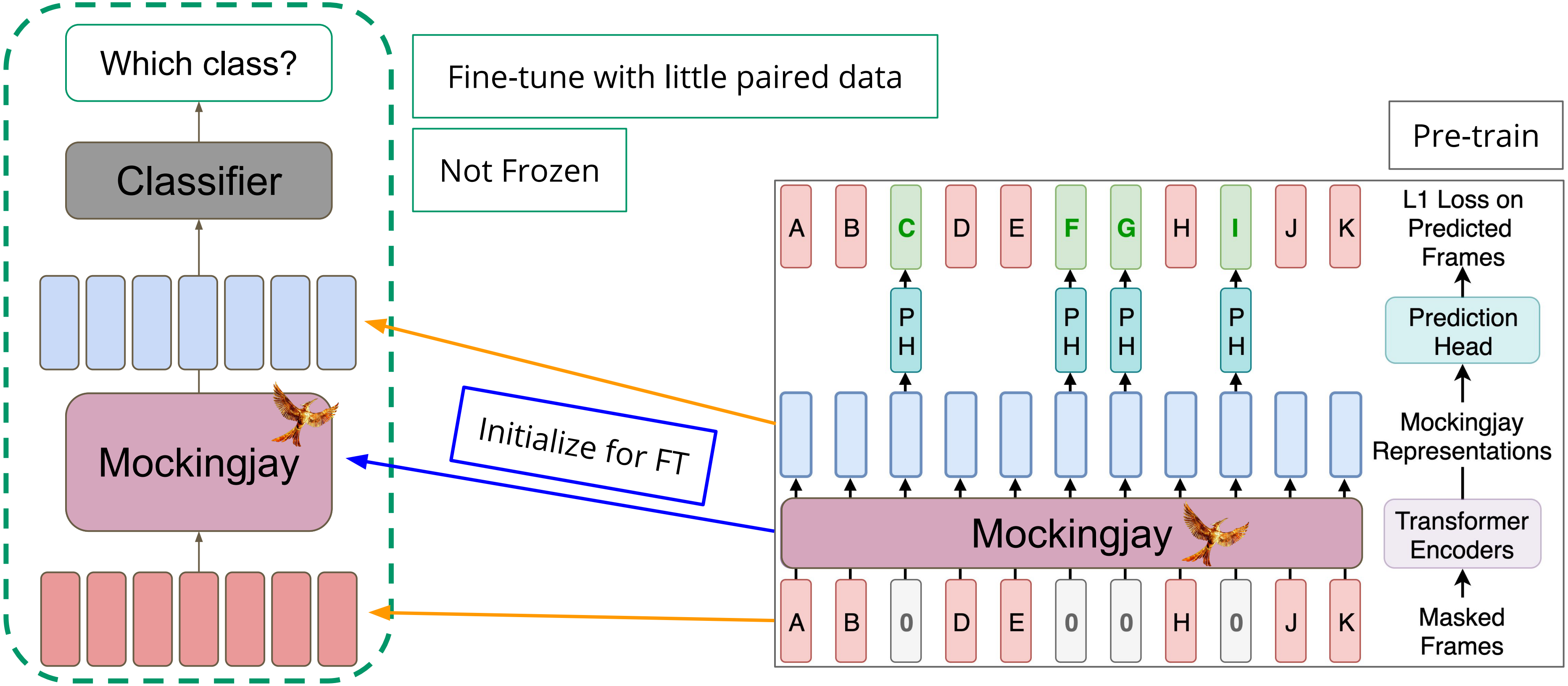
Incorporating with Downstream Tasks

3) Fine-tune (FT2)



Incorporating with Downstream Tasks

3) Fine-tune (FT2)



Experiments

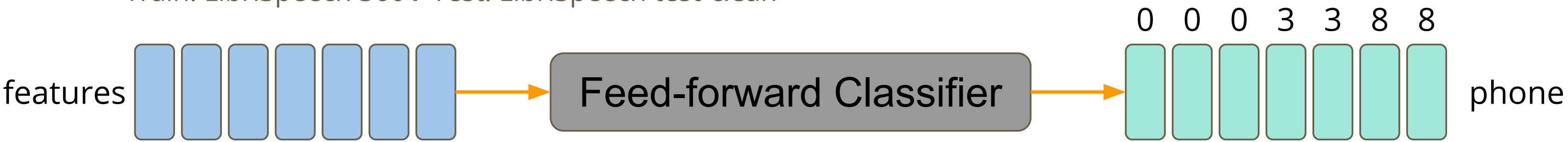
We report results on 3 different downstream tasks:

- Phoneme Classification
- Speaker Recognition
- Sentiment Classification on spoken content

Experiments

We report results on 3 different downstream tasks:

- Phoneme Classification (72 classes):
  Train: LibriSpeech 360 / Test: LibriSpeech test-clean



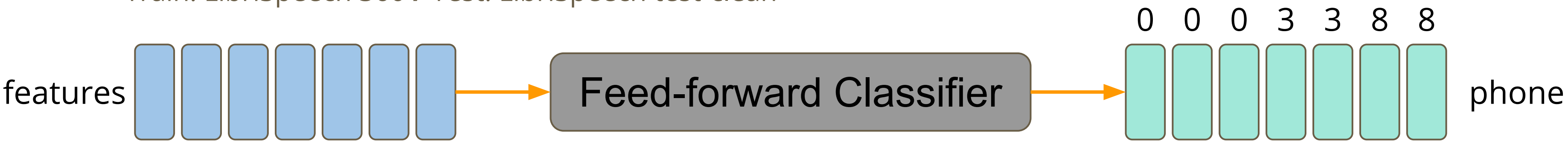
- Speaker Recognition
- Sentiment Classification on spoken content

Experiments

[3] Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph

We report results on 3 different downstream tasks:

- Phoneme Classification (72 classes):
  Train: LibriSpeech 360 / Test: LibriSpeech test-clean



- Speaker Recognition (63 classes):
  Train: 90% of LibriSpeech 100 / Test: 10% of LibriSpeech 100

- Sentiment Classification on spoken content (2 classes):
  To demonstrate domain invariant transferability, we use another dataset: MOSEI [3]

Experiments - 1/3

Acoustic Features	Phoneme Classification	Speaker Recognition	Sentiment Classification
Mel Features	49.1	70.1	64.6
BASE	60.9	94.5	67.4
LARGE	**64.3**	**96.3**	**70.1**

Consistent results over all three tasks:
Mel < BASE < LARGE

Experiments - 2/3

Acoustic Features	Phoneme Classification	Speaker Recognition	Sentiment Classification
Mel Features	49.1	70.1	64.6
BASE	60.9	94.5	67.4
LARGE	64.3	96.3	70.1
LARGE-WS	**69.9**	**96.4**	**71.1**

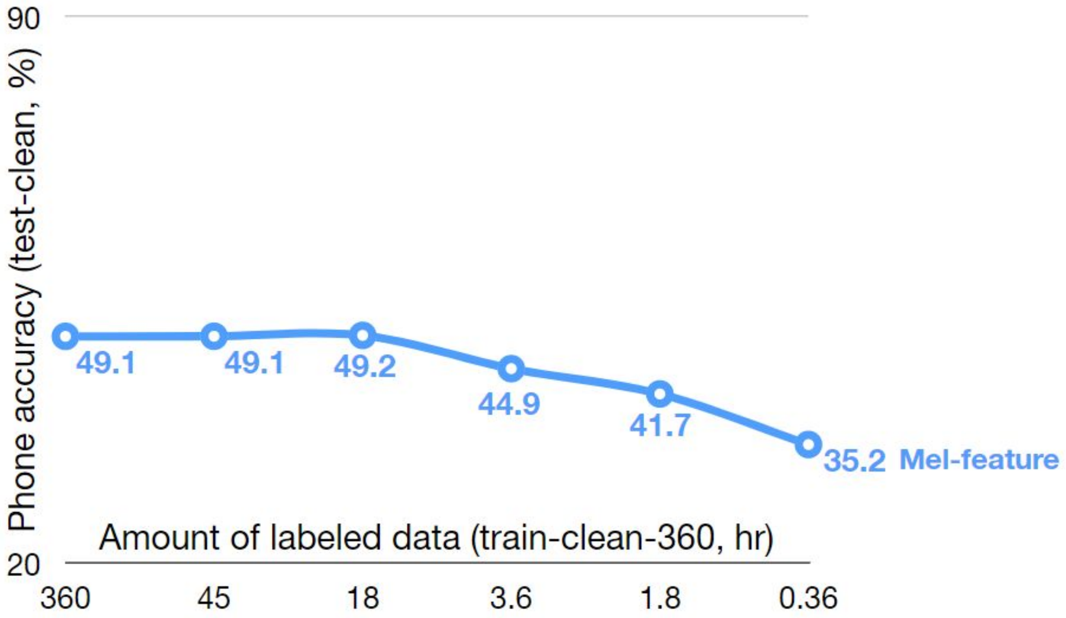
Consistent results over all three tasks:
LARGE < LARGE-WS

Experiments - 3/3

Acoustic Features	Phoneme Classification	Speaker Recognition	Sentiment Classification
Mel Features	49.1	70.1	64.6
BASE	60.9	94.5	67.4
LARGE	64.3	96.3	70.1
LARGE-WS	69.9	96.4	**71.1**
BASE-FT2	**84.3**	**98.1**	68.5
APC [2]	74.1	85.9	66.0

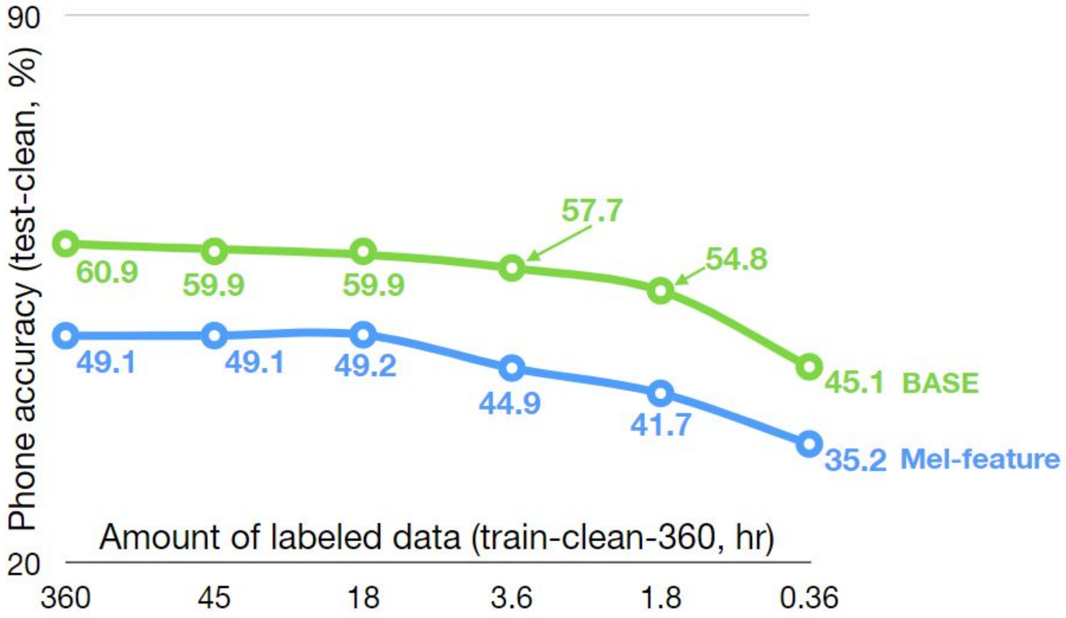
[2] An Unsupervised Autoregressive Model for Speech Representation Learning

Low-Resource Experiments - 1/6



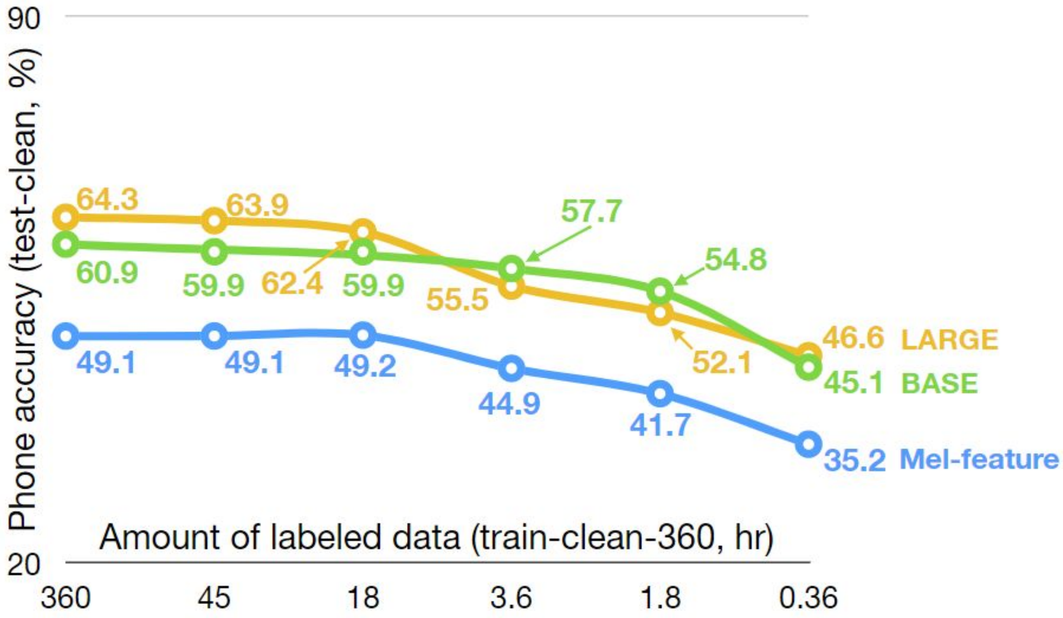
We demonstrate how pre-training on speech can improve supervised training in low resource scenarios, we train with reduced amount of labels.

Low-Resource Experiments - 2/6



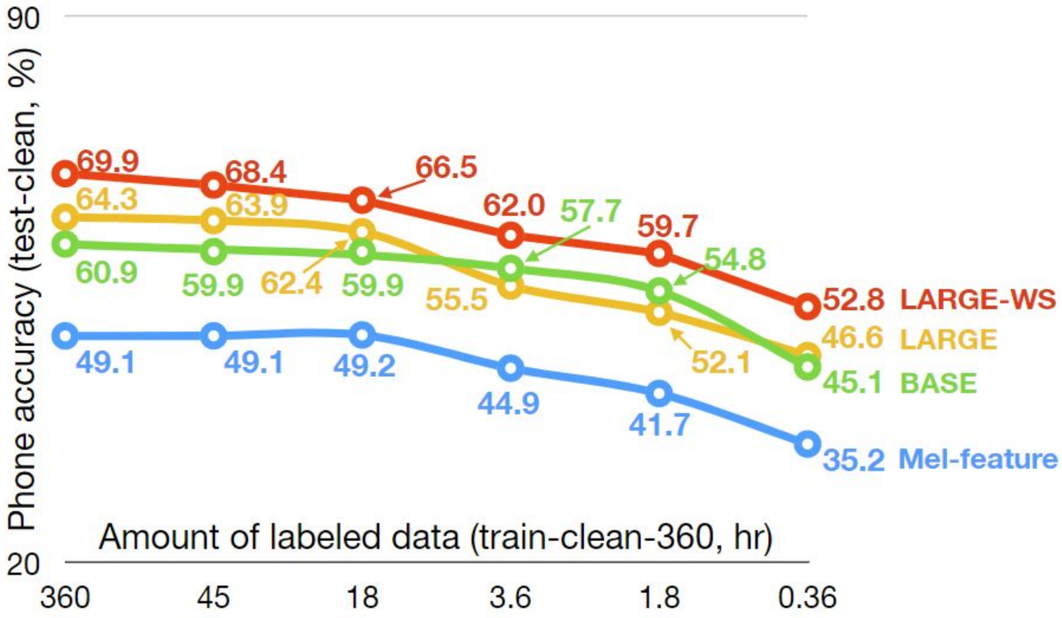
Mel < BASE

Low-Resource Experiments - 3/6



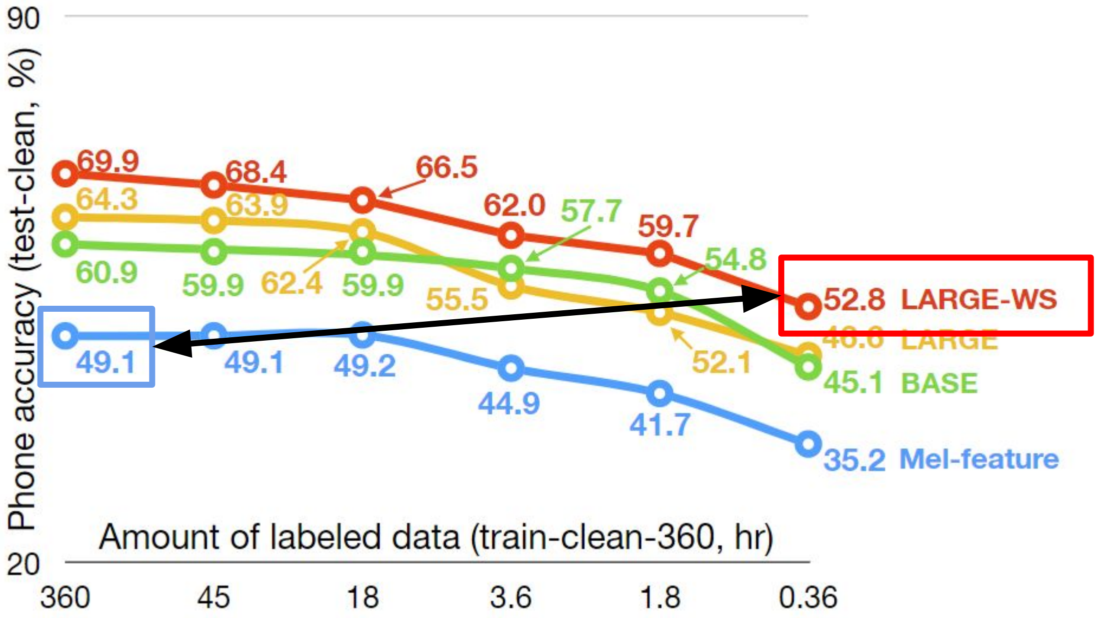
Mel < BASE < LARGE

Low-Resource Experiments - 4/6



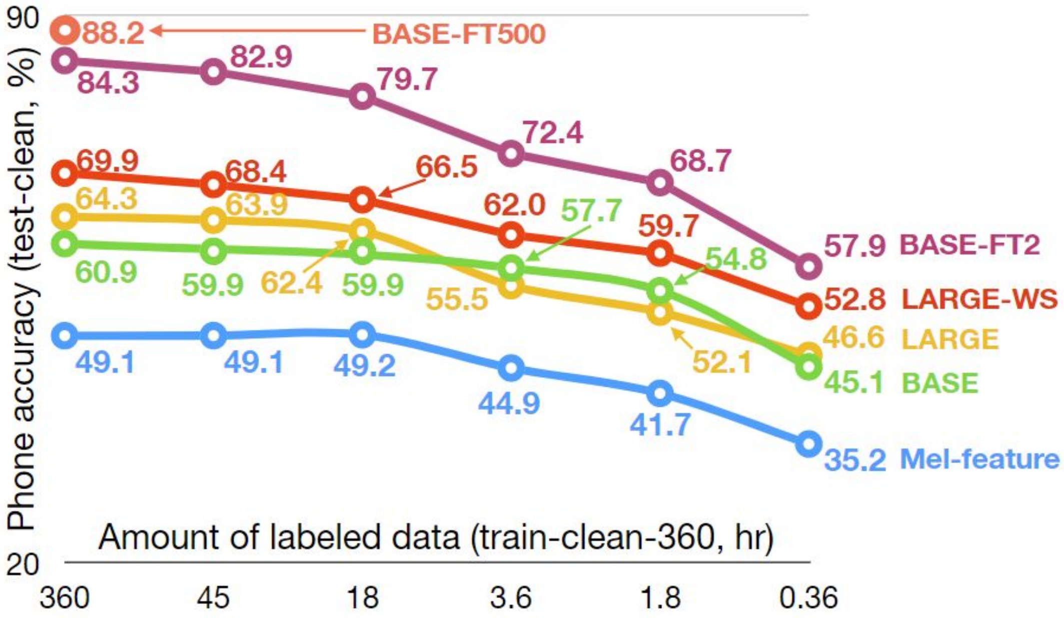
LARGE < LARGE-WS
with an avg 5.75% improvement

Low-Resource Experiments - 4/6



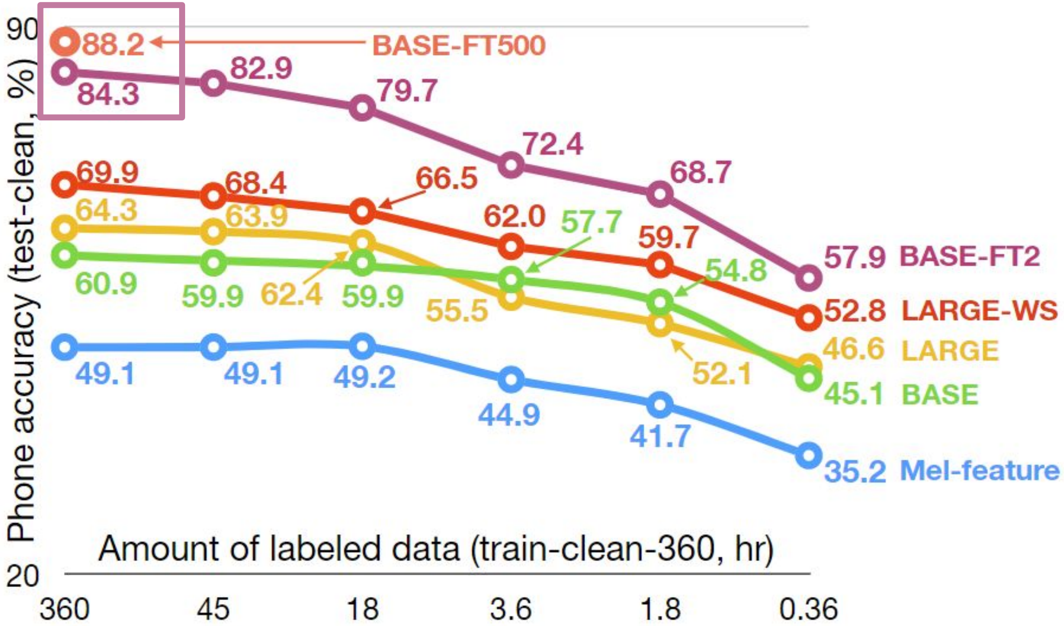
With 0.1% of labels,
LARGE-WS (52.8%) outperformed Mel (49.1%) that uses all 100% hours of labeled data.

Low-Resource Experiments - 5/6



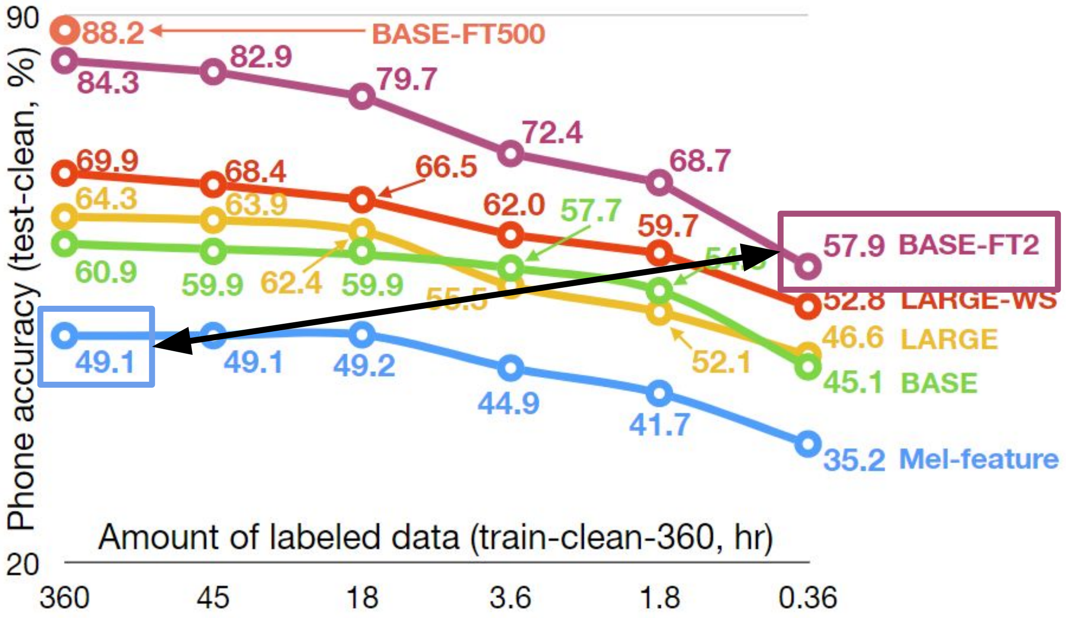
All < BASE-FT2

Low-Resource Experiments - 5/6



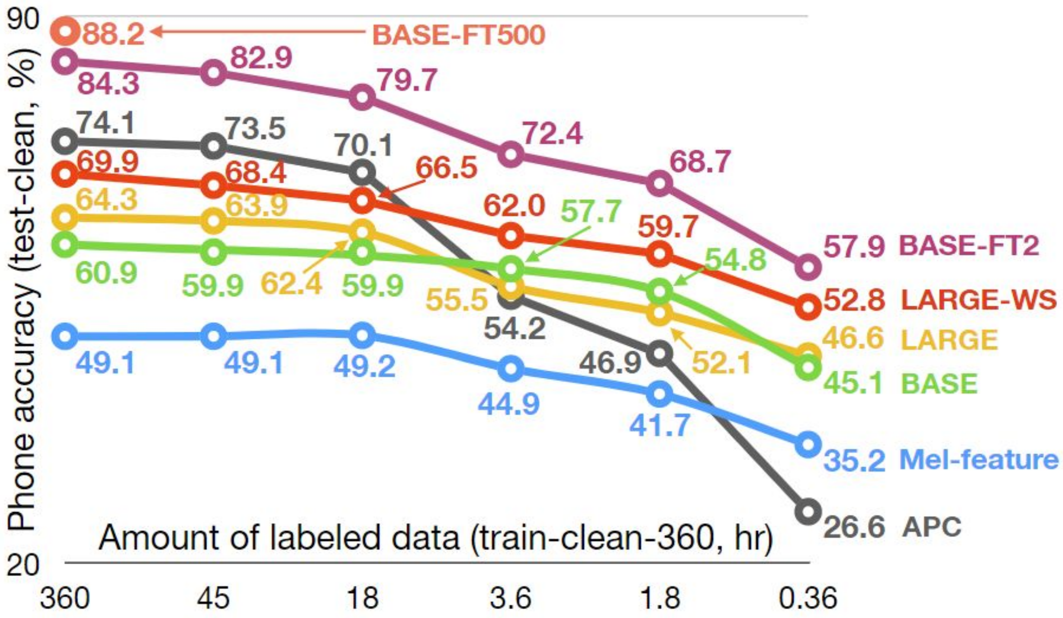
All < BASE-FT2

Low-Resource Experiments - 5/6



With 0.1% of labels,
BASE-FT2 (57.9%) outperformed Mel (49.1%) that uses all 100% hours of labeled data.

Low-Resource Experiments - 6/6



APC works well on full resource but fails to generalize for limited labeled data.

Conclusion

We conclude that unsupervised Mockingjay improves supervised training!

Links

This slide (with speaker notes) can be found here:
https://bit.ly/icassp2020-mockingjay

Our code and implementation can be found here:
https://github.com/andi611/Mockingjay-Speech-Representation