Self-Supervised Learning

For Speech

Andy T. Liu

2020/06/03

Overview

What is Self-Supervised Learning?

An analogy

How do Infants Learn?

Can Machine do the same?

Looks at a lot of books!

Listens to a lot of conversations!

Then their parents teach them.

How do Infants Learn? Yes! Self-Supervised Learning

Unlabeled Text

Unlabeled Speech

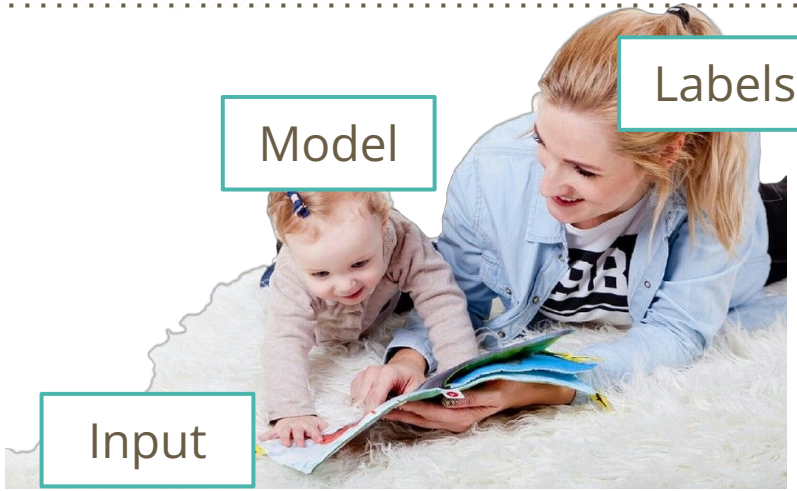
Self-Supervised Pre-training

How do Infants Learn? Yes! Self-Supervised Learning

Unlabeled Text

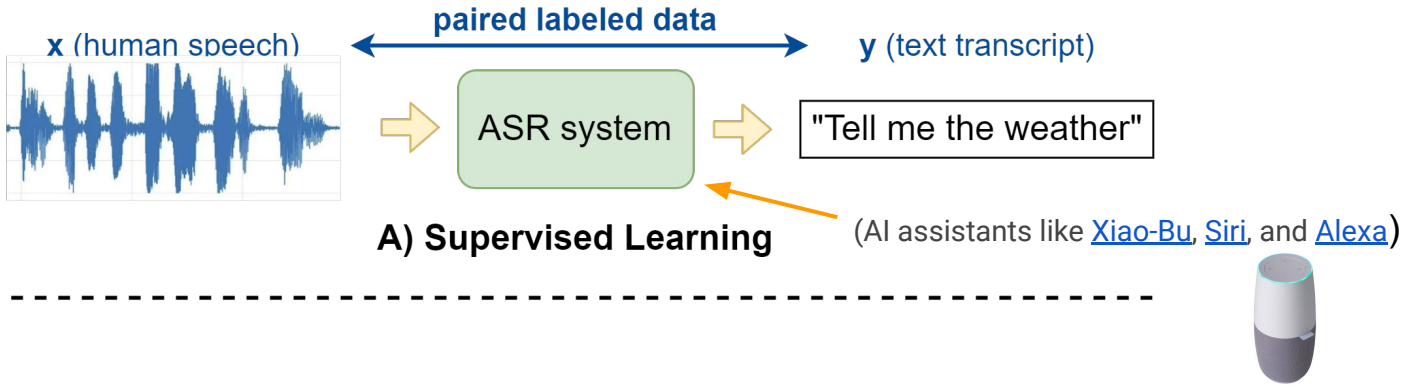
Unlabeled Speech

Self-Supervised Pre-training



Supervised Learning

Self-Supervised Learning for Speech



Self-Supervised Learning for Speech

x (human speech)

paired labeled data

y (text transcript)

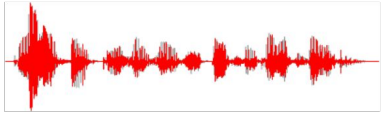
ASR system

"Tell me the weather"

A) Supervised Learning

(AI assistants like Xiao-Bu, Siri, and Alexa)

x' (lots of human speech)

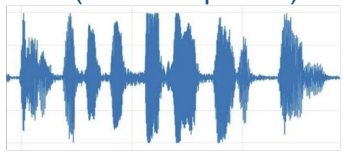


Pre-trained Model

leverage the knowledge of unlabeled data

Pre-trained models are evaluated on downstream tasks.

x (human speech)



ASR system

y (text transcript)

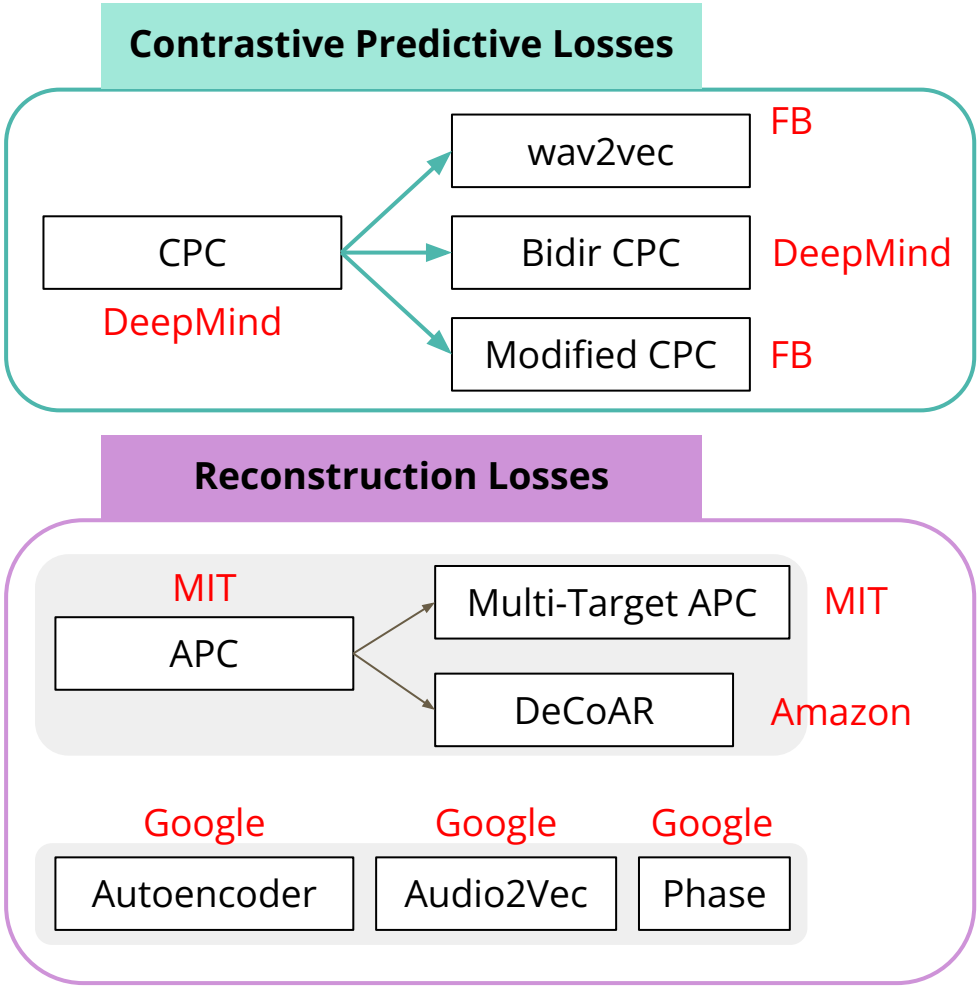
"Tell me the weather"

B) Self-Supervised Learning for Improving Supervised Systems

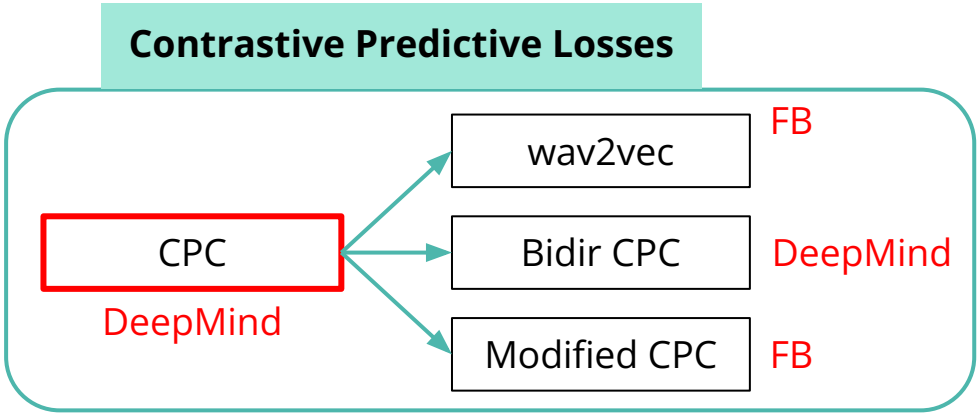
A Broad Introduction

of All Recent Methods

A Broad Introduction



A Broad Introduction



A Broad Introduction

Contrastive Predictive Losses

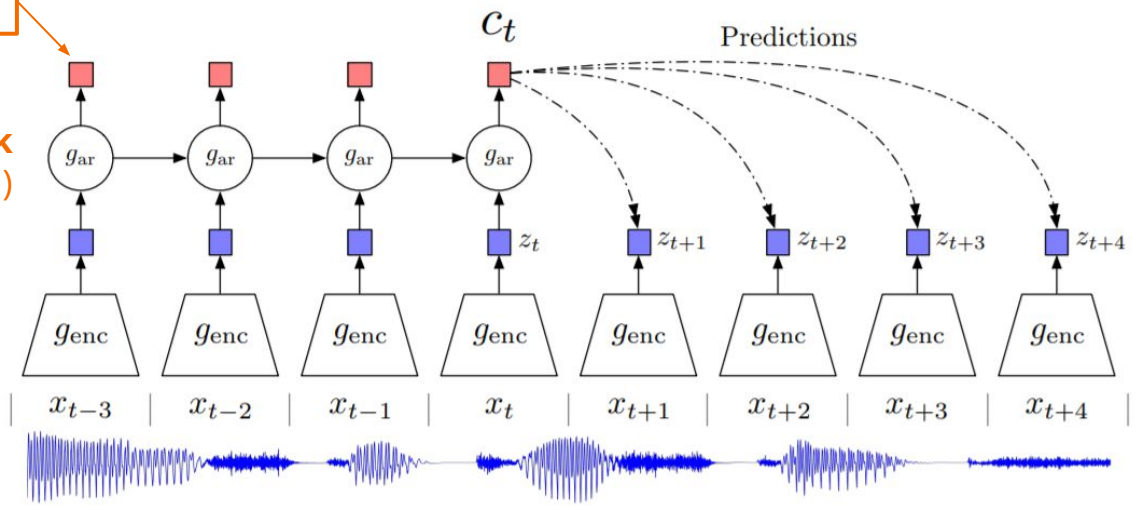
- CPC — DeepMind
- wav2vec — FB
- Bidir CPC — DeepMind
- Modified CPC — FB

Intuition:
Pulls temporally nearby representations closer and pushes temporally distant ones further.

Learned representations

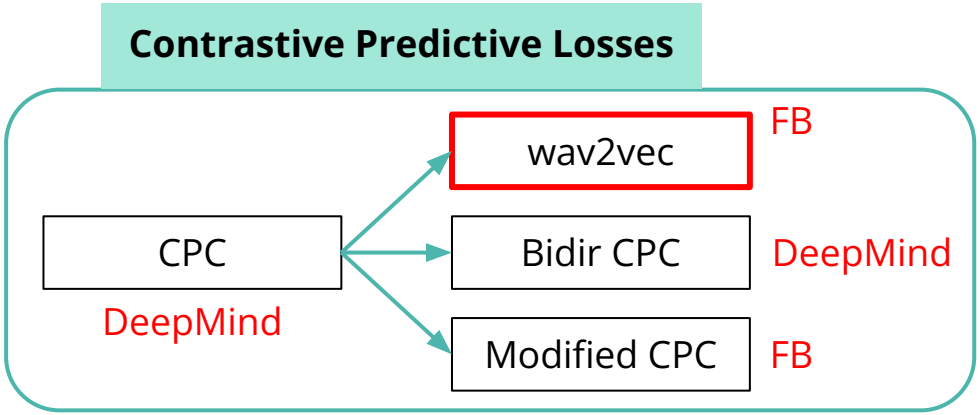
Context Network
1-layer Gated Recurrent Units (GRU)

Encoder Network
5-layer convolutional neural network (CNN)

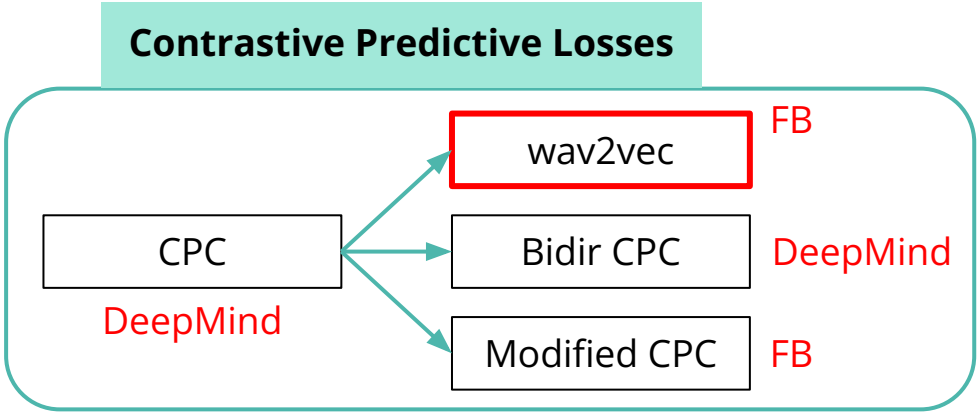


Important exps: Phone / Speaker Classification

A Broad Introduction



A Broad Introduction

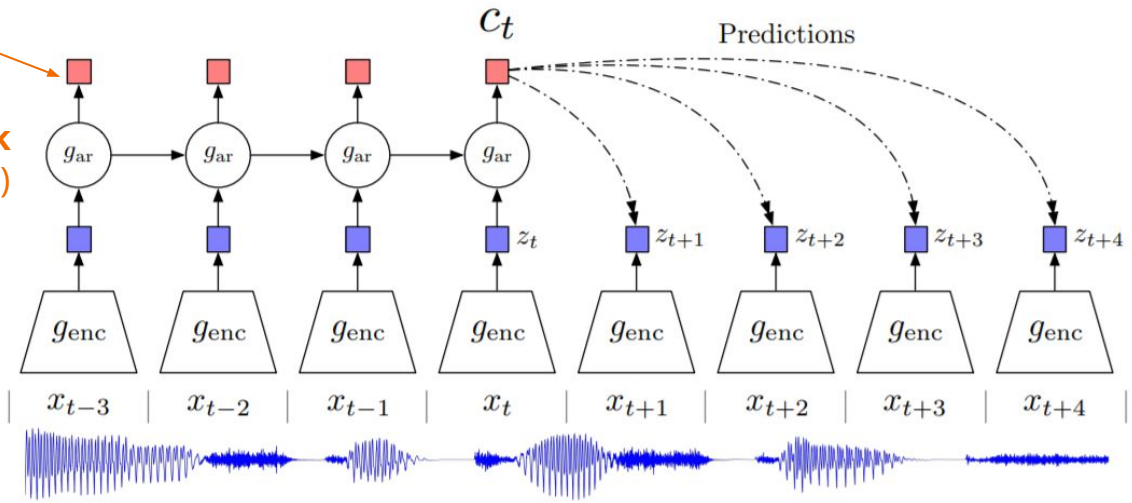


Contribution:
self-supervised pre-training is shown to improve supervised **ASR**

Used as input for ASR models, replace acoustic features

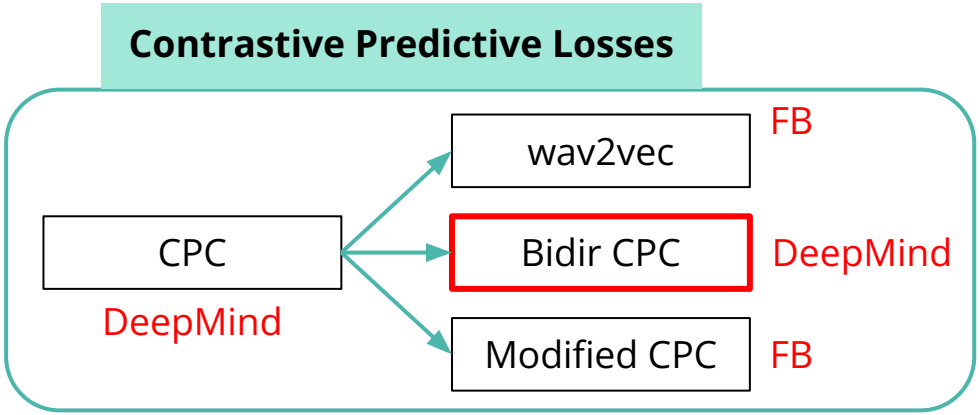
Context Network
9-layer convolutional neural network (CNN)

Encoder Network
5-layer convolutional neural network (CNN)

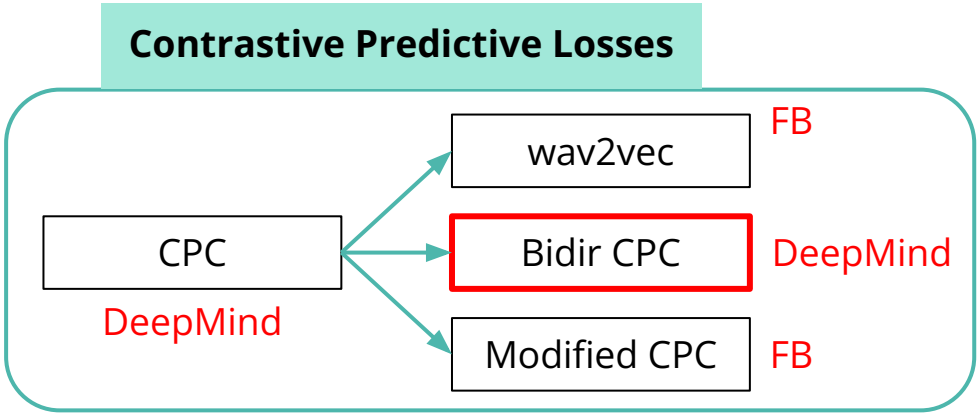


Important exps: ASR on WSJ / TIMIT

A Broad Introduction

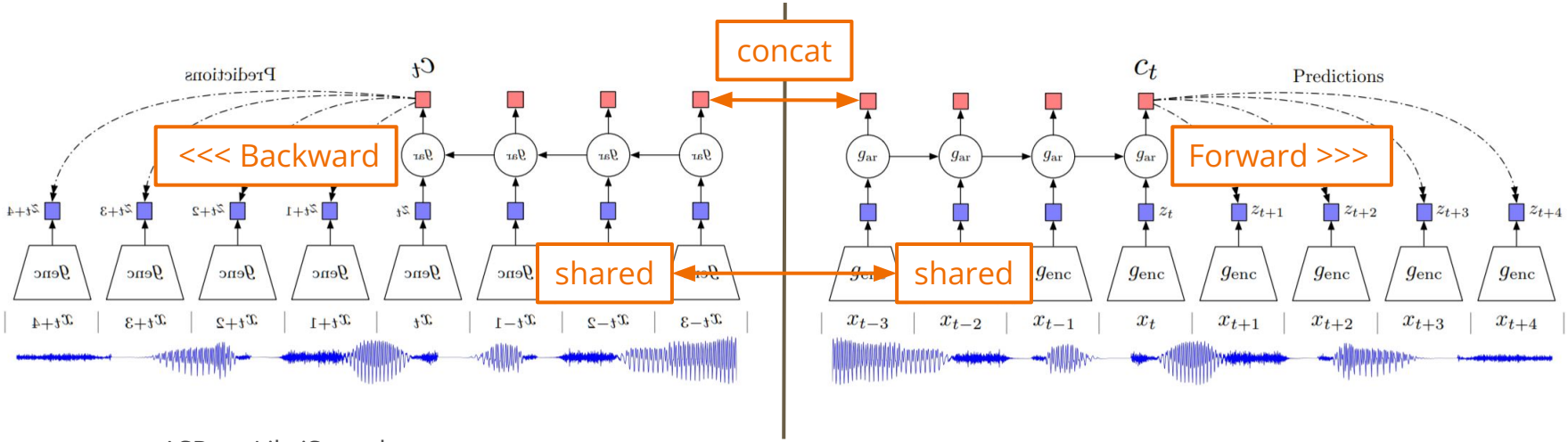


A Broad Introduction



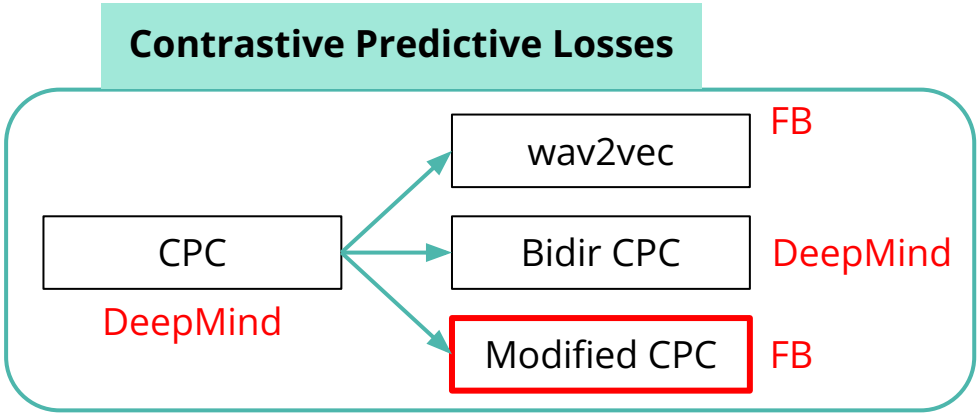
Contribution:
bidirectional context + ASR

learning representations from large amount of unlabeled data (8000 hrs) can provide improvements for out-of-domain transfer (different datasets / cross-lingual).

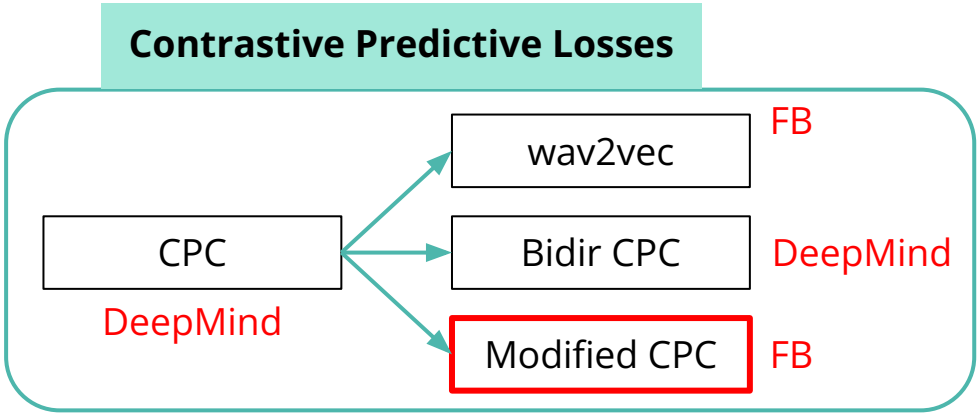


Important exps: ASR on LibriSpeech

A Broad Introduction



A Broad Introduction

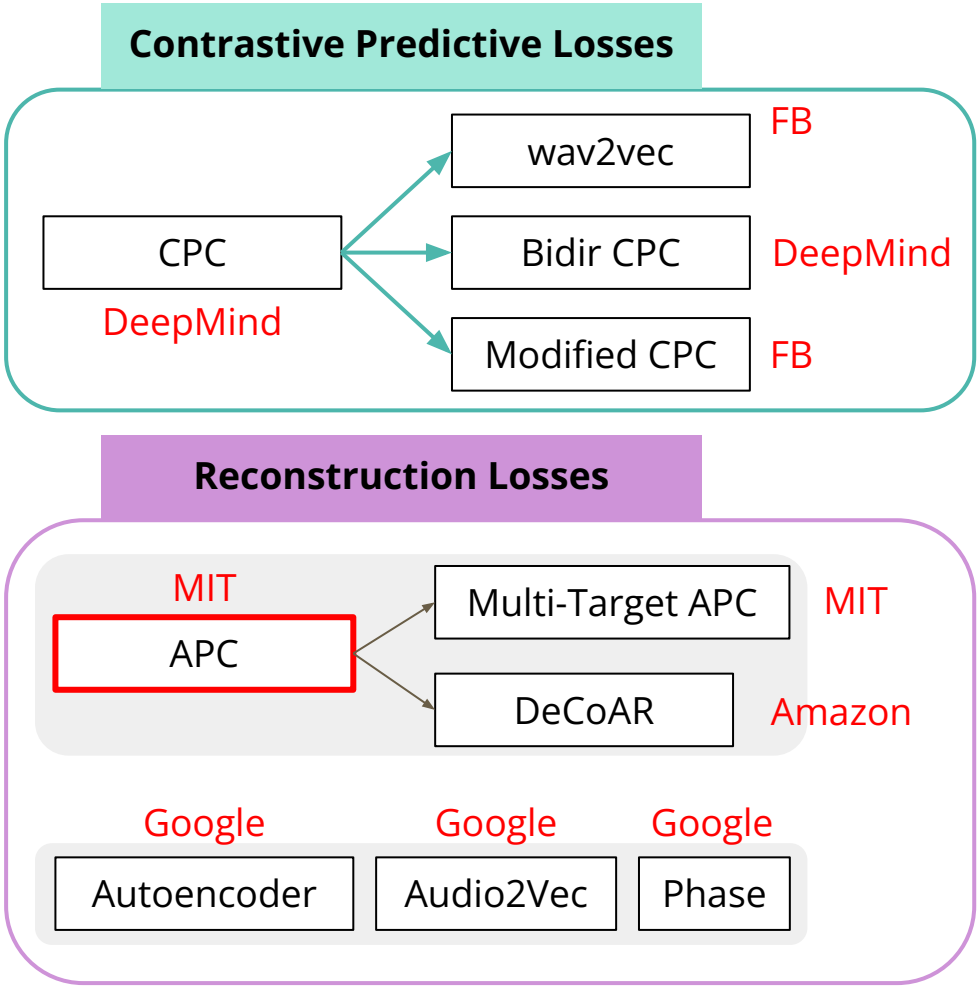


Contributions:

1) changing the batch normalization to channel-wise normalization
2) replace the linear prediction layer to a Transformer layer
3) and replacing the context network of GRUs with Long Short-Term Memory (LSTM) cells

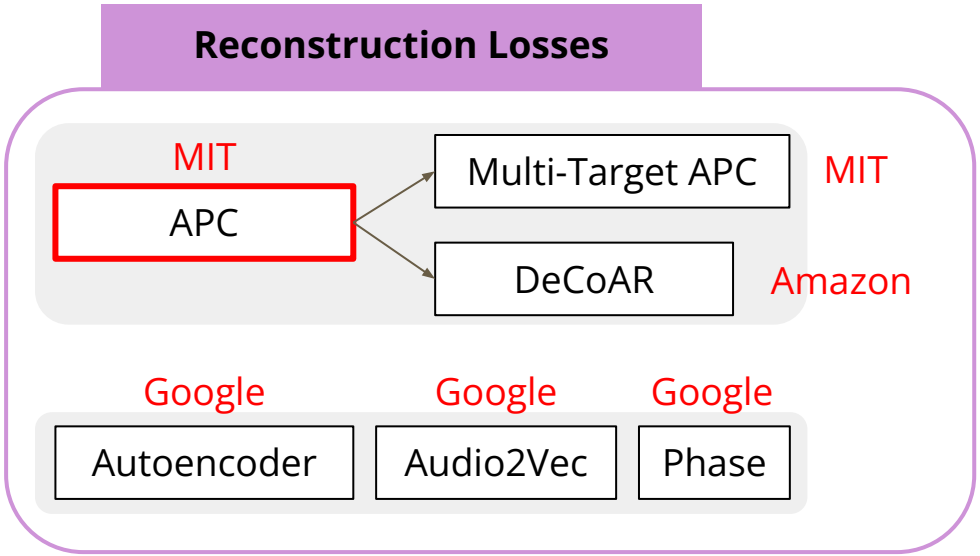
Important exps: Phone Classification

A Broad Introduction

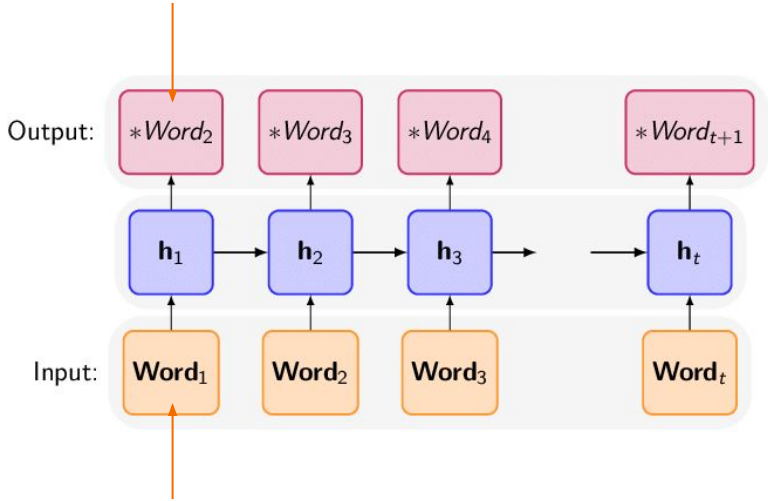


A Broad Introduction

Intuition:
Speech Version of a RNN Language Model

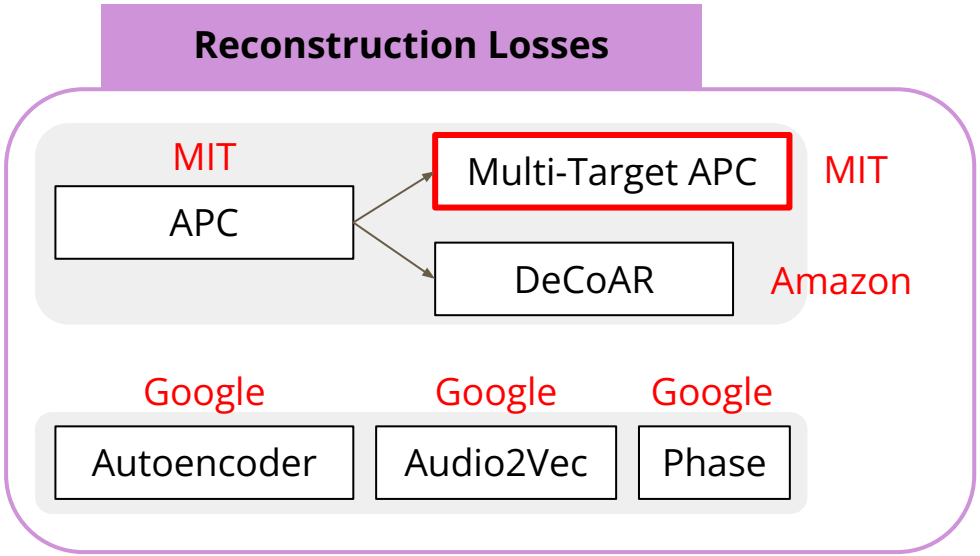


Change the softmax layer to
regression layer for reconstruction



Instead of operating on word tokens,
change them to acoustic frames

A Broad Introduction

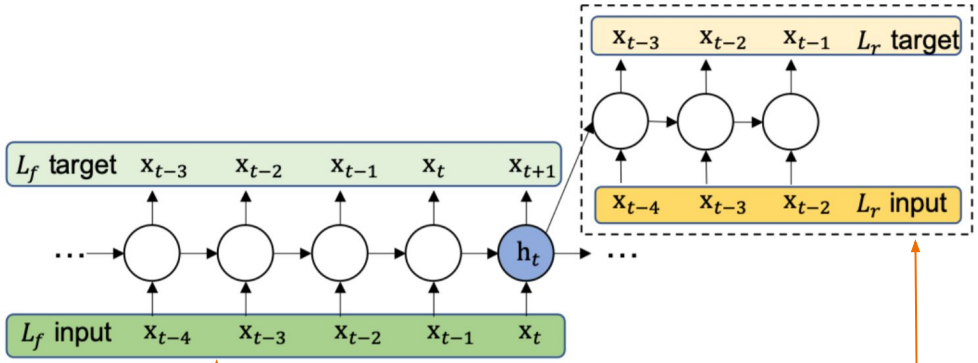


Important exps: Phone Classification, ASR on WSJ
They use settings that are not conventional.

A Broad Introduction

Intuition:
The APC objective is extended to bidirectional.

An auxiliary RNN is used to refresh current hidden states with the knowledge learned in the past, allowing the model to remember more from the past.



predicts the future frame conditioning on previous context,

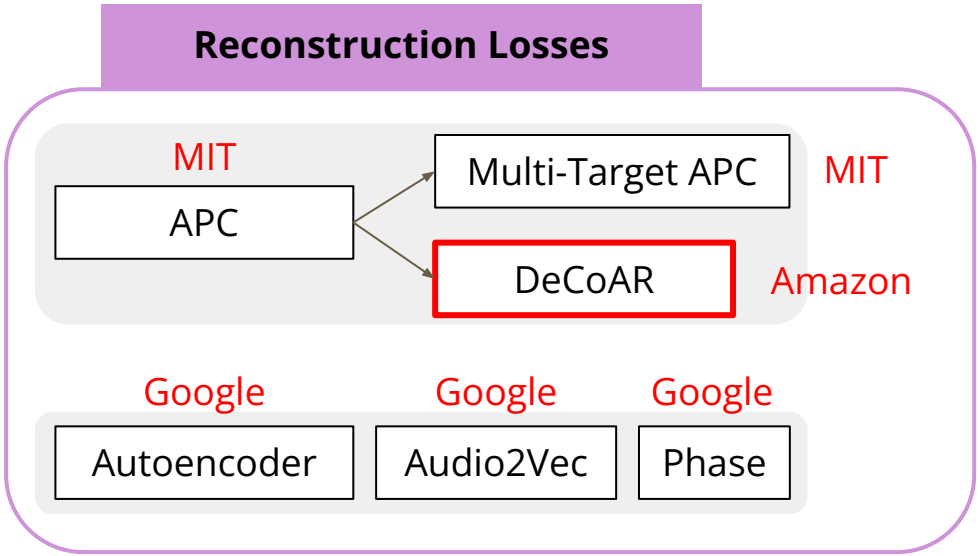
but also predicts the past memory through reconstruction.

Reconstruction Losses

MIT
APC
Multi-Target APC
MIT
DeCoAR
Amazon

Google
Autoencoder
Google
Audio2Vec
Google
Phase

A Broad Introduction

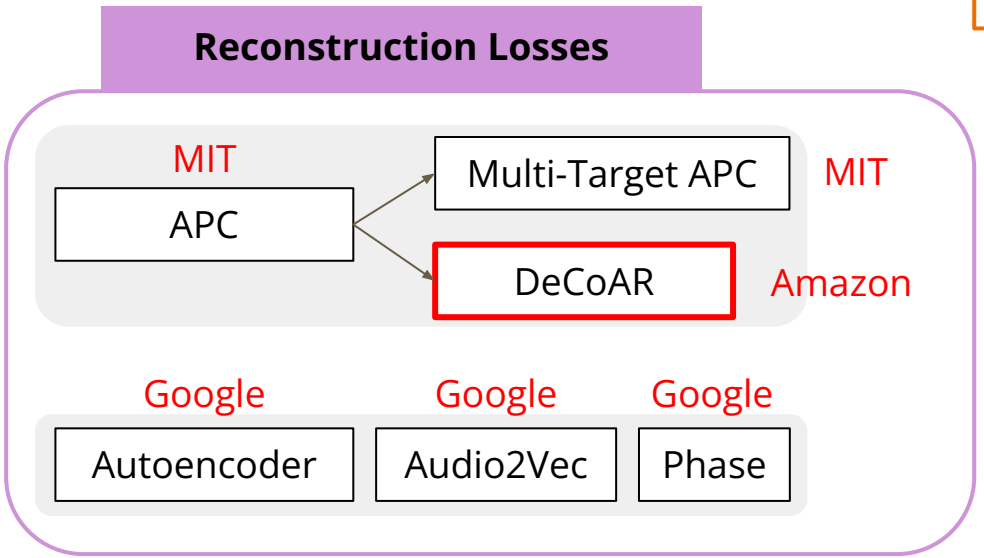


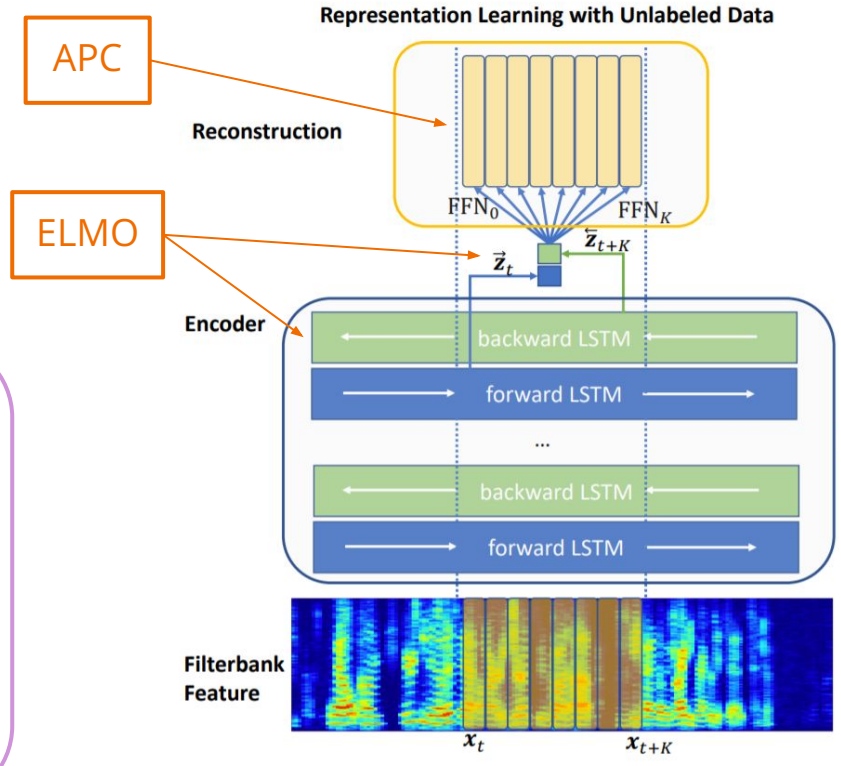
Important exps: ASR on WSJ / LibriSpeech

A Broad Introduction

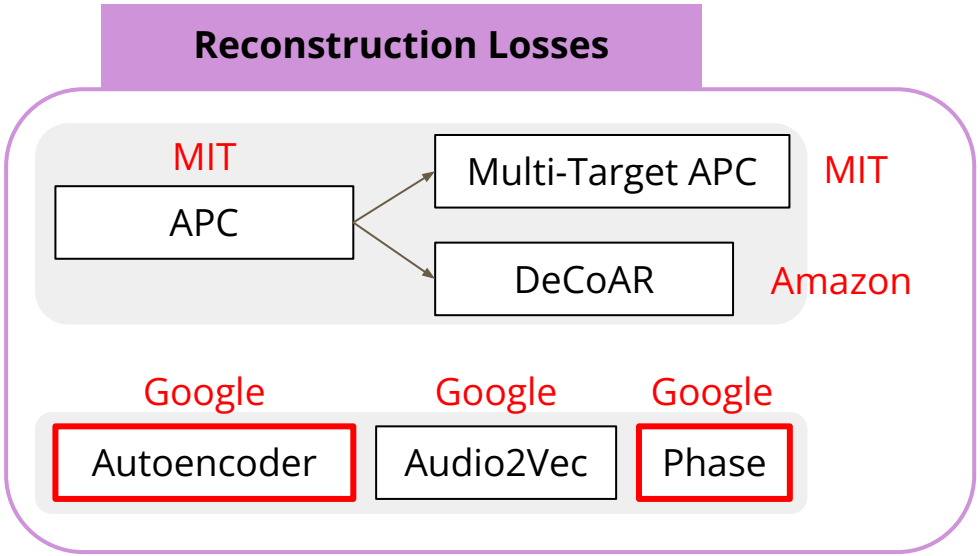
Intuition:
Deep Contextualized Acoustic Representations

Combining the bidirectionality of ELMo and the reconstruction objective of APC. Reconstruction loss is summed over all possible slices in the entire sequence.



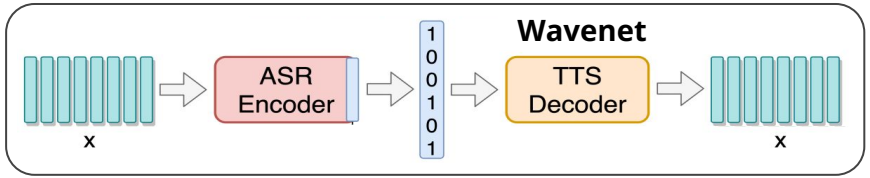


A Broad Introduction

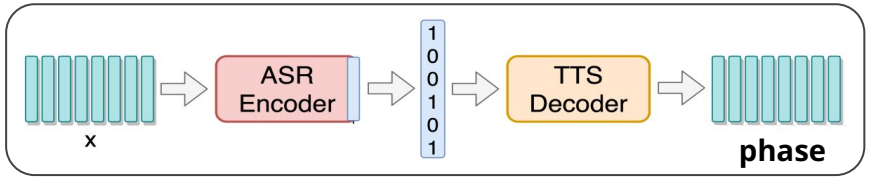


A Broad Introduction

Important exps: Phone Classification



Important exps: Linear Classifications



Reconstruction Losses

MIT

APC

Multi-Target APC

MIT

DeCoAR

Amazon

Google

Autoencoder

Google

Audio2Vec

Google

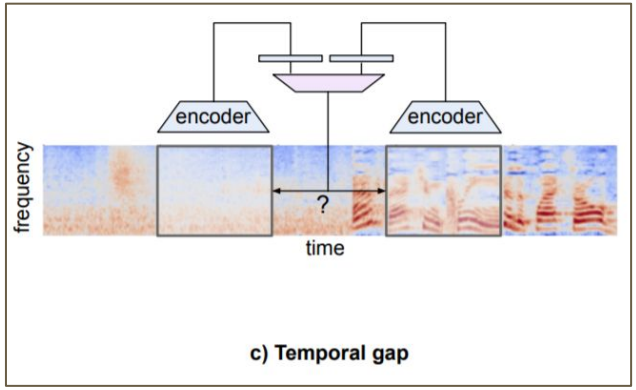
Phase

Intuition:

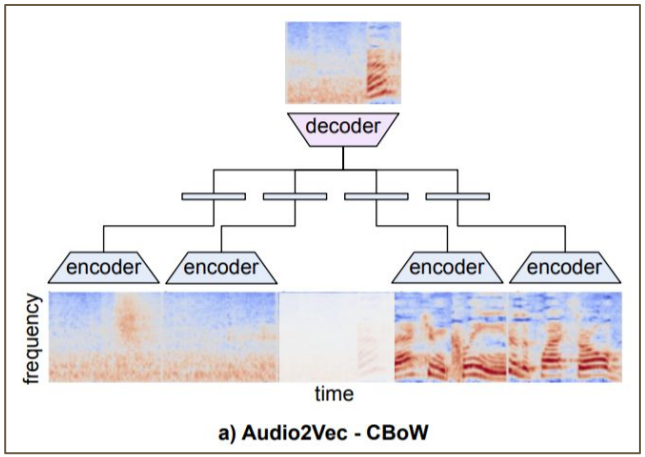
Very similar to the VC structure, learn through autoencoder bottleneck and reconstruction

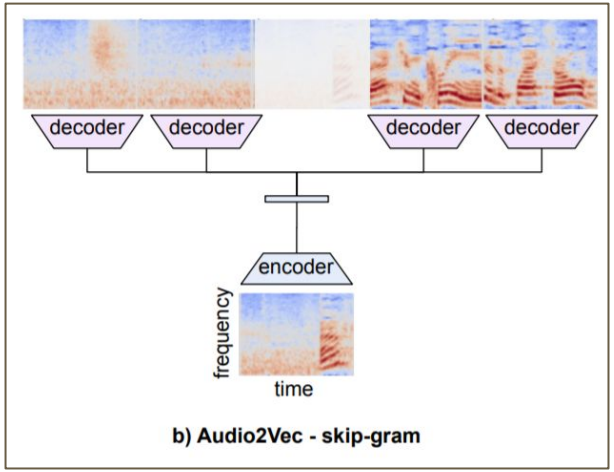
A Broad Introduction

Important exps: Linear Classifications



Intuition:
Audio version of Word2Vec



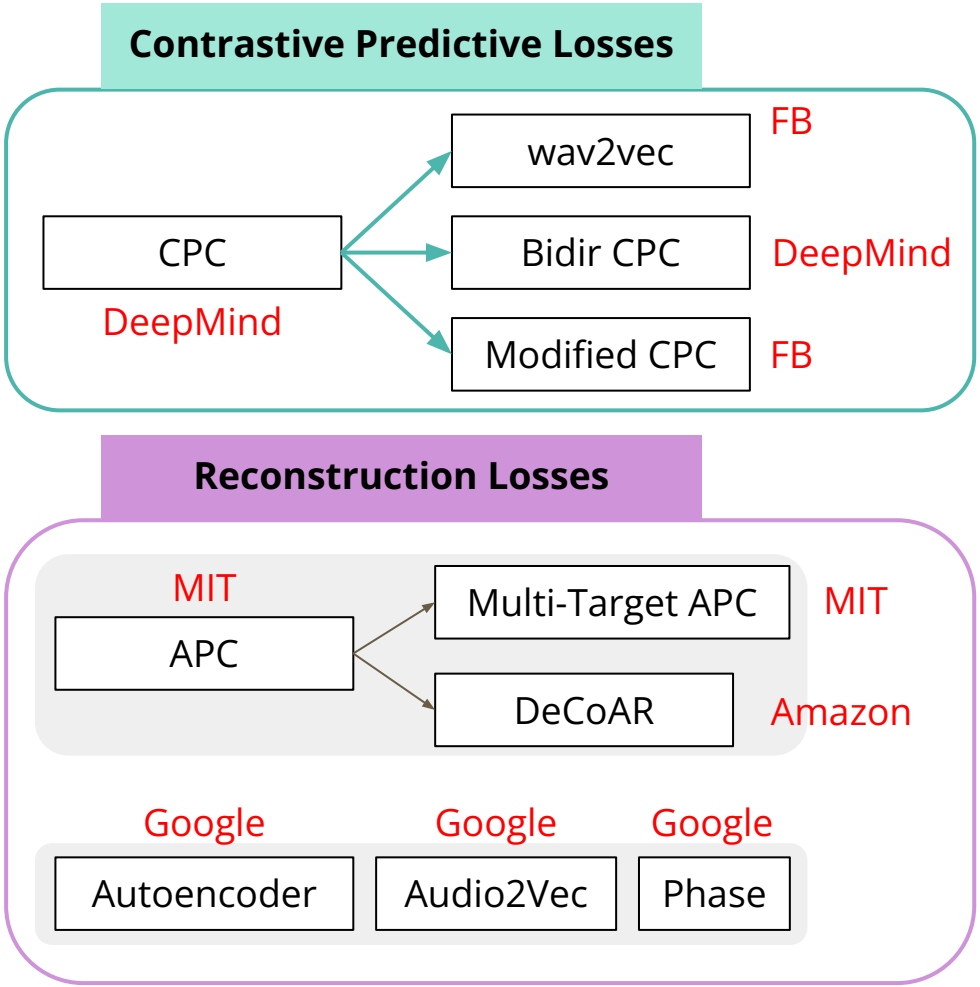


Reconstruction Losses

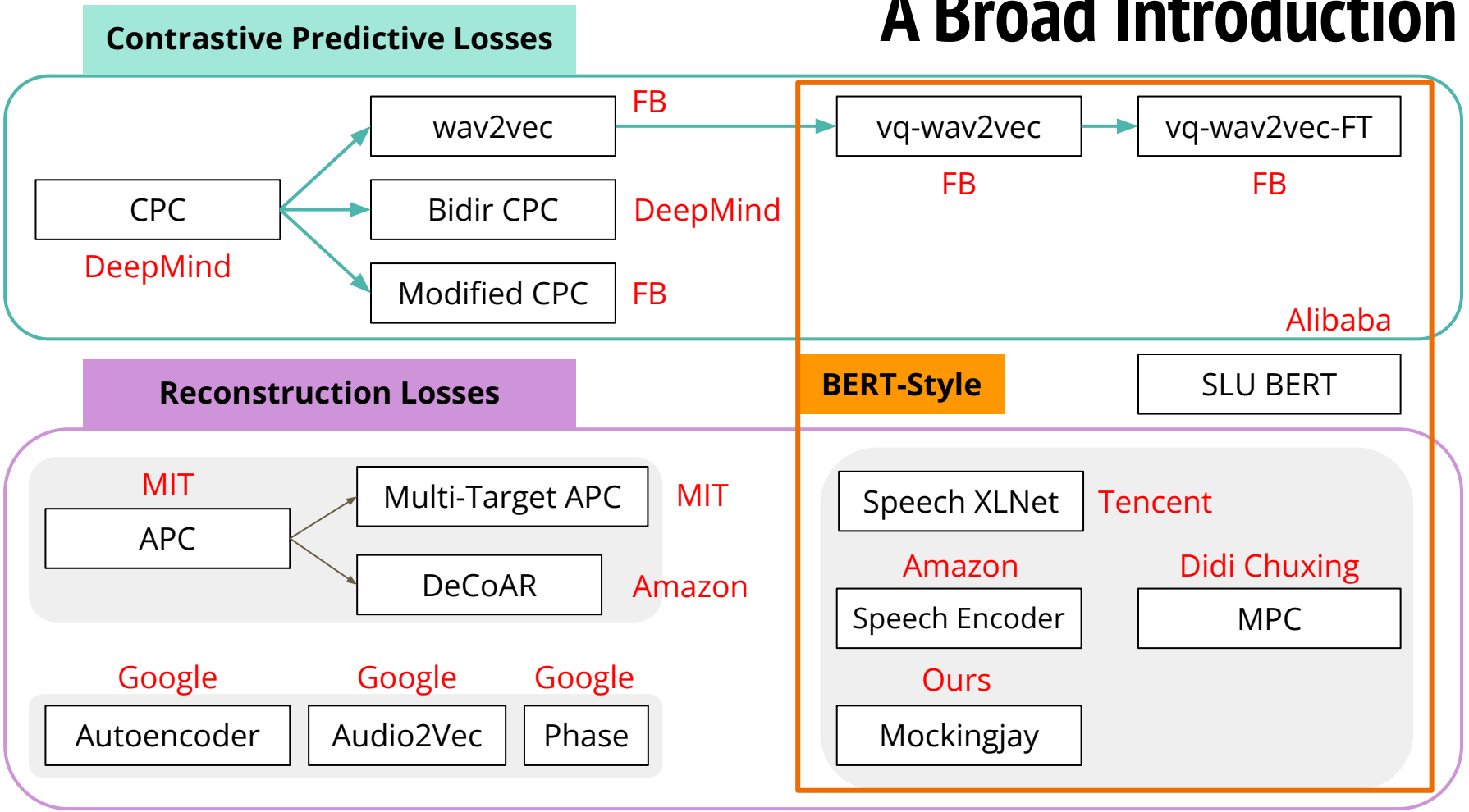
MIT
APC
Multi-Target APC
MIT
DeCoAR
Amazon

Google
Autoencoder
Google
Audio2Vec
Google
Phase

A Broad Introduction



A Broad Introduction

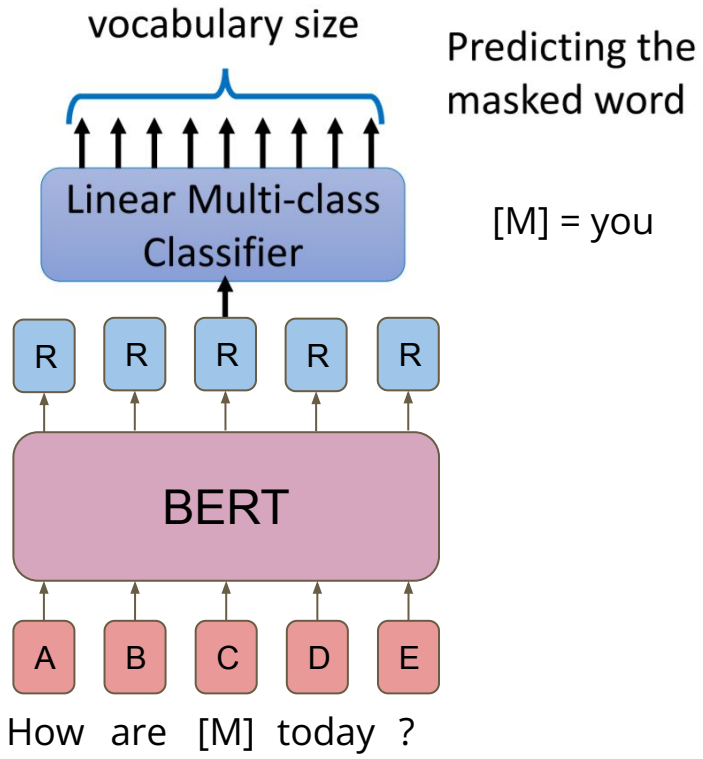


BERT (Bidirectional Encoder Representations from Transformers)

- Achieved **11 SOTA** when published.
- A technique for NLP pre-training developed by Google.

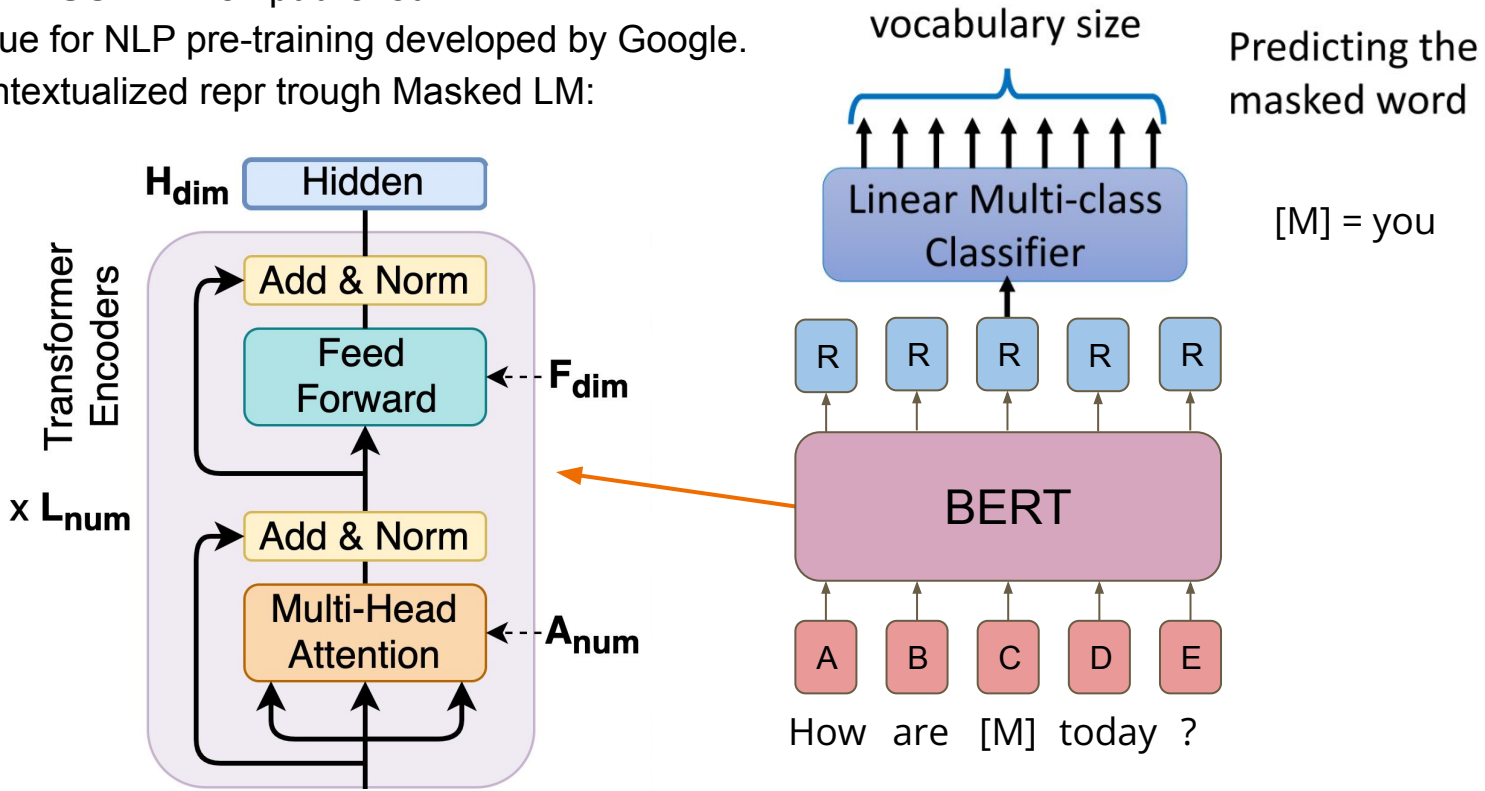
BERT (Bidirectional Encoder Representations from Transformers)

- Achieved **11 SOTA** when published.
- A technique for NLP pre-training developed by Google.
- Learn contextualized repr trough Masked LM:

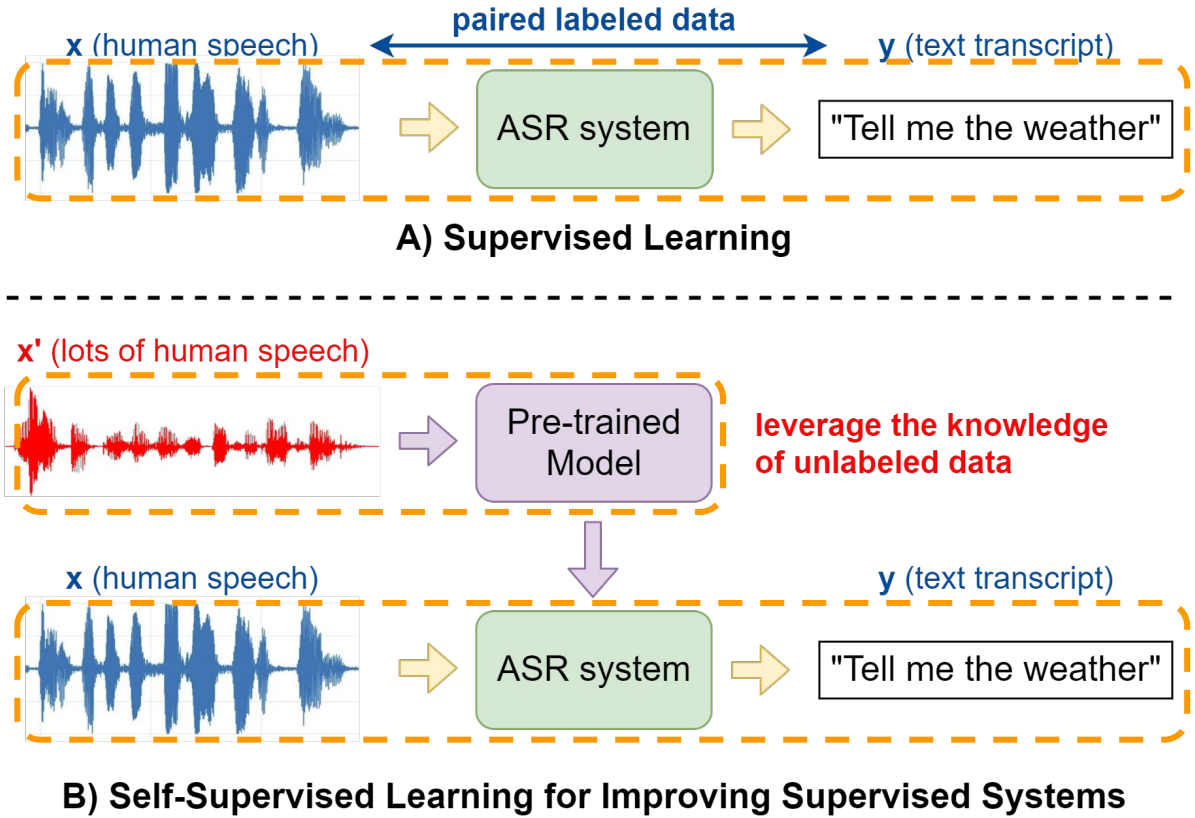


BERT (Bidirectional Encoder Representations from Transformers)

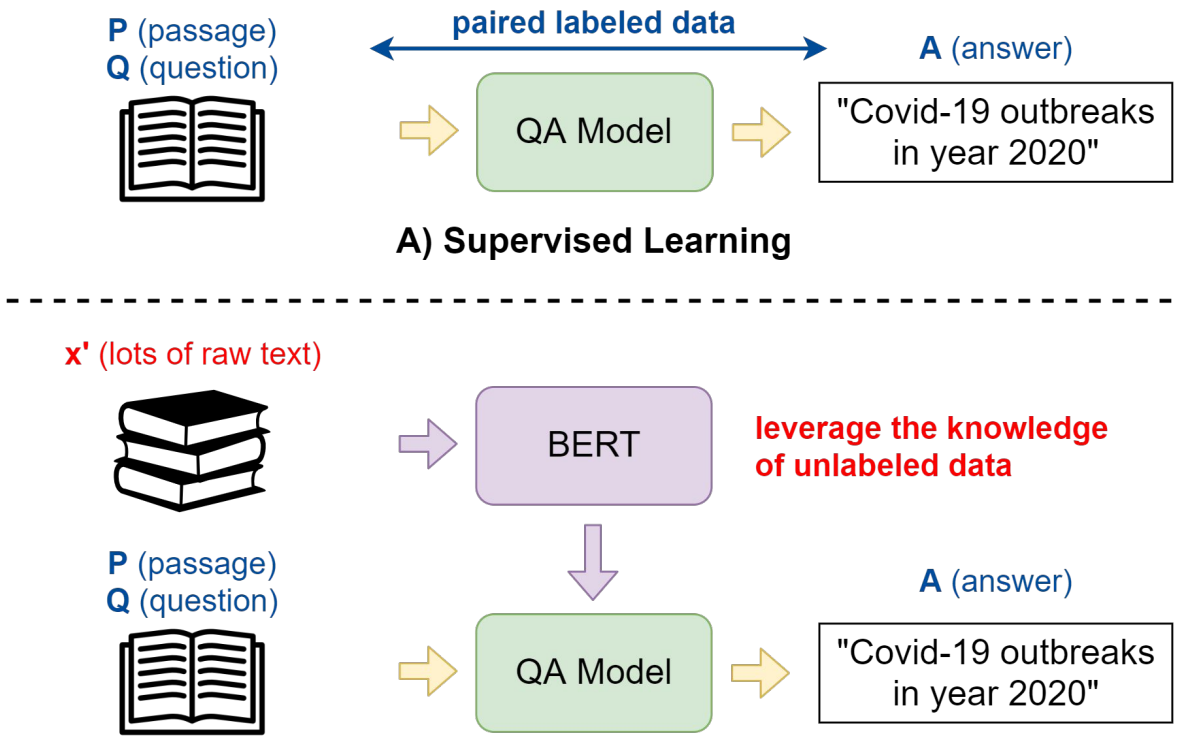
- Achieved **11 SOTA** when published.
- A technique for NLP pre-training developed by Google.
- Learn contextualized repr trough Masked LM:



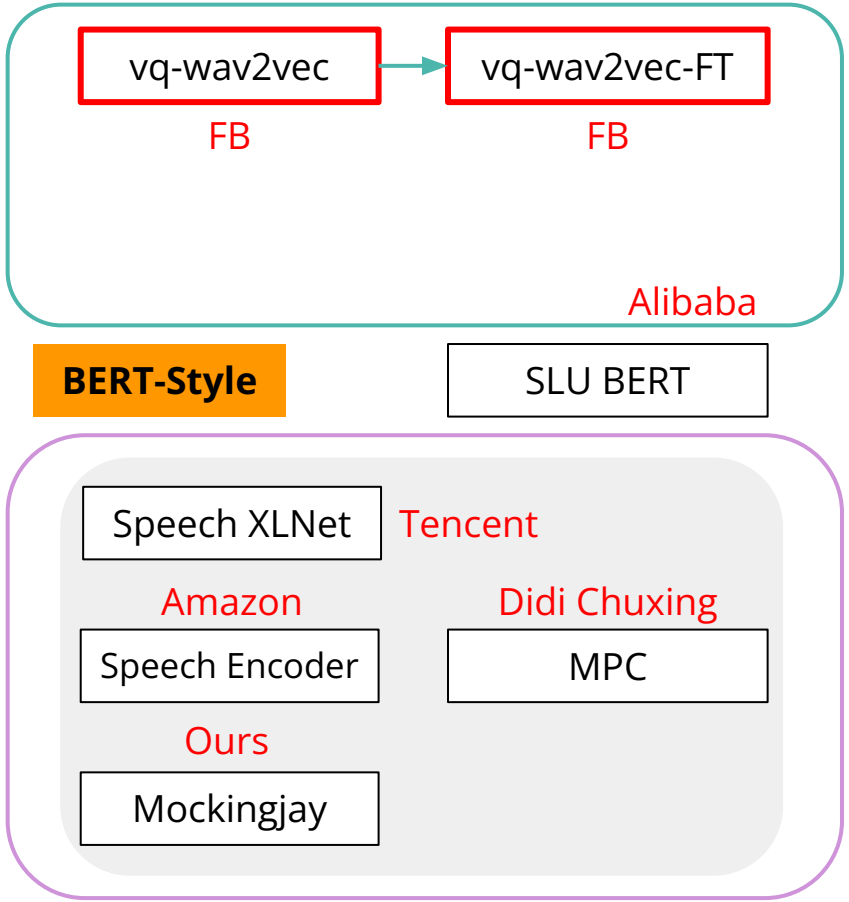
Recall: Self-Supervised Learning for Speech



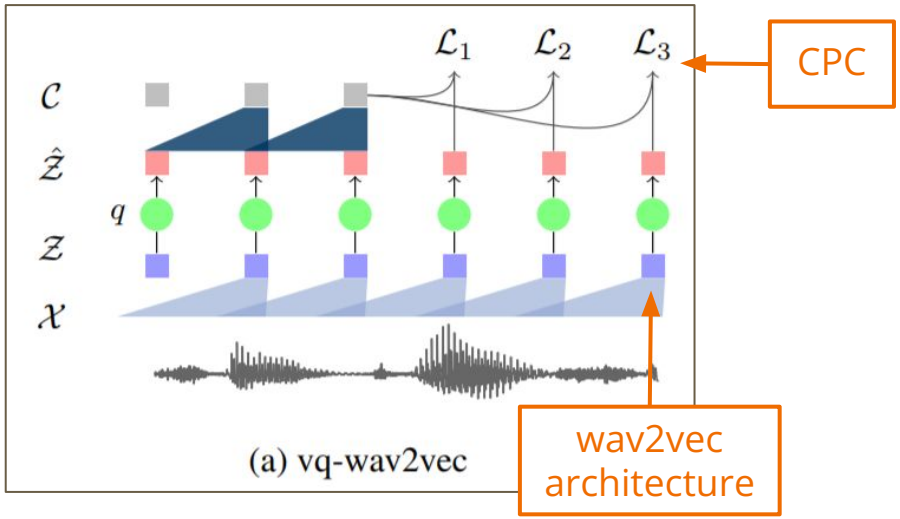
Self-Supervised Learning: BERT



A Broad Introduction



A Broad Introduction



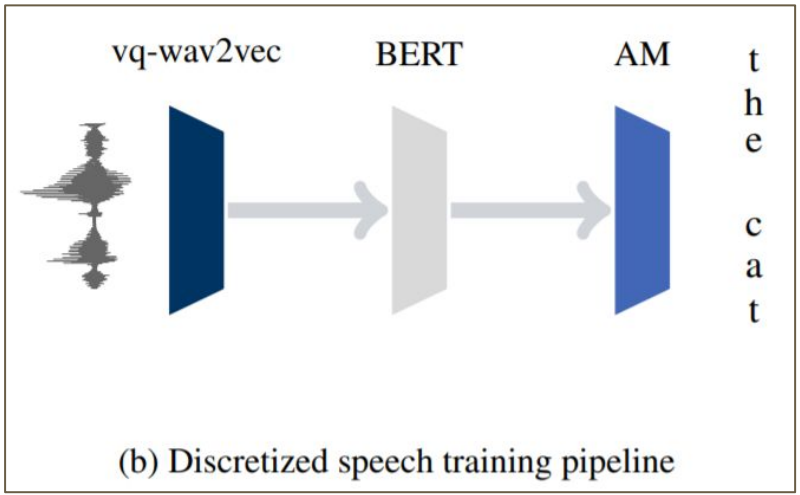
CPC

wav2vec architecture

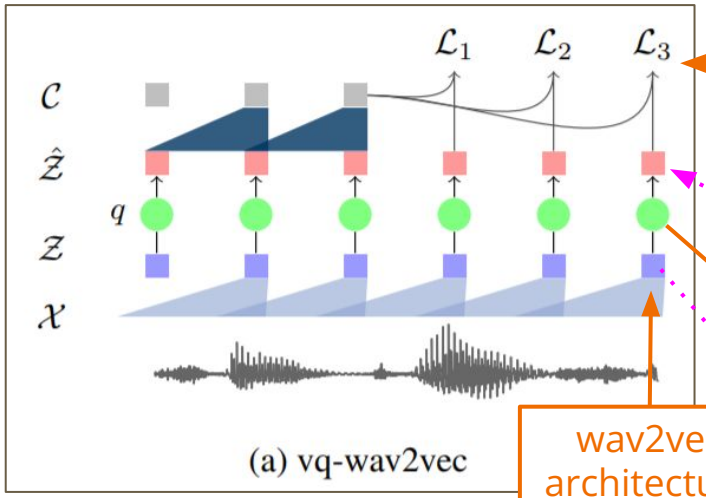
vq-wav2vec → vq-wav2vec-FT

FB FB

BERT-Style



A Broad Introduction



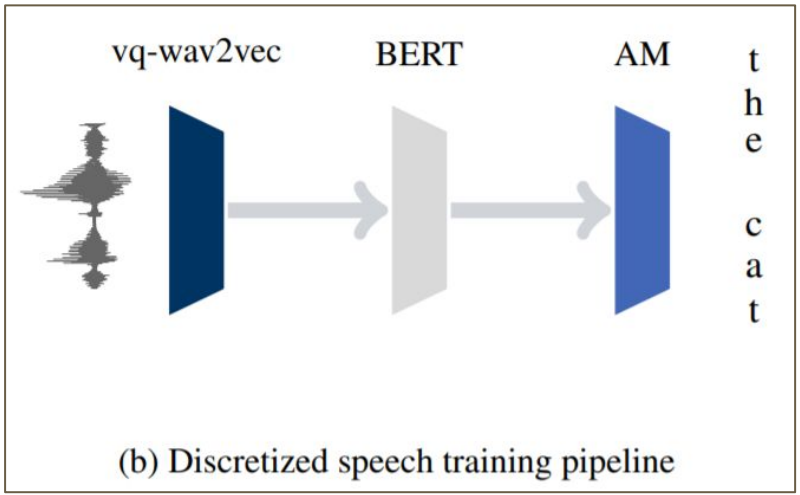
CPC

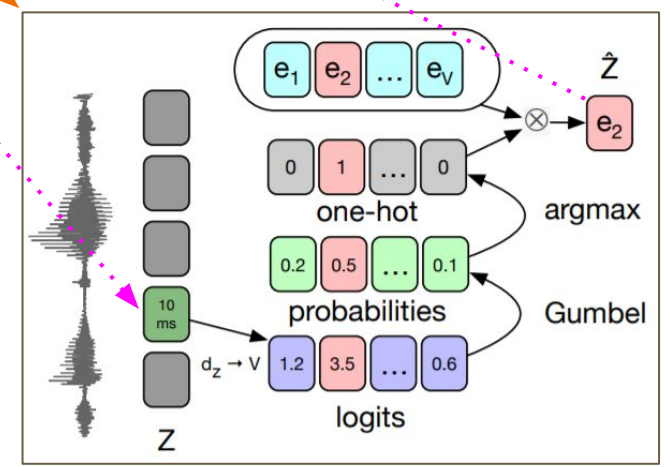
wav2vec architecture

vq-wav2vec → vq-wav2vec-FT

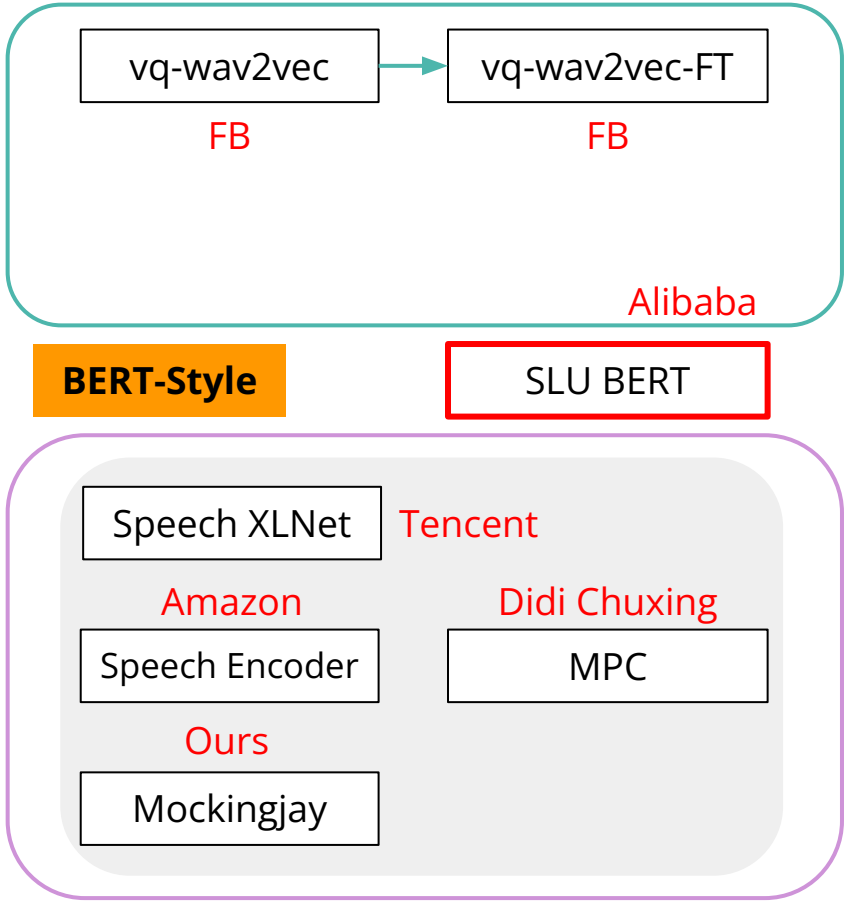
FB FB

BERT-Style





A Broad Introduction



A Broad Introduction

masked frames are filled with a default posterior vectors with the probability of a placeholder phoneme “[PAD]” equals to 1.

CTC loss training over ground-truth phonemes

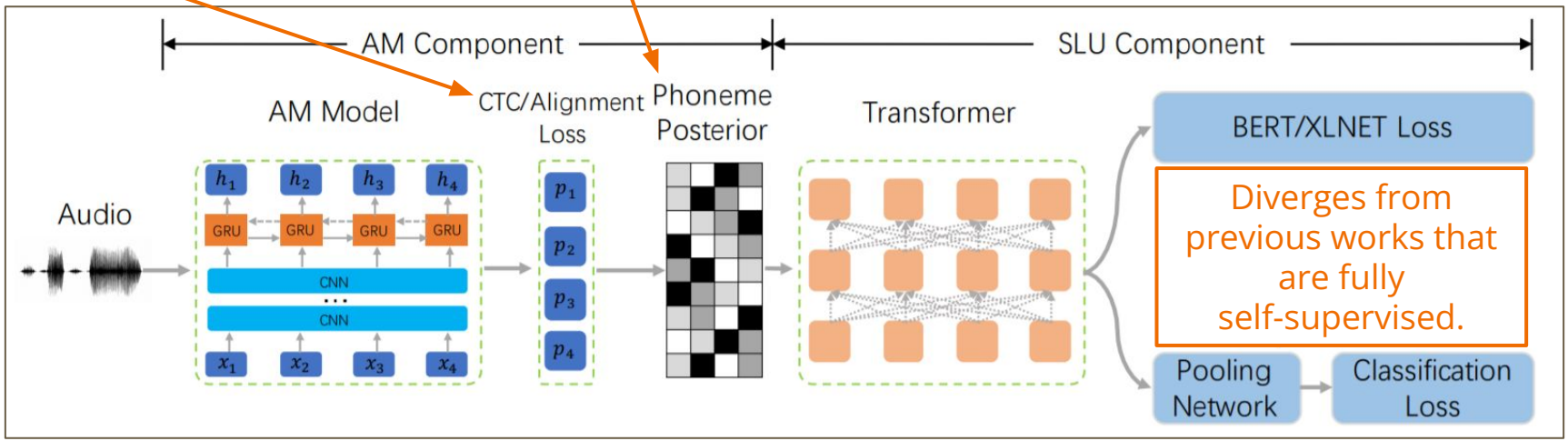
vq-wav2vec → vq-wav2vec-FT

FB FB

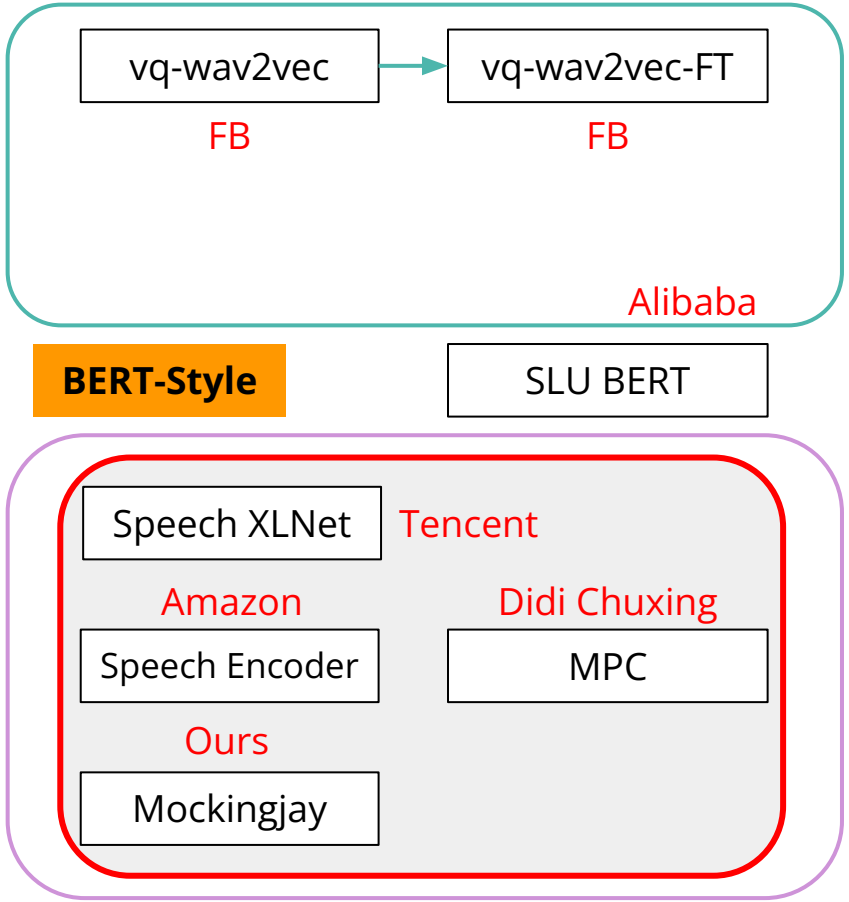
Alibaba

BERT-Style

SLU BERT



A Broad Introduction



The Trend:
All of these works emerges around October, 2019.
All submitted to ICASSP 2020

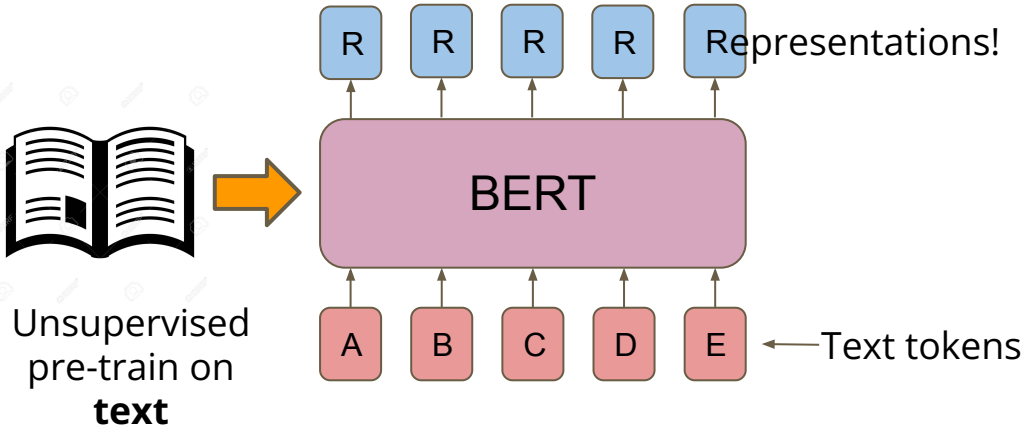
(Speech XLNet and MPC did not make it)

Mockingjay

From BERT to Speech BERT

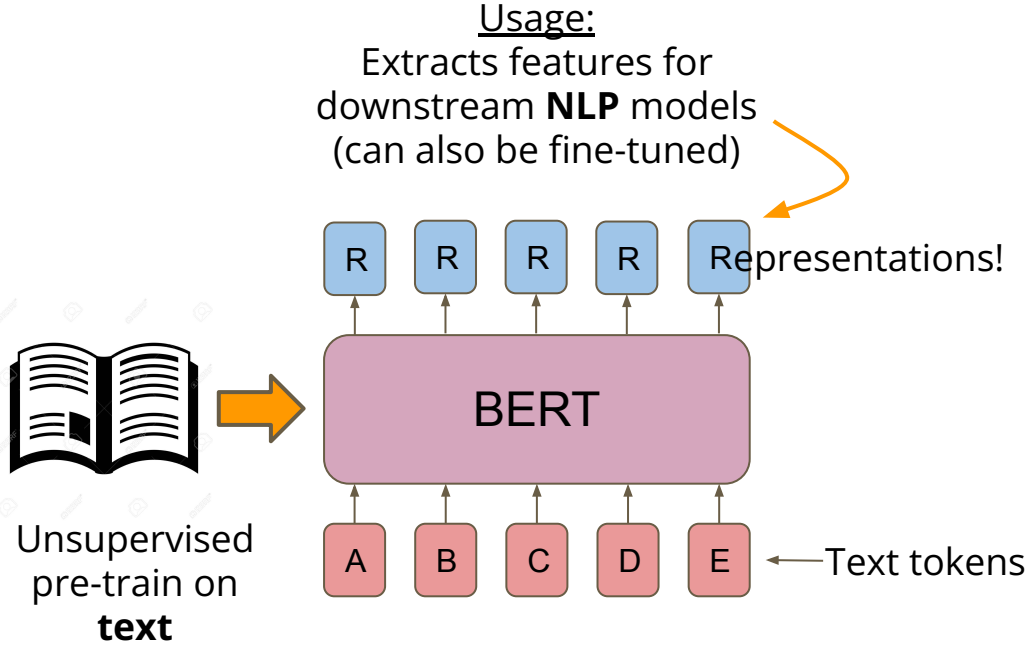
From BERT to Speech BERT

NLP BERT:
Language Representation Learning



From BERT to Speech BERT

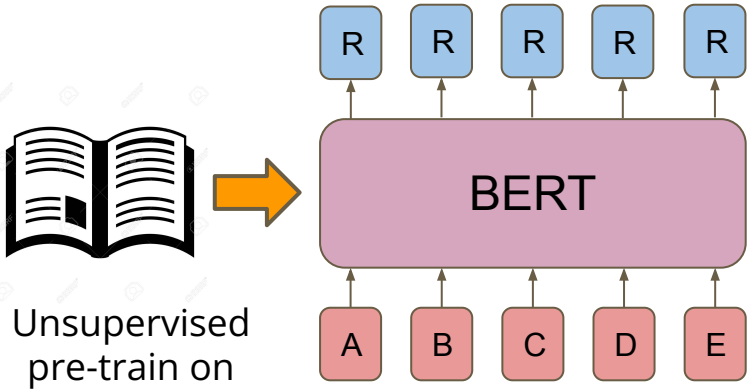
NLP BERT:
Language Representation Learning



From BERT to Speech BERT

NLP BERT:
Language Representation Learning

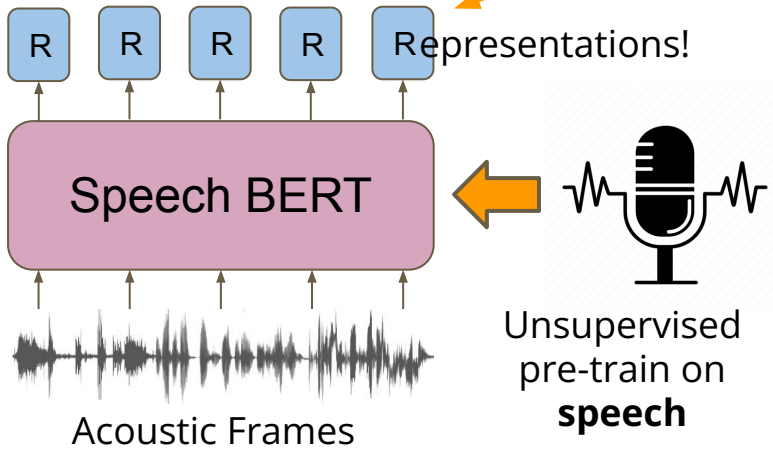
Usage:
Extracts features for downstream **NLP** models (can also be fine-tuned)



Unsupervised pre-train on **text**

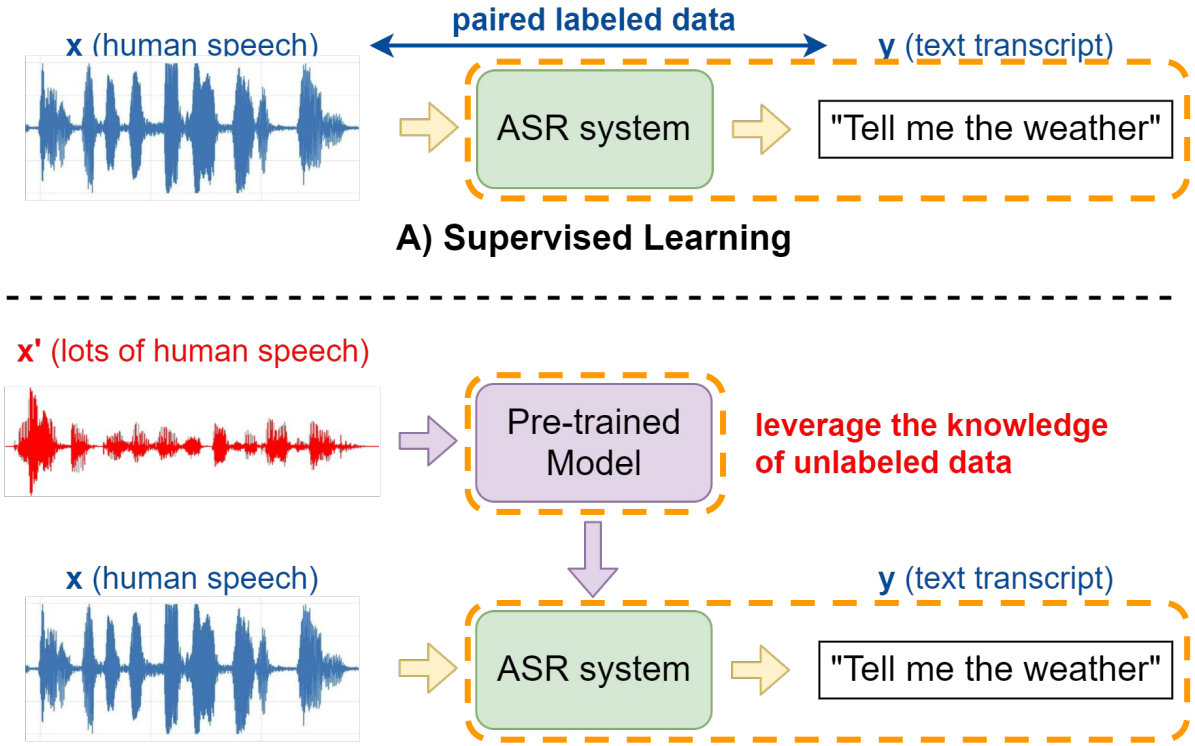
Speech BERT:
Speech Representation Learning

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

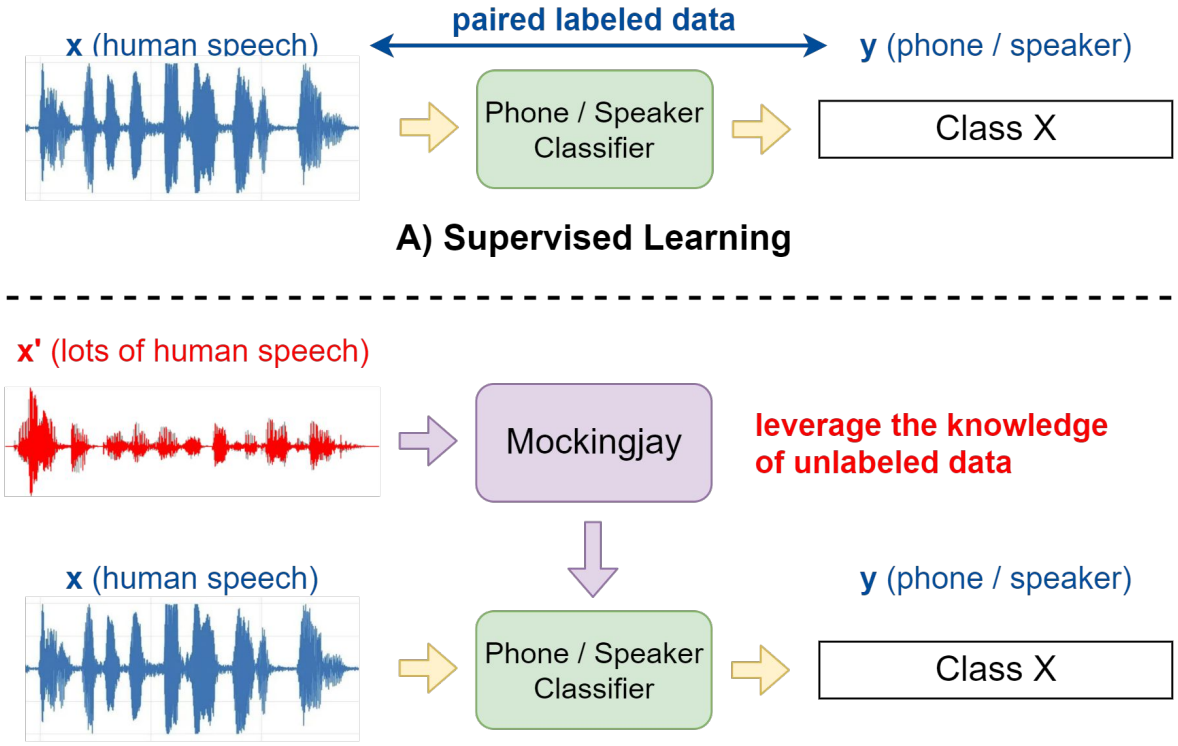


Unsupervised pre-train on **speech**

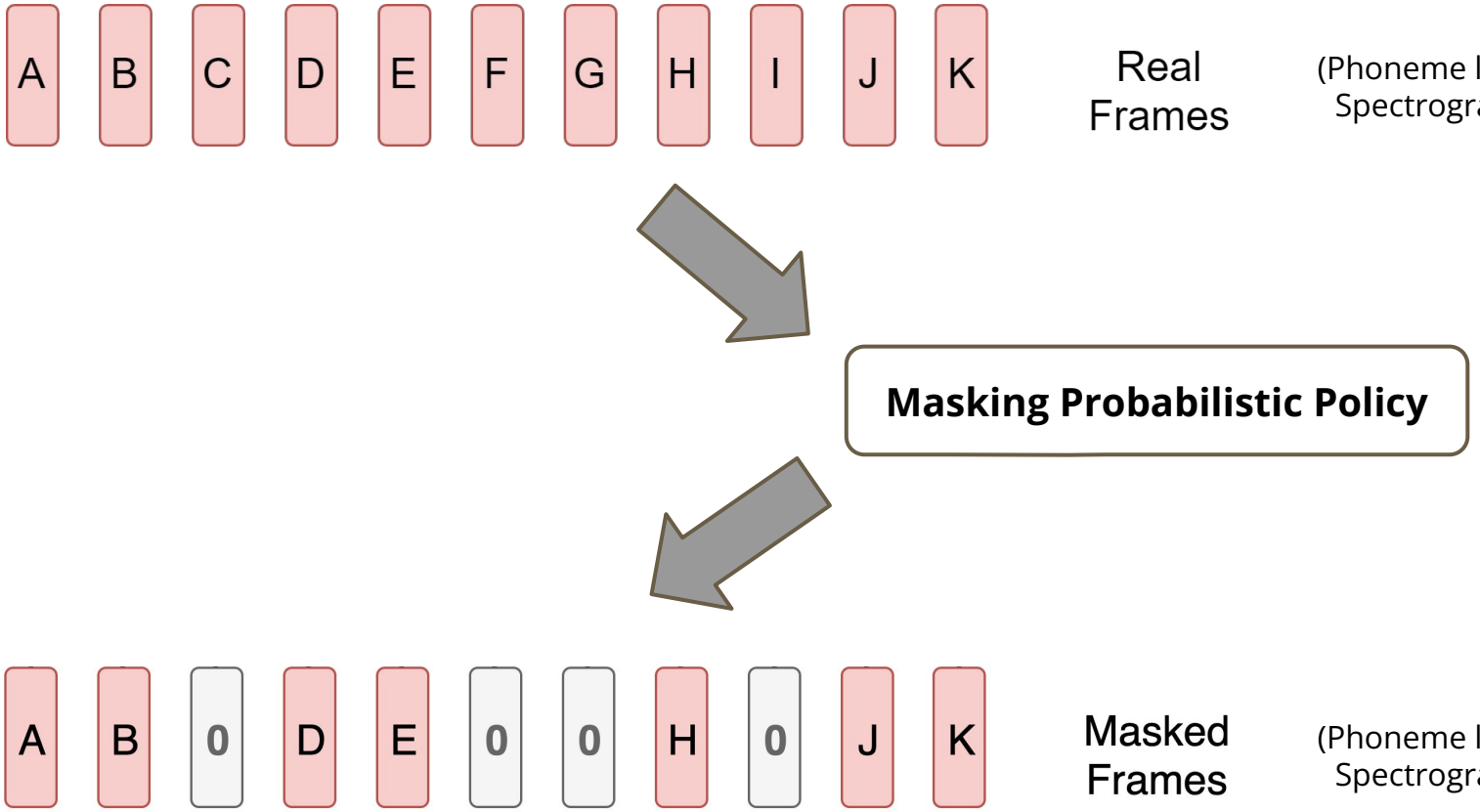
Recall: Self-Supervised Learning for Speech



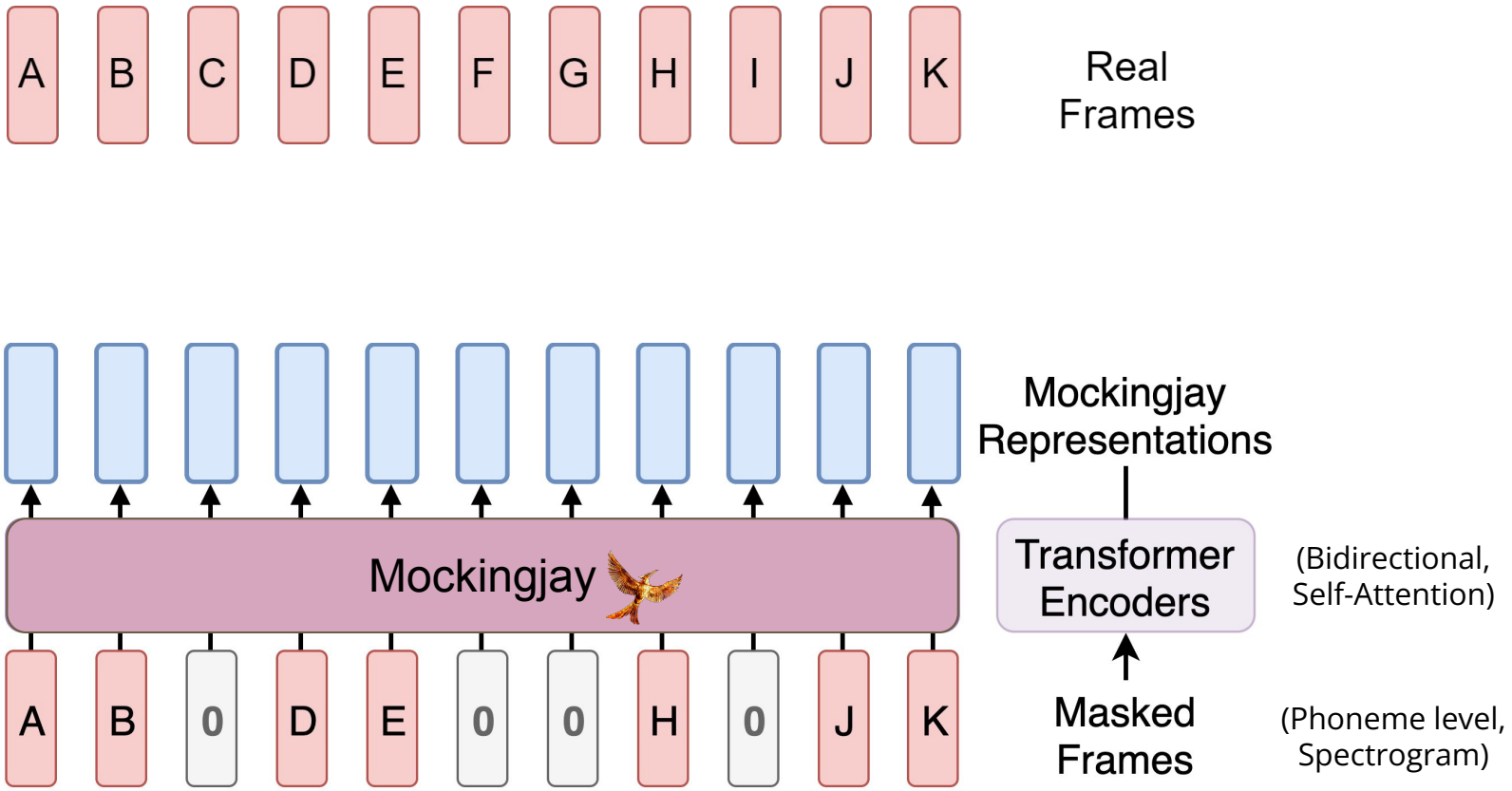
Self-Supervised Learning: Mockingjay



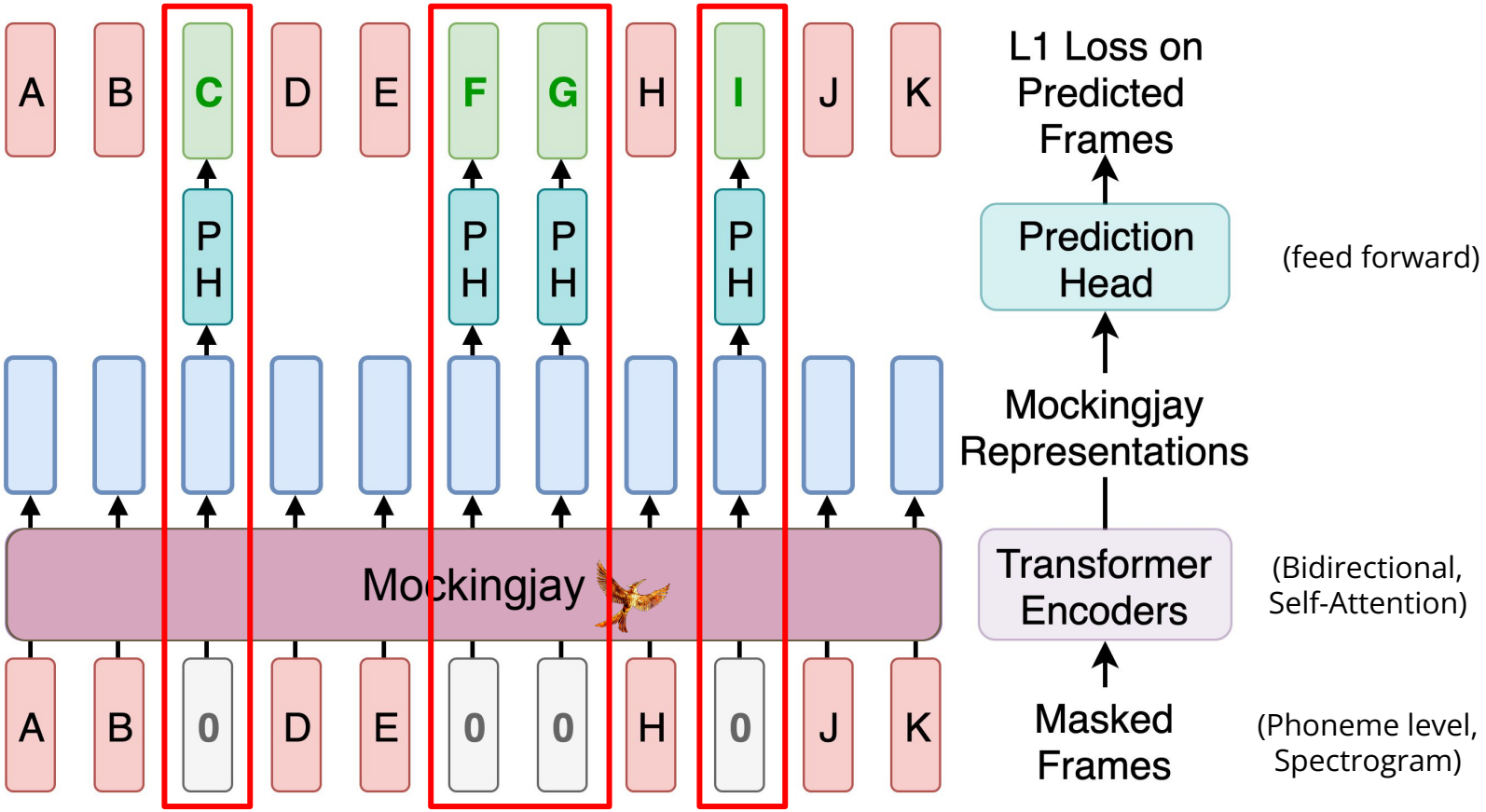
Pre-Training Task: Masked Acoustic Model



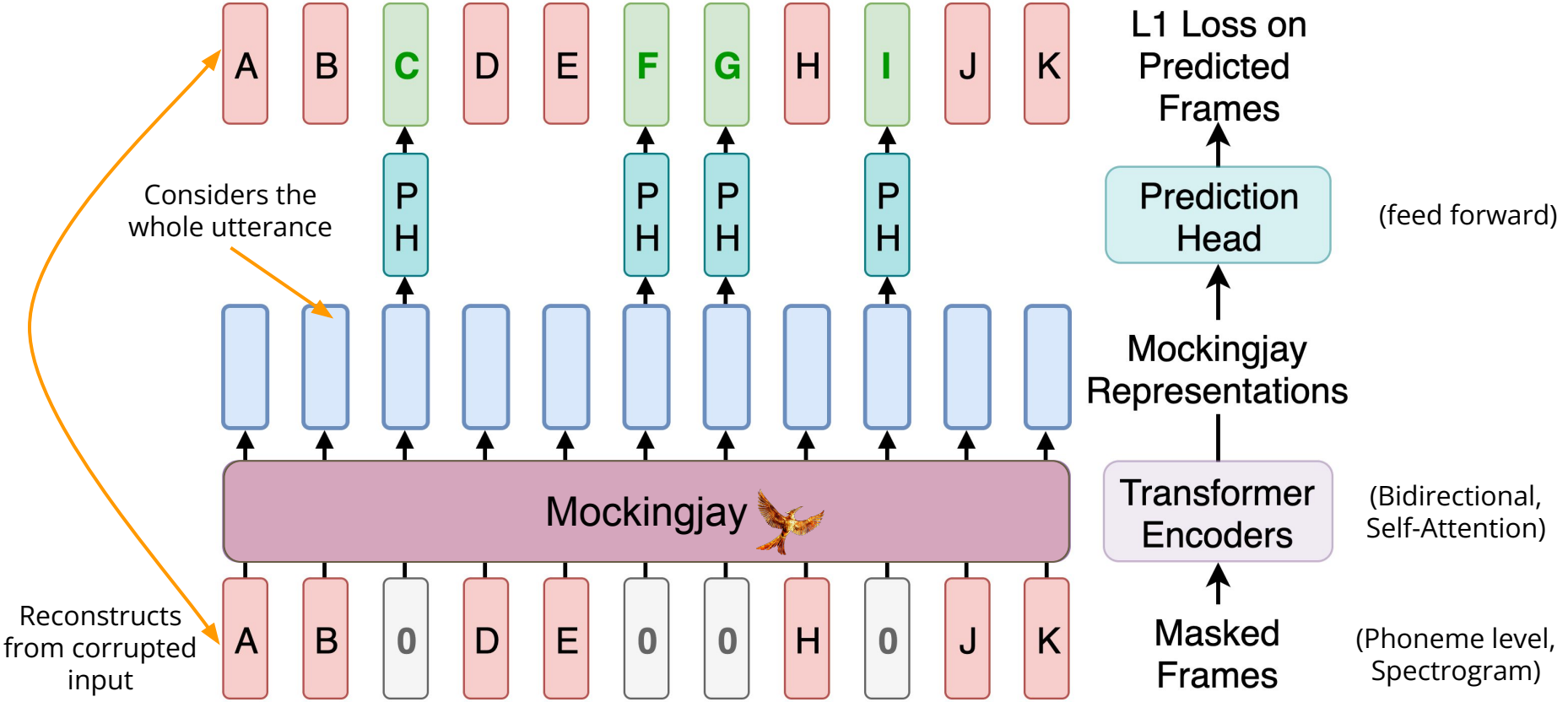
Pre-Training Task: Masked Acoustic Model



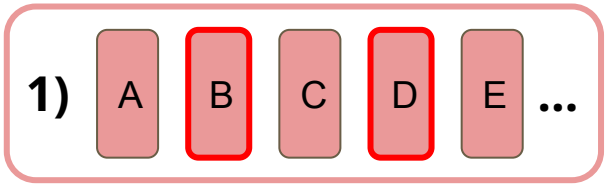
Pre-Training Task: Masked Acoustic Model



Pre-Training Task: Masked Acoustic Model

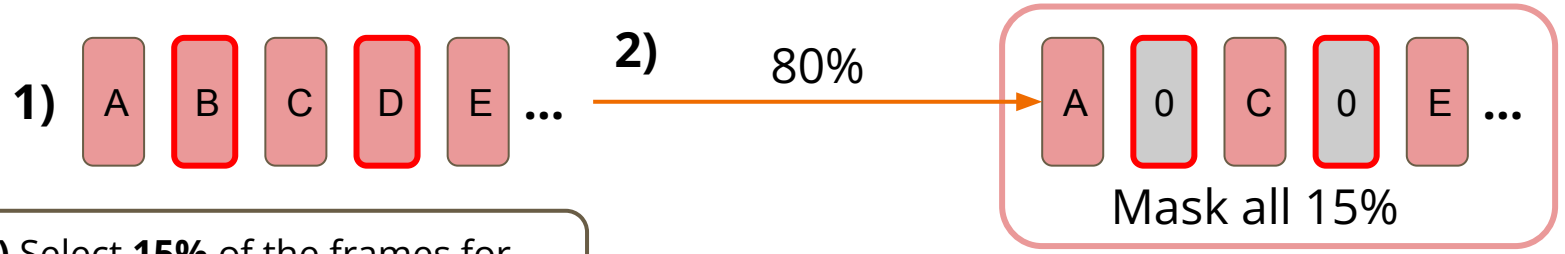


Probabilistic Policy for Masking Frames



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

Probabilistic Policy for Masking Frames

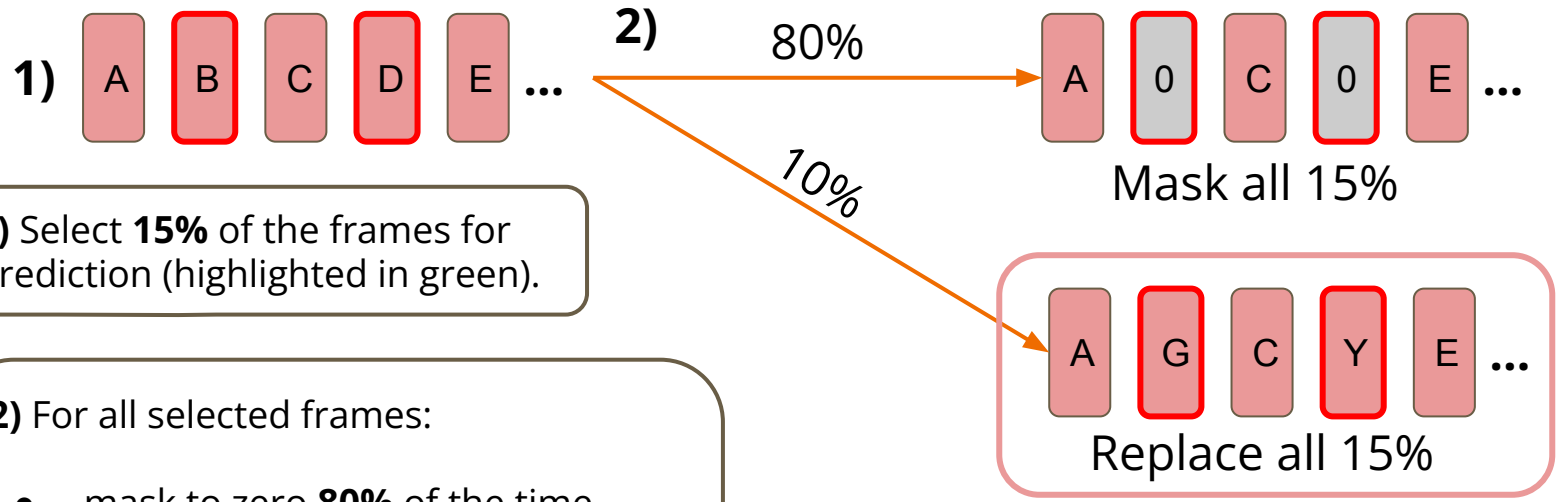


1) Select **15%** of the frames for prediction (highlighted in green).

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

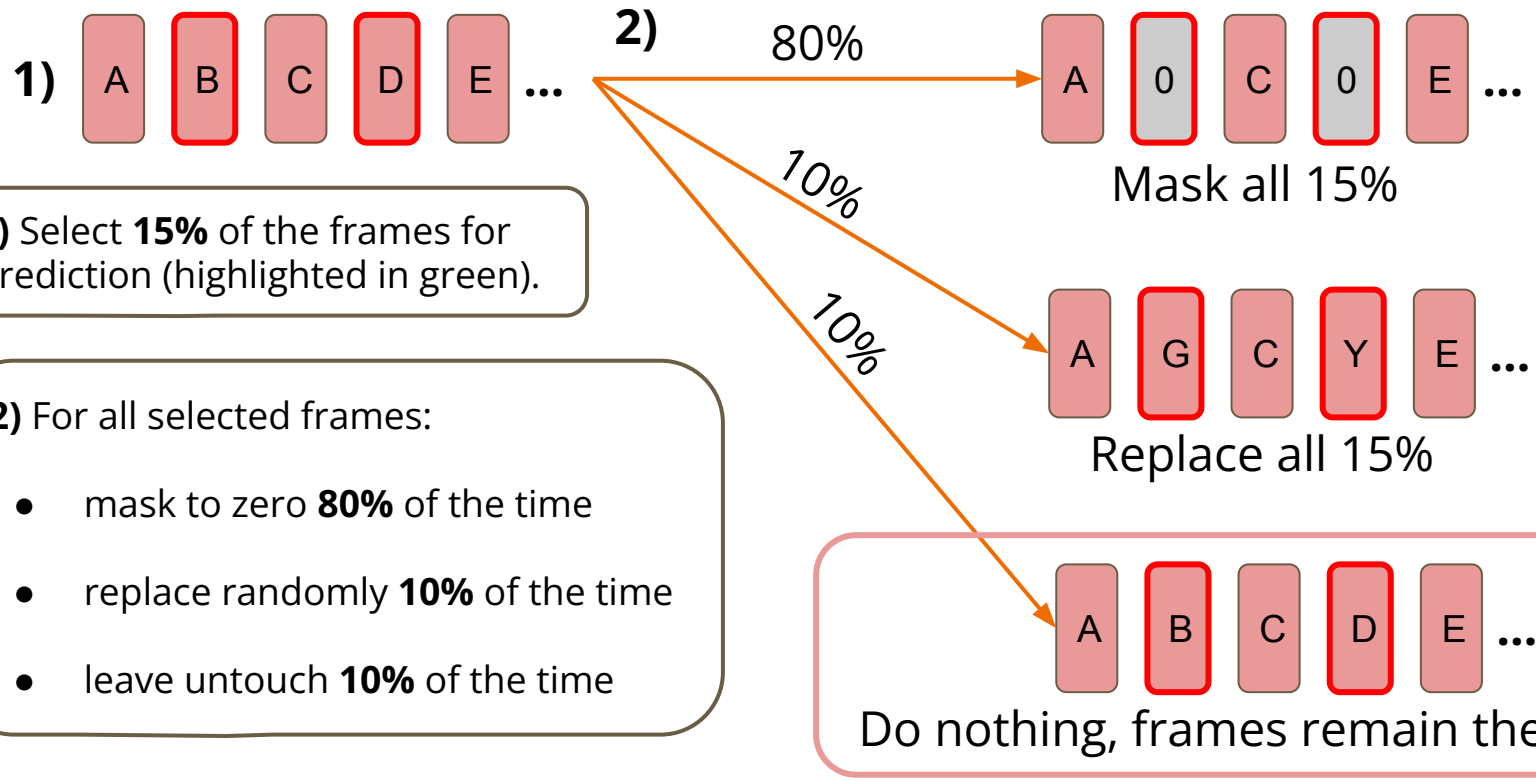


1) Select **15%** of the frames for prediction (highlighted in green).

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

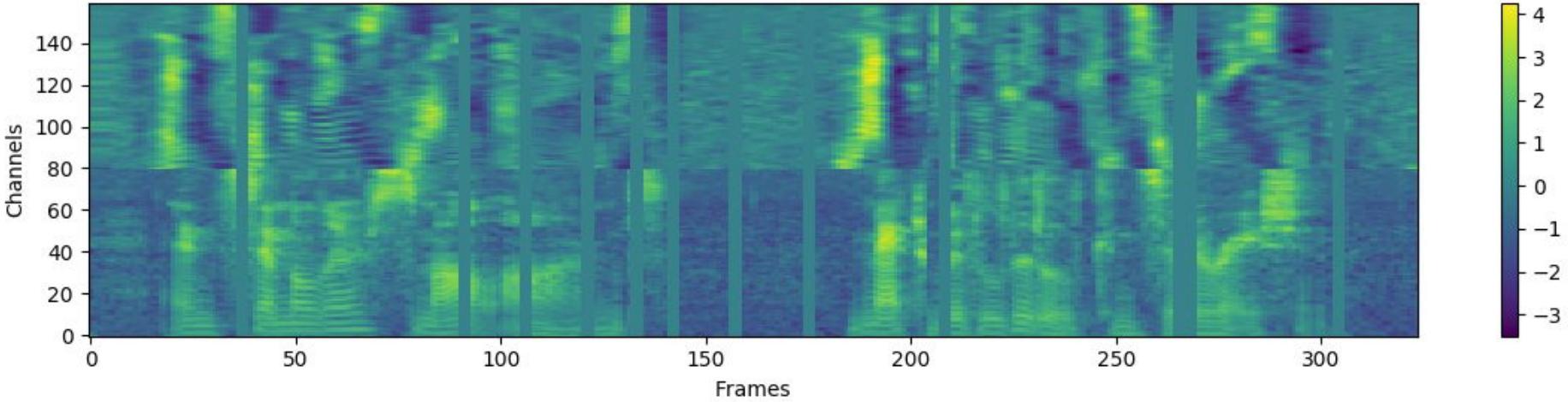


1) Select **15%** of the frames for prediction (highlighted in green).

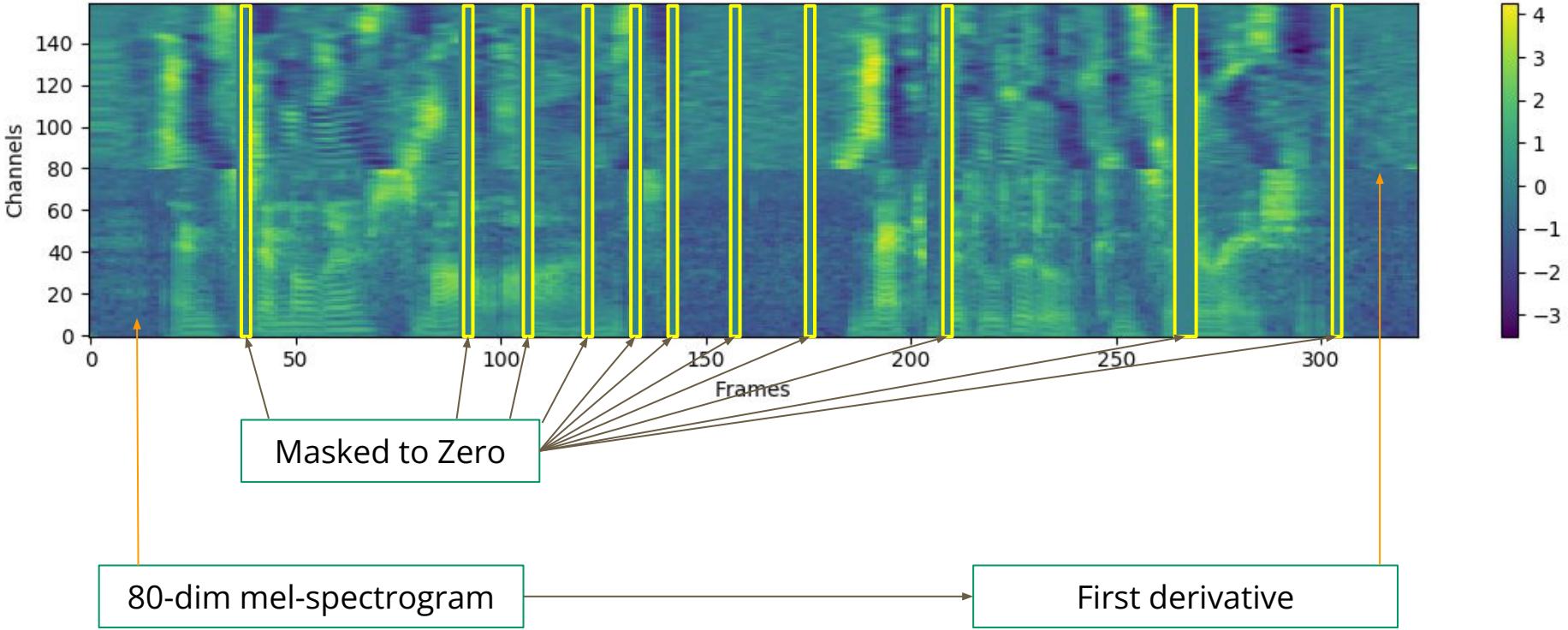
2) For all selected frames:

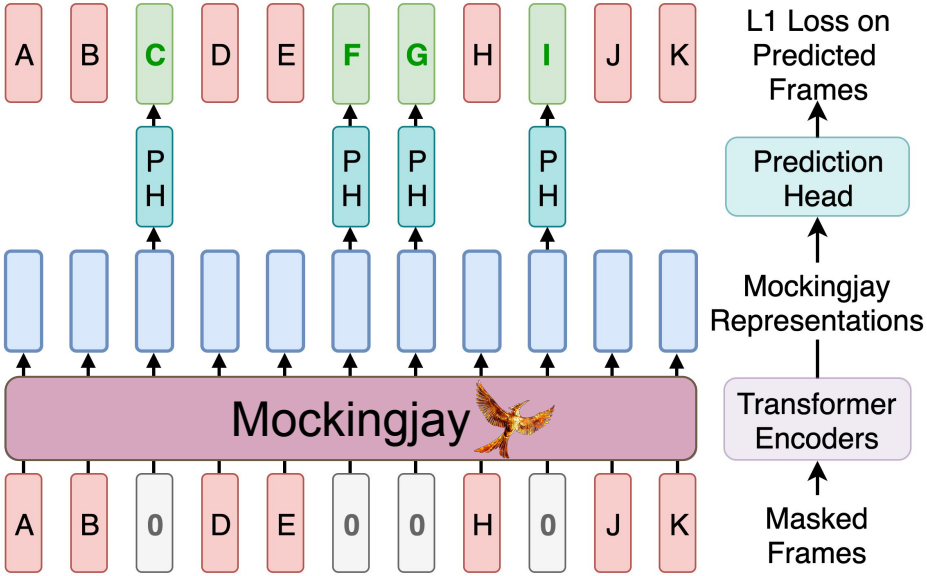
- mask to zero **80%** of the time
- replace randomly **10%** of the time
- leave untouch **10%** of the time

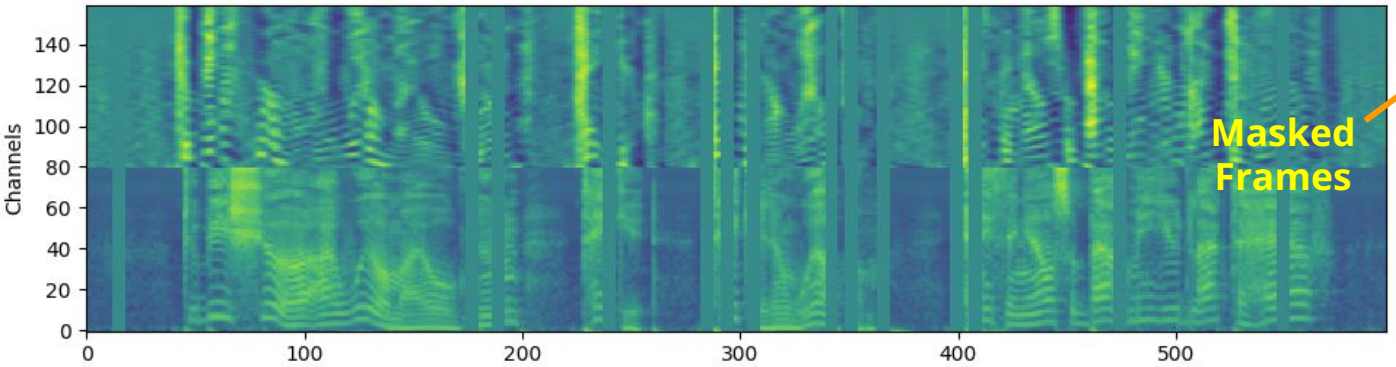
Input Feature: Masked Spectrogram



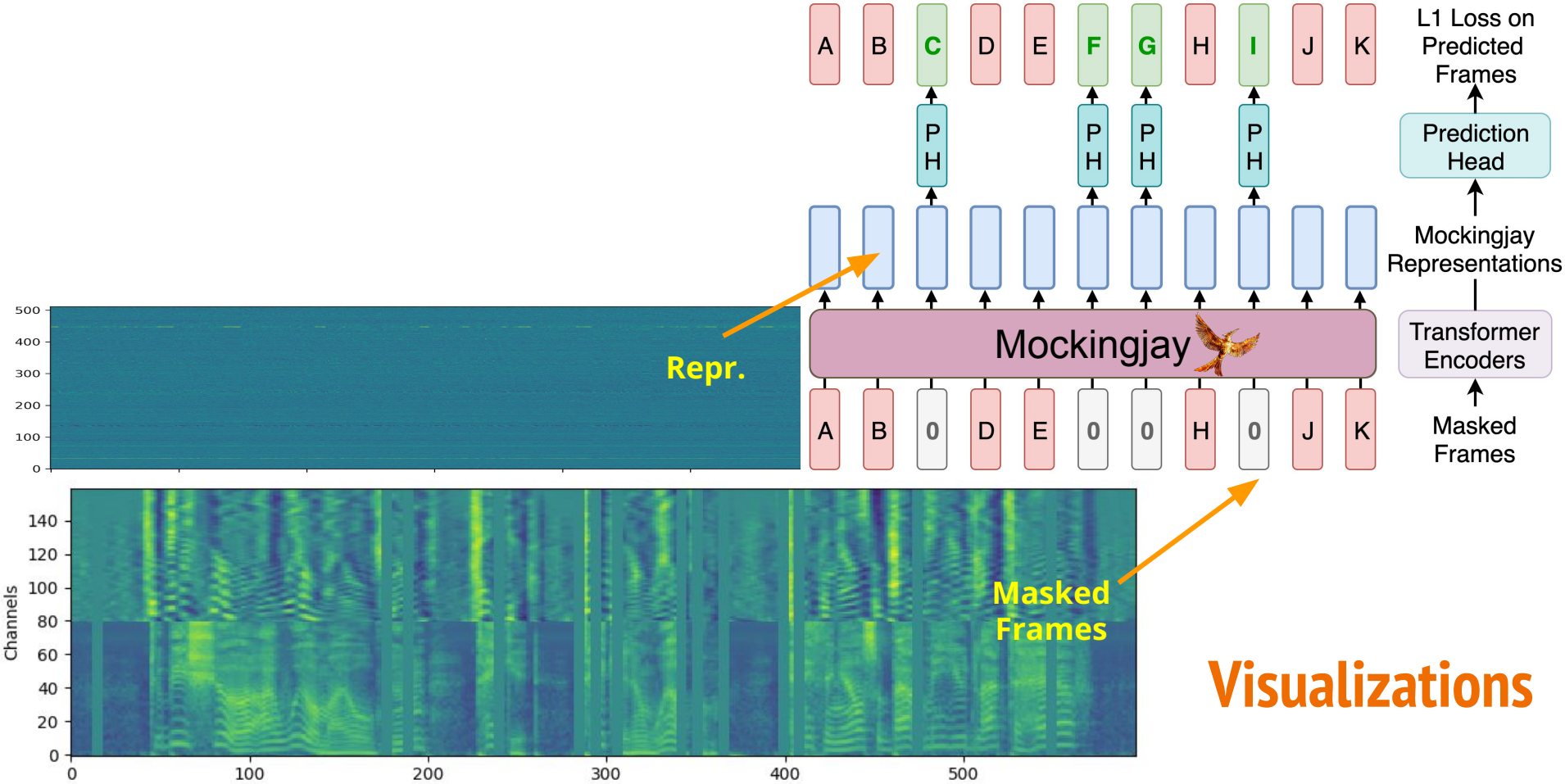
Input Feature: Masked Spectrogram



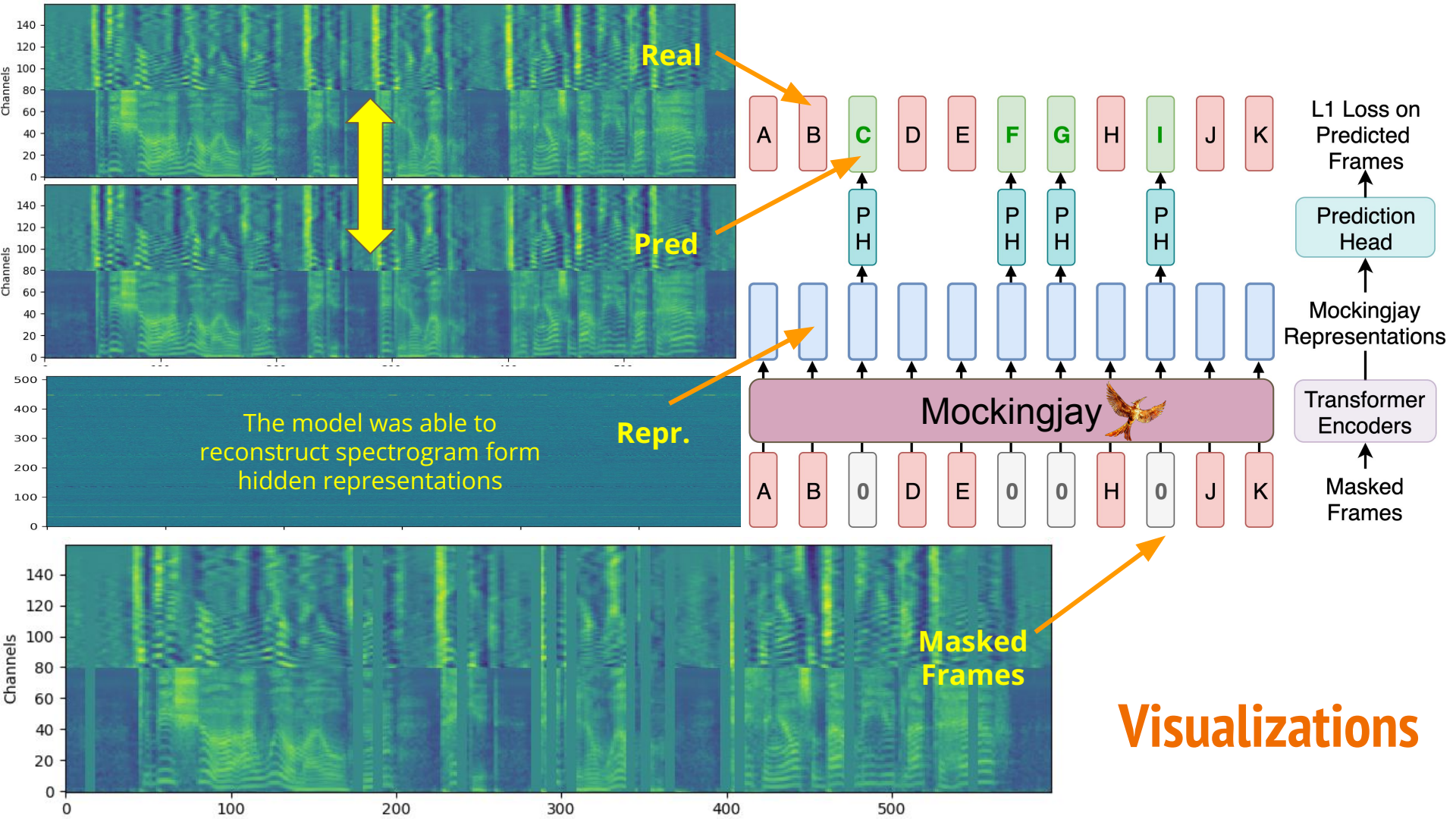




Visualizations



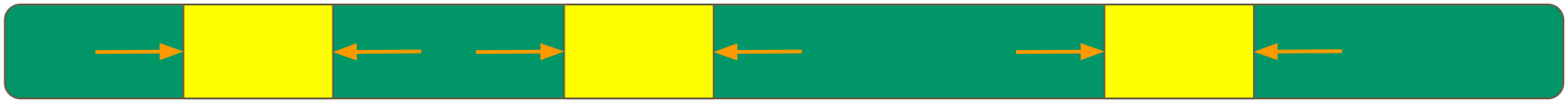
Visualizations


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

Migrating from text to speech

Acoustic Features: long and locally smooth in nature,

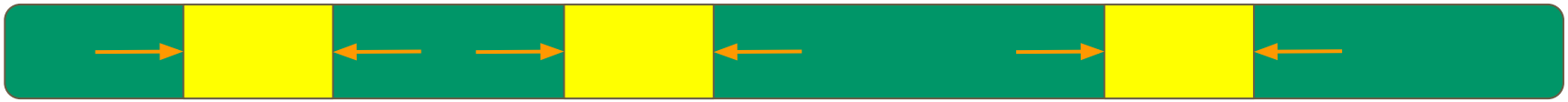
need to 1) shorten the sequence and 2) mask over a longer span



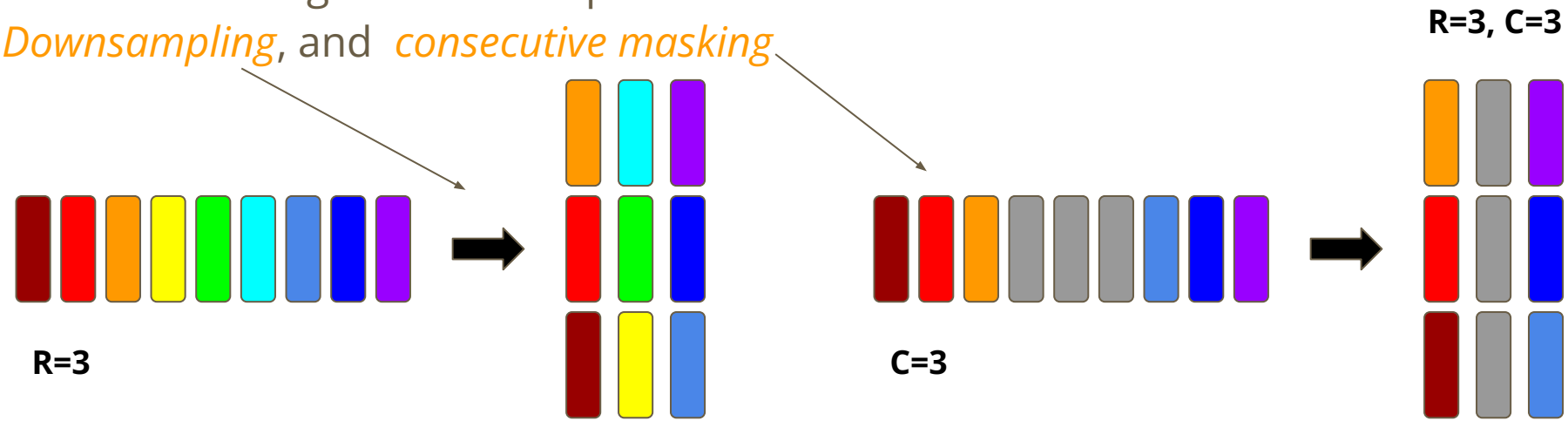
Migrating from text to speech

Acoustic Features: long and locally smooth in nature,

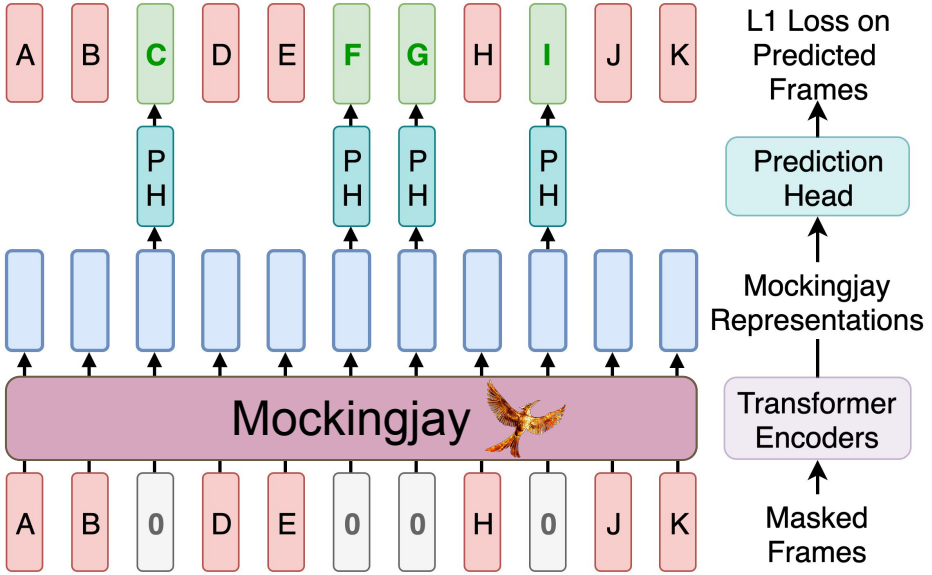
need to 1) shorten the sequence and 2) mask over a longer span



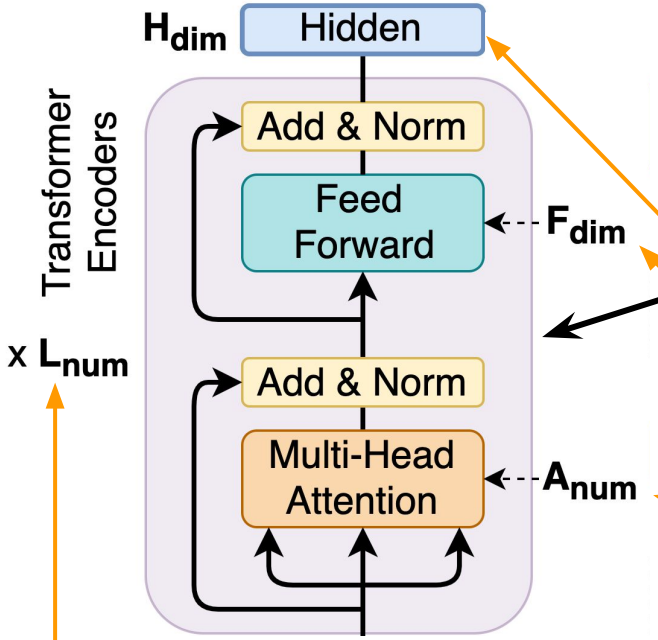
Address the long and smooth problem with:
Downsampling, and *consecutive masking*



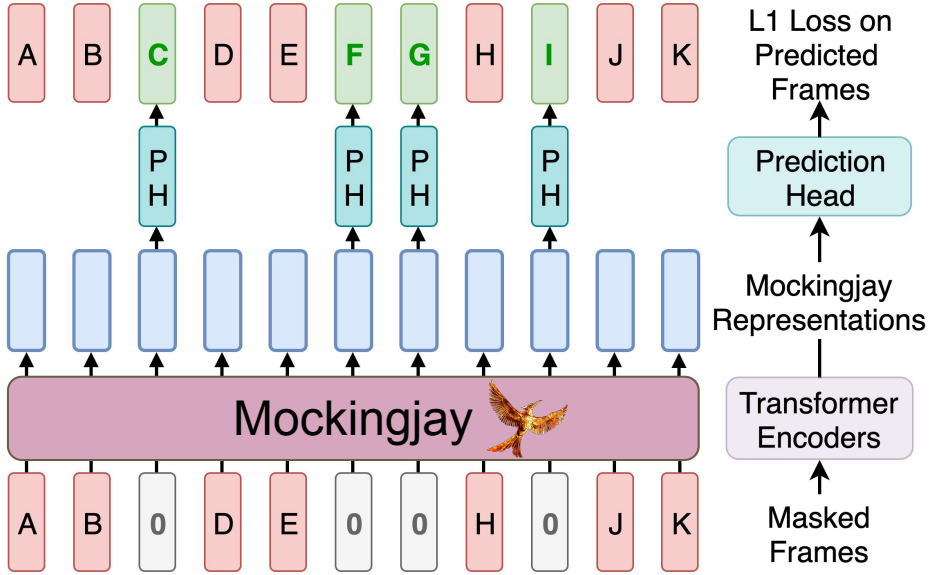
Model Architecture



Model Architecture



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

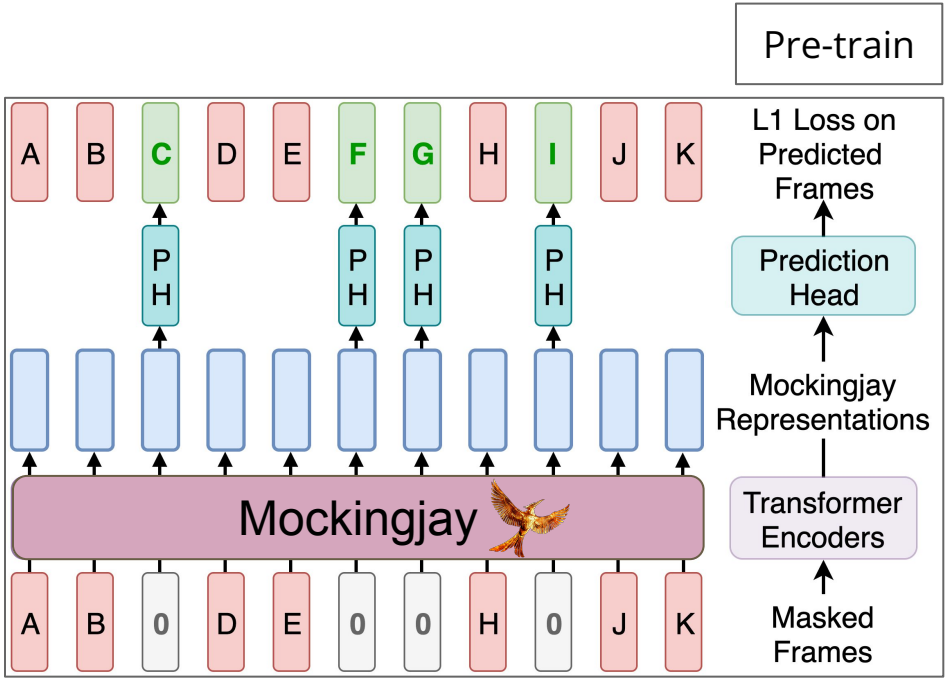


- H_dim = **768**
- F_dim = **3072**
- A_num = **12**

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

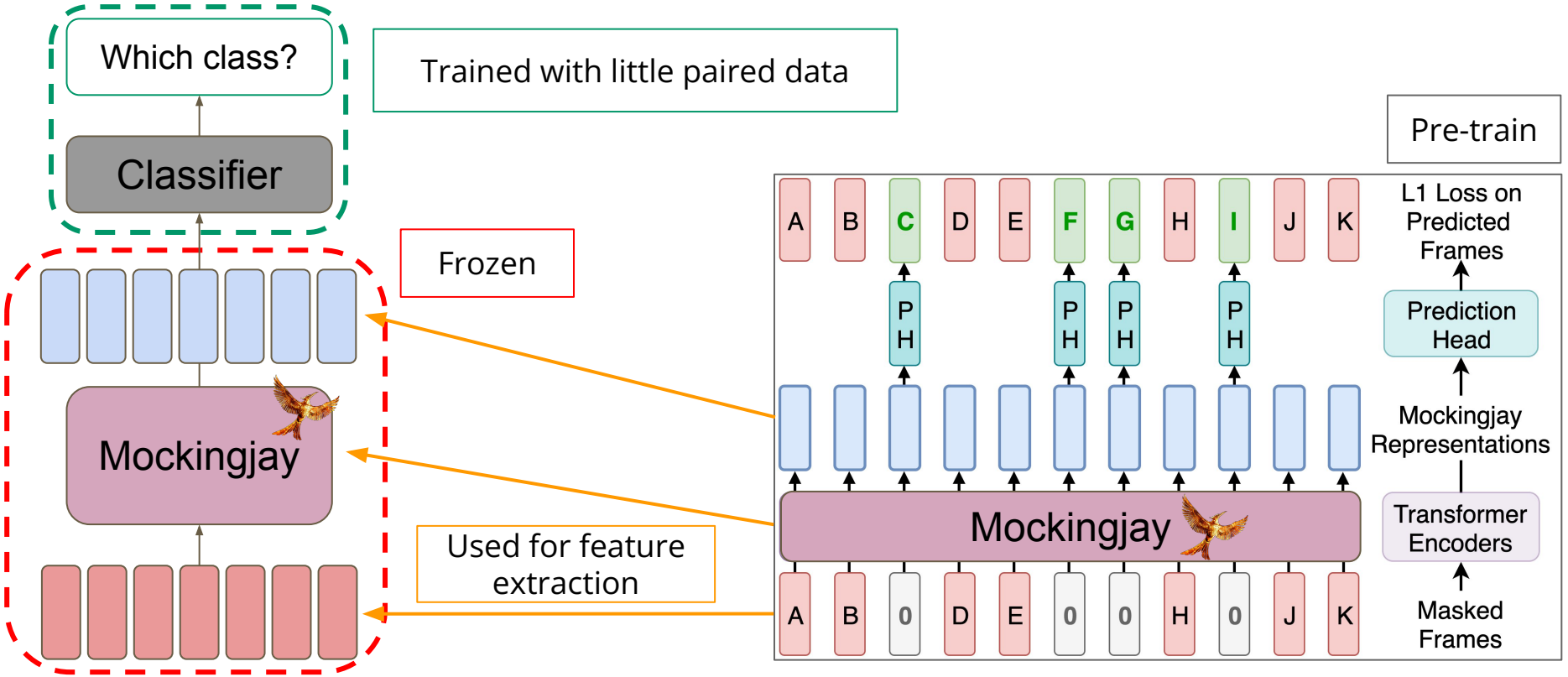
Incorporating with Downstream Tasks

1) Feature Extraction



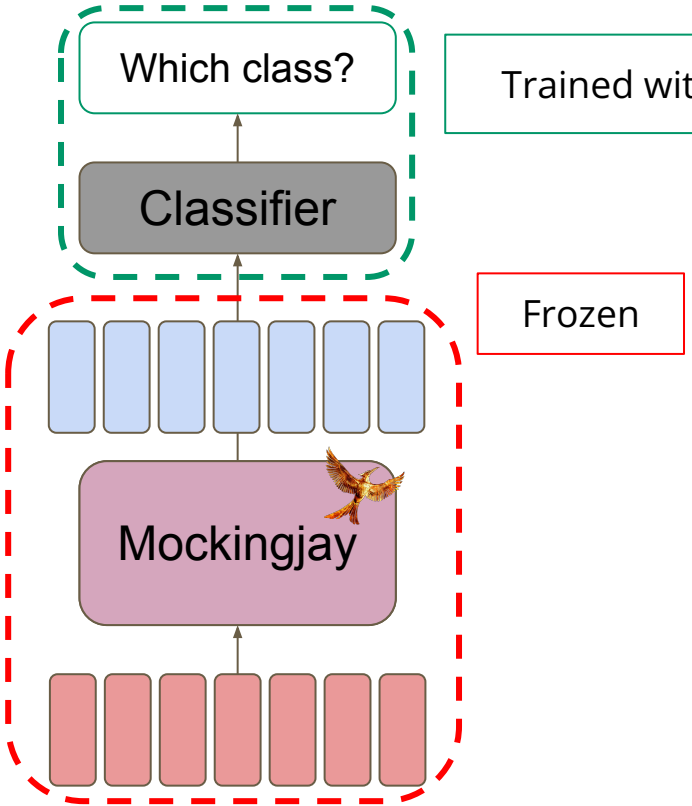
Incorporating with Downstream Tasks

1) Feature Extraction



Incorporating with Downstream Tasks

2) Weighted Sum from All Layers (WS)

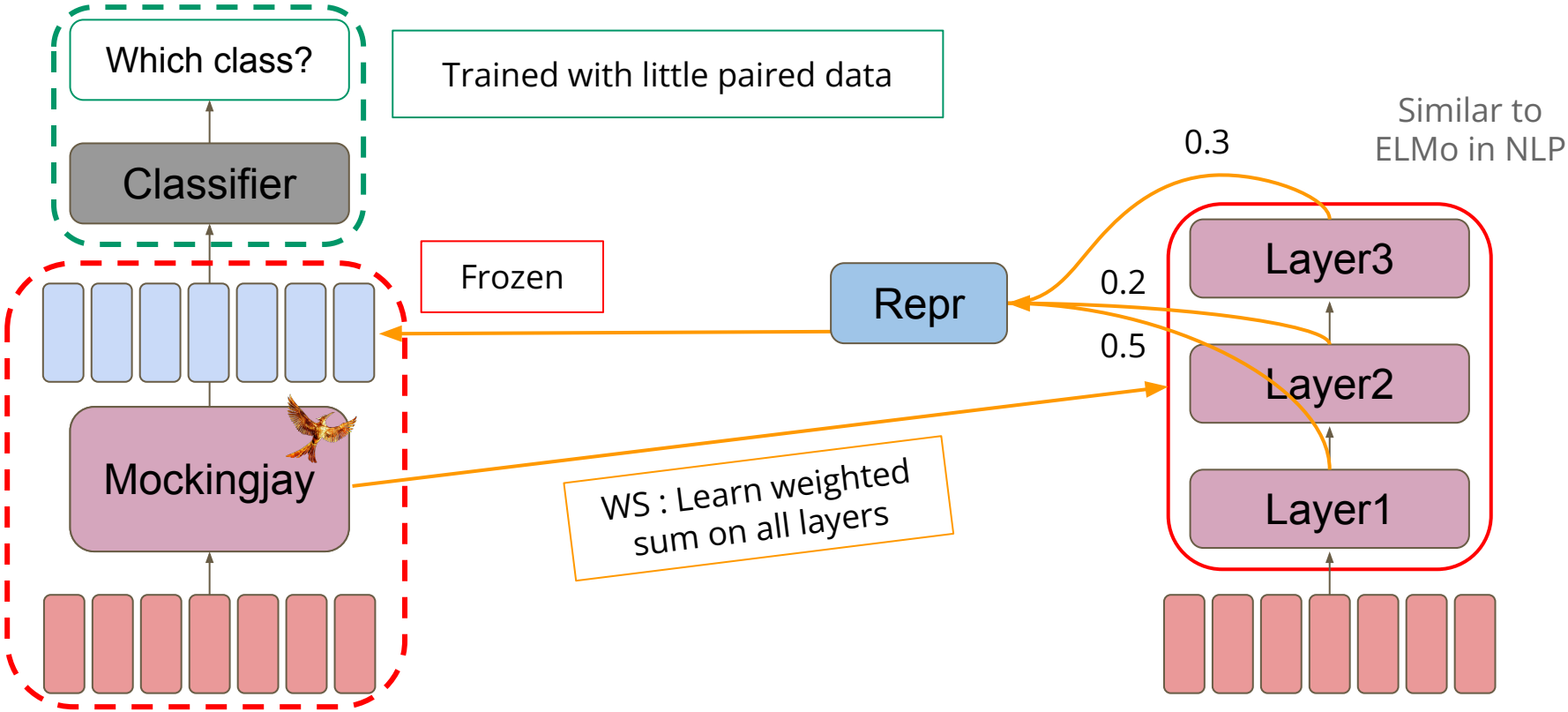


Trained with little paired data

Frozen

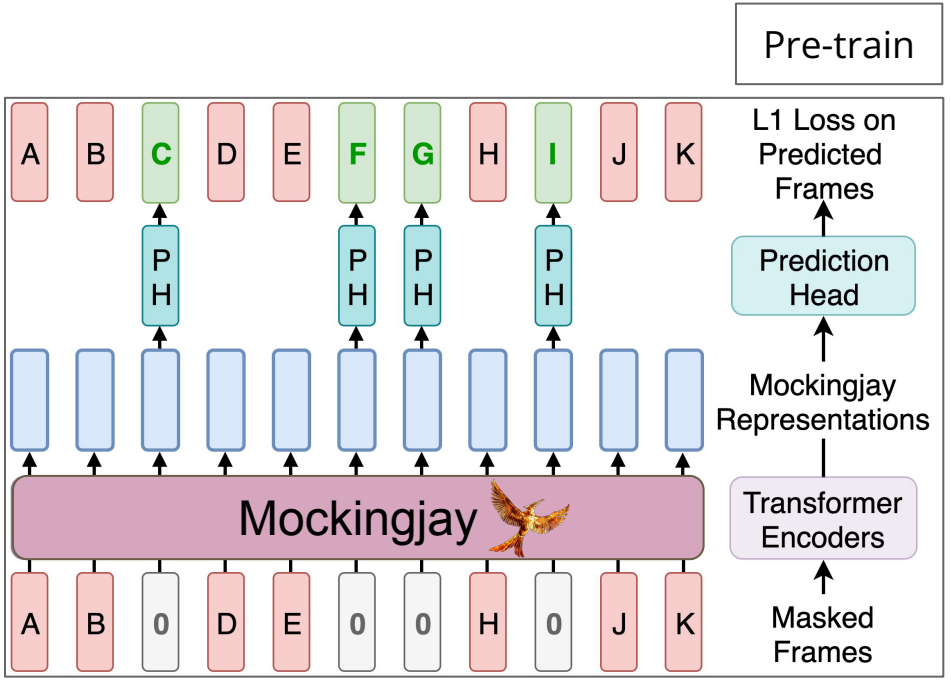
Incorporating with Downstream Tasks

2) Weighted Sum from All Layers (WS)



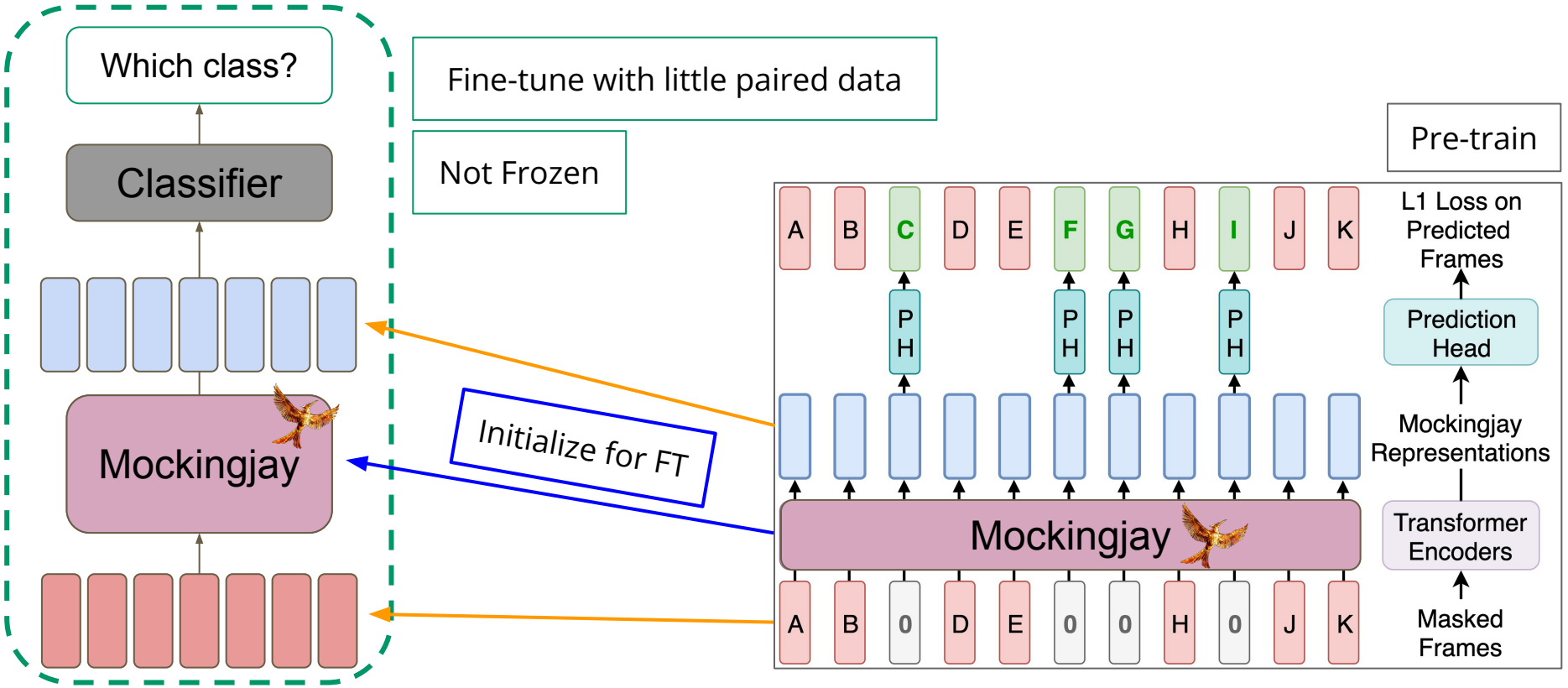
Incorporating with Downstream Tasks

3) Fine-tune (FT2)



Incorporating with Downstream Tasks

3) Fine-tune (FT2)



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**

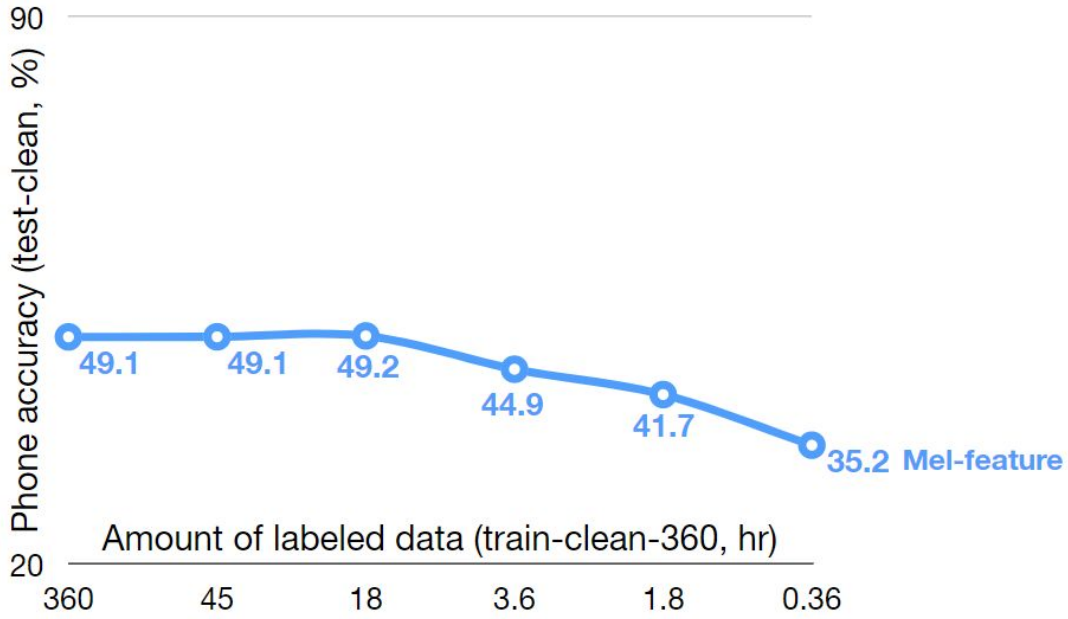
Consistent results over all three tasks:
Mel < BASE < LARGE

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**

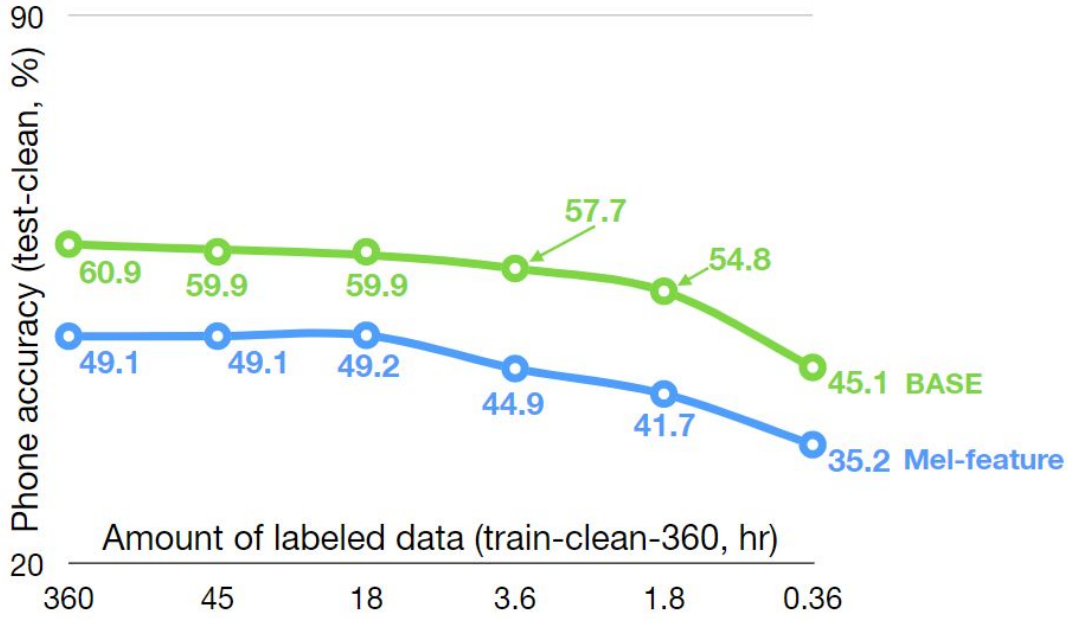
Consistent results over all three tasks:
LARGE < LARGE-WS

Low-Resource Experiments - 1/6



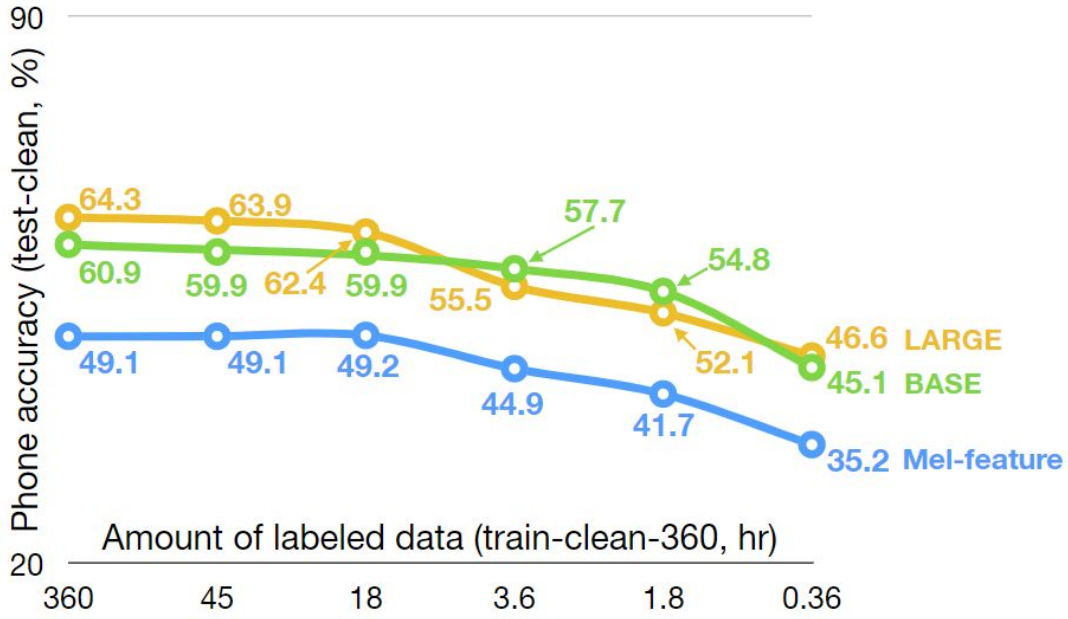
We demonstrate how pre-training on speech can improve supervised training in low resource scenarios, we train with reduced amount of labels.

Low-Resource Experiments - 2/6



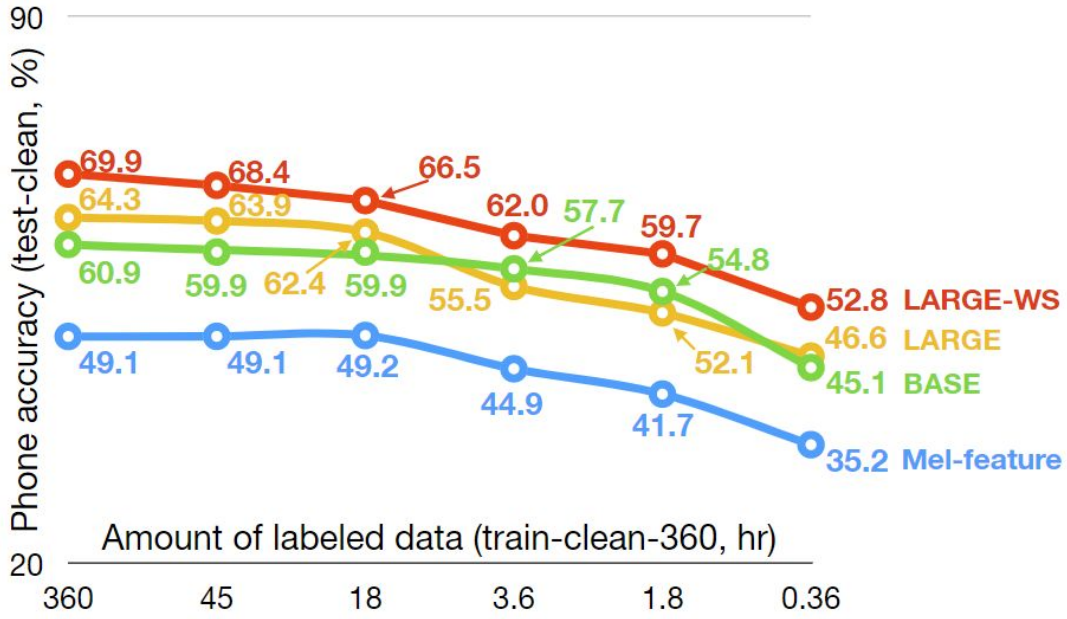
Mel < BASE

Low-Resource Experiments - 3/6



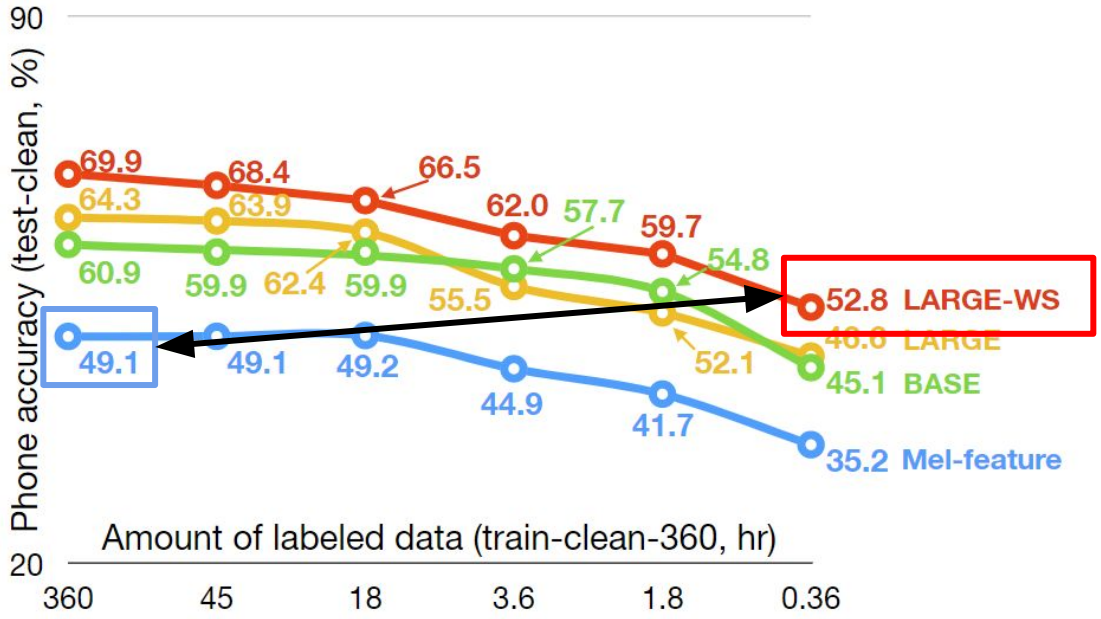
Mel < BASE < LARGE

Low-Resource Experiments - 4/6



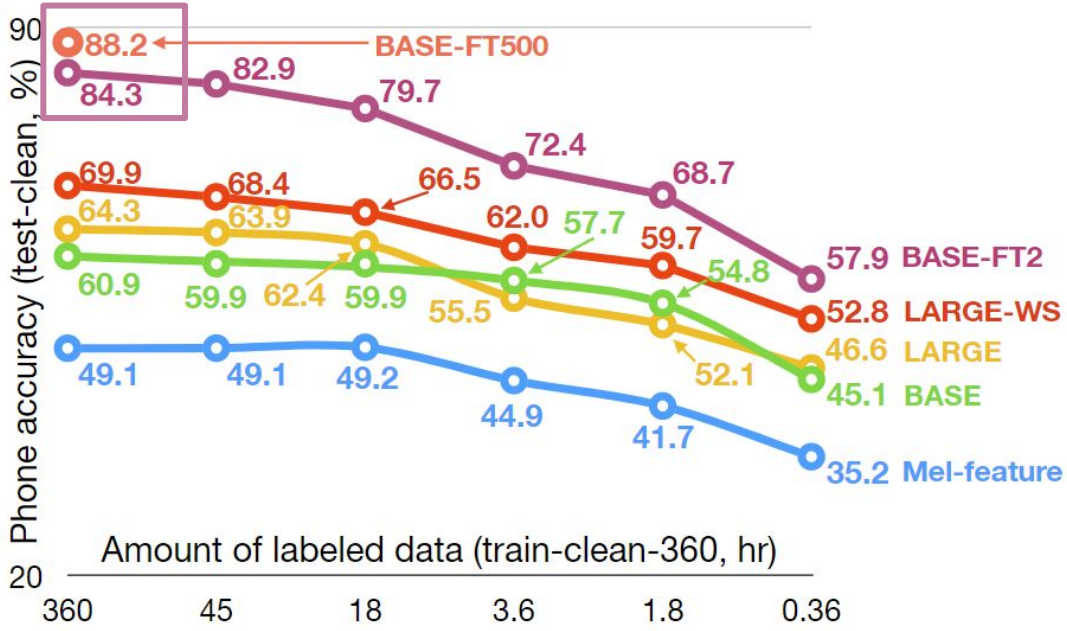
LARGE < LARGE-WS
with an avg 5.75% improvement

Low-Resource Experiments - 4/6



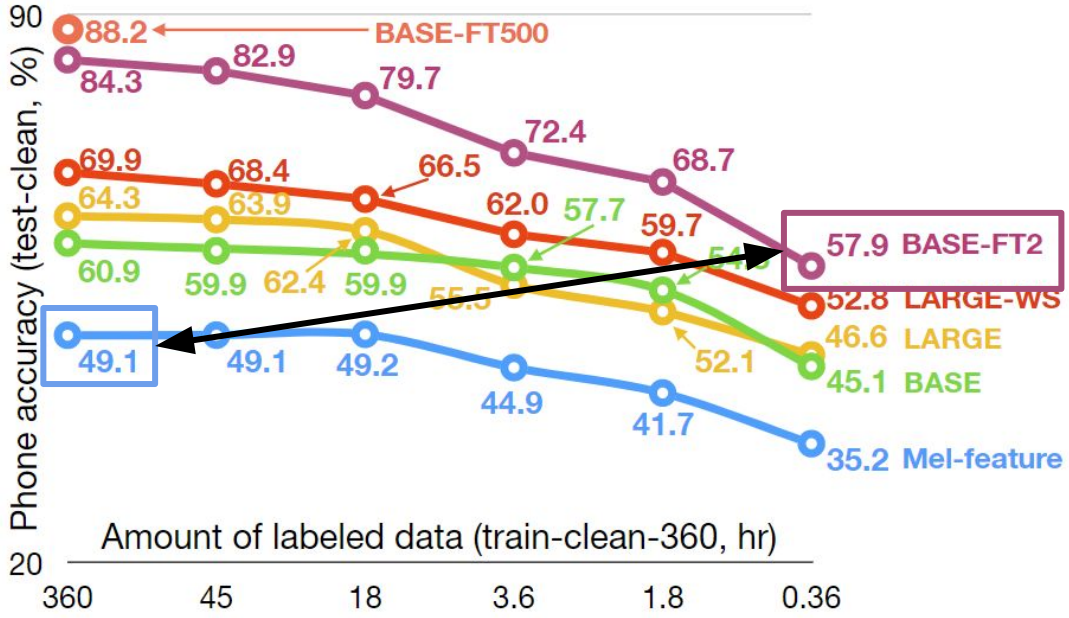
With 0.1% of labels,
LARGE-WS (52.8%) outperformed Mel (49.1%) that uses all 100% hours of labeled data.

Low-Resource Experiments - 5/6



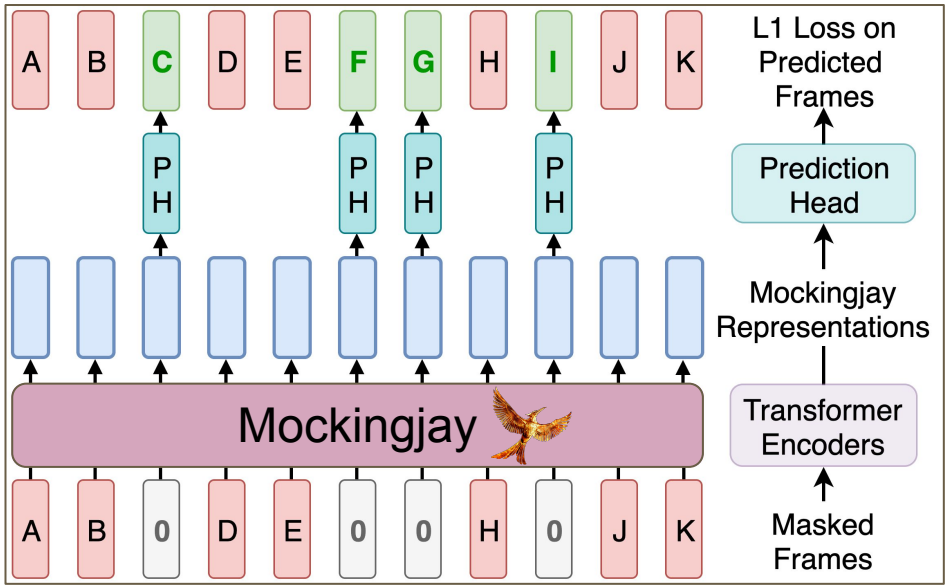
All < BASE-FT2

Low-Resource Experiments - 6/6



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

A Broad Introduction



Intuition:
A model that can predict the partial loss of small segments of speech, should provide a contextualized understanding of previous and later content.

vq-wav2vec → vq-wav2vec-FT

FB FB

Alibaba

BERT-Style

SLU BERT

Speech XLNet — Tencent

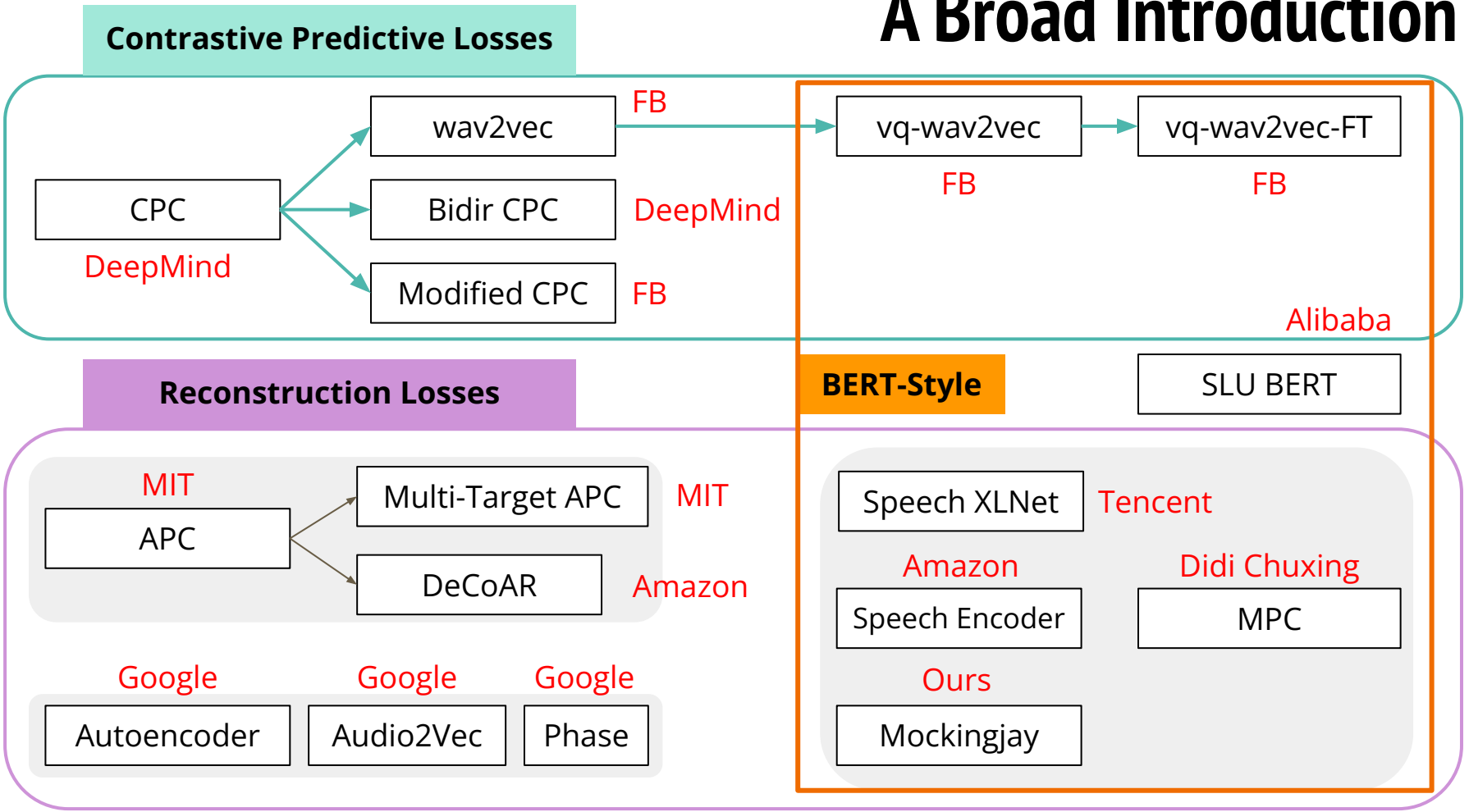
Amazon — Didi Chuxing

Speech Encoder — MPC

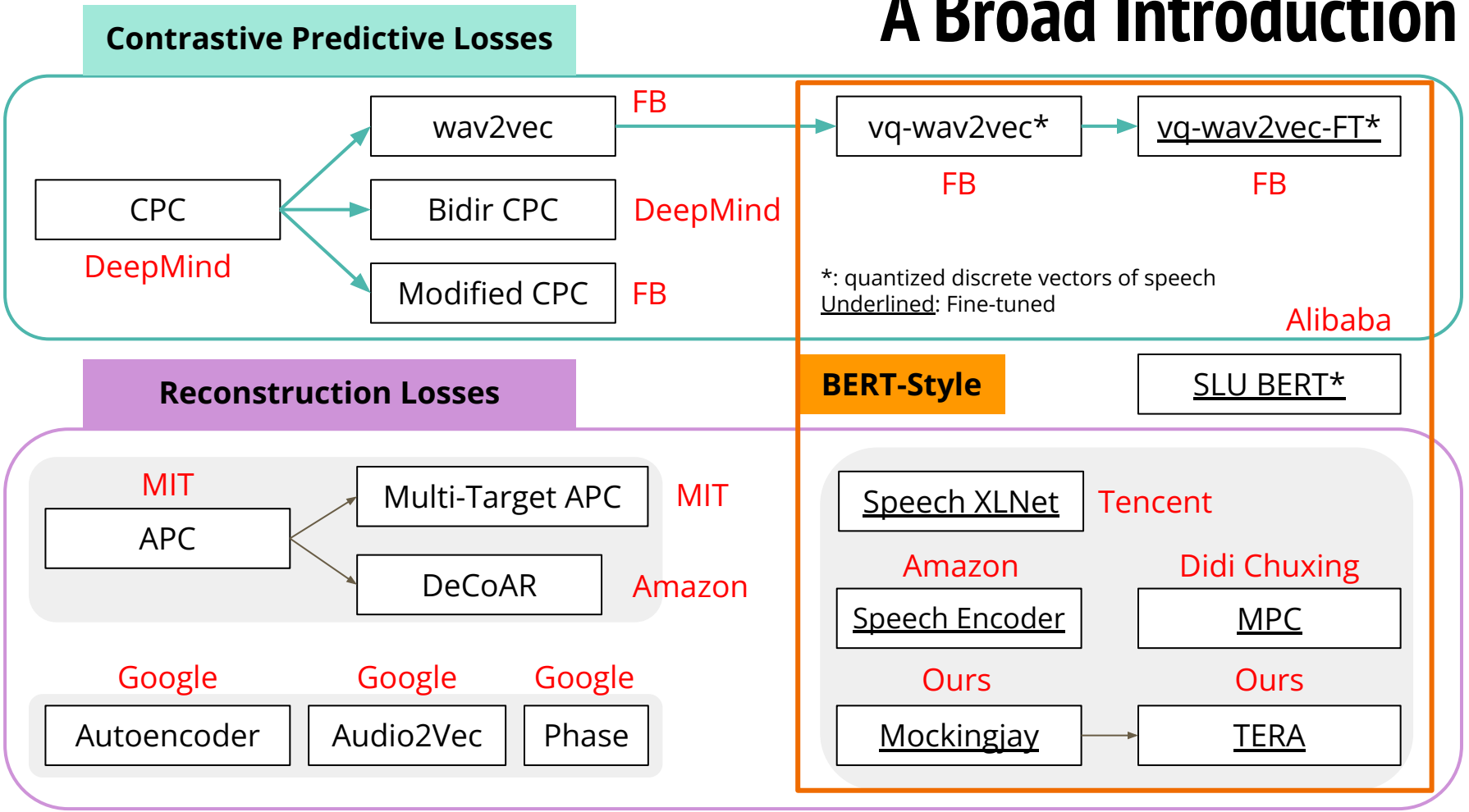
Ours

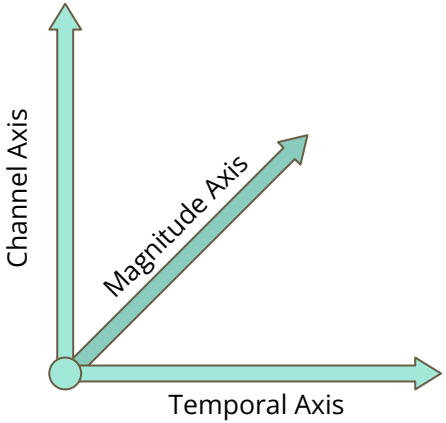
Mockingjay

A Broad Introduction



A Broad Introduction



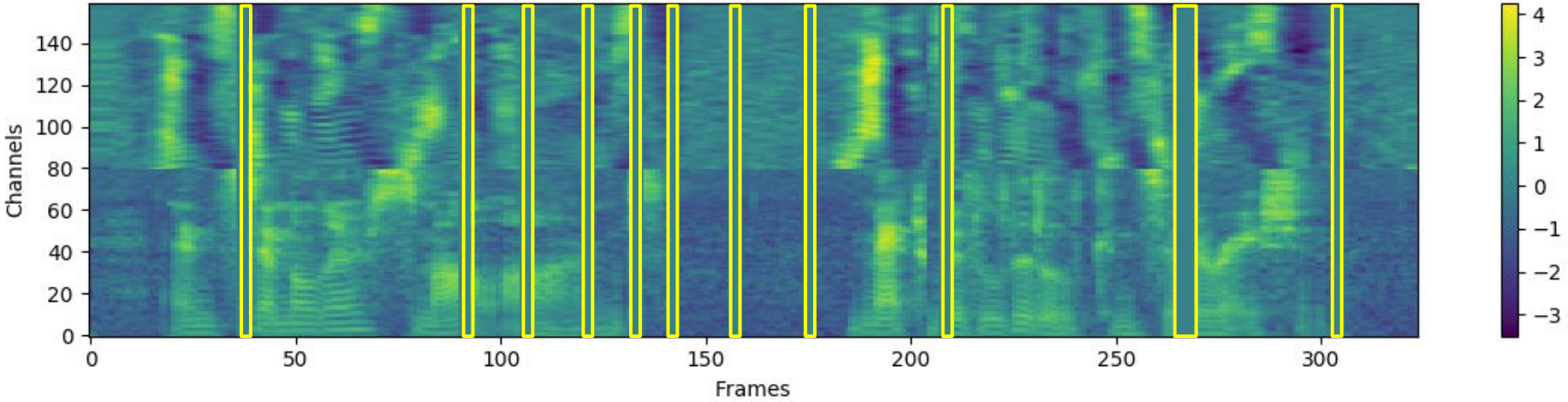


TERA

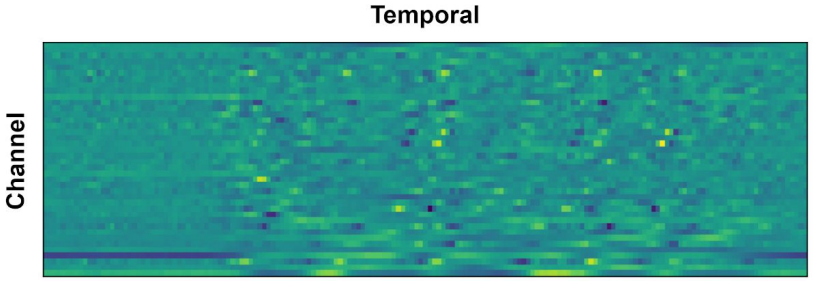
Transformer Encoder Representations from Alteration

Extending Mockingjay to multi-target learning on three dimensions

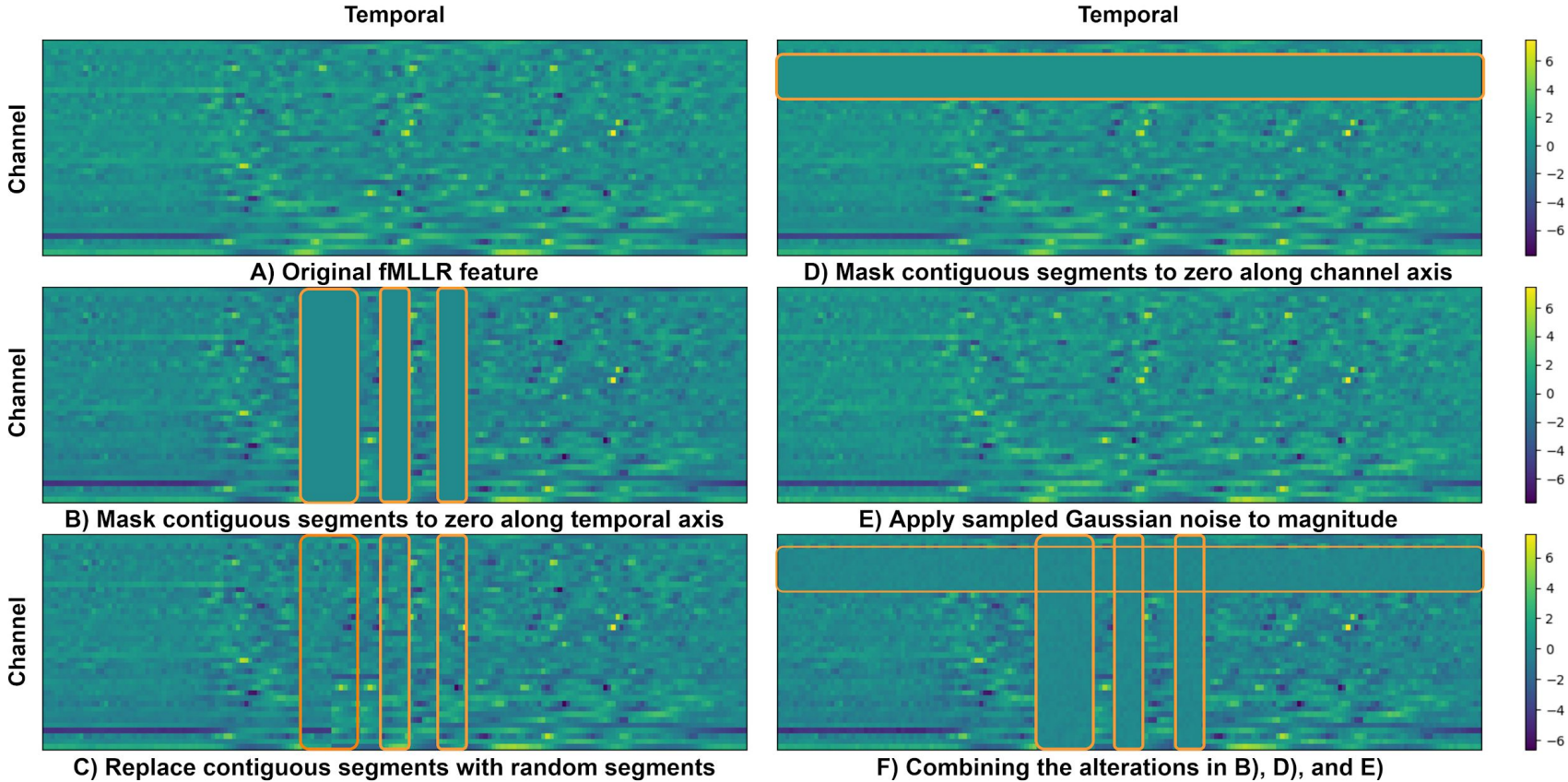
Recall: we mask mel spectrogram on time axis



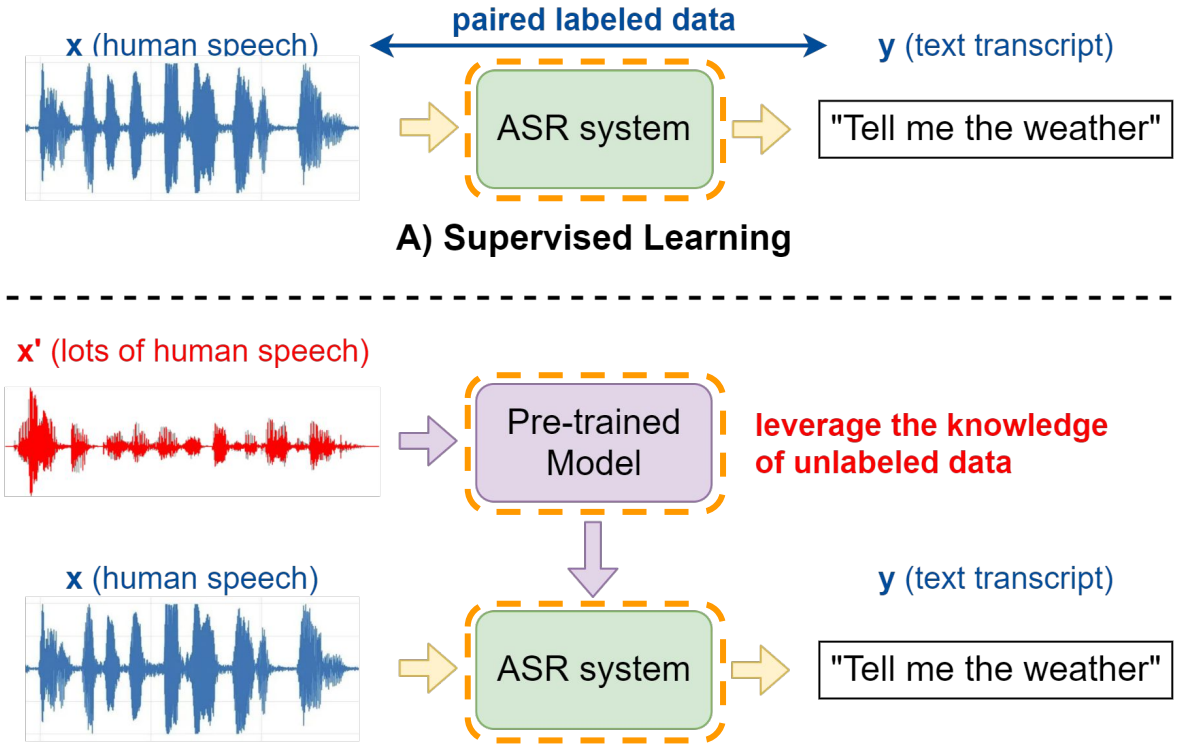
Consider fMLLR on 3 Axis:



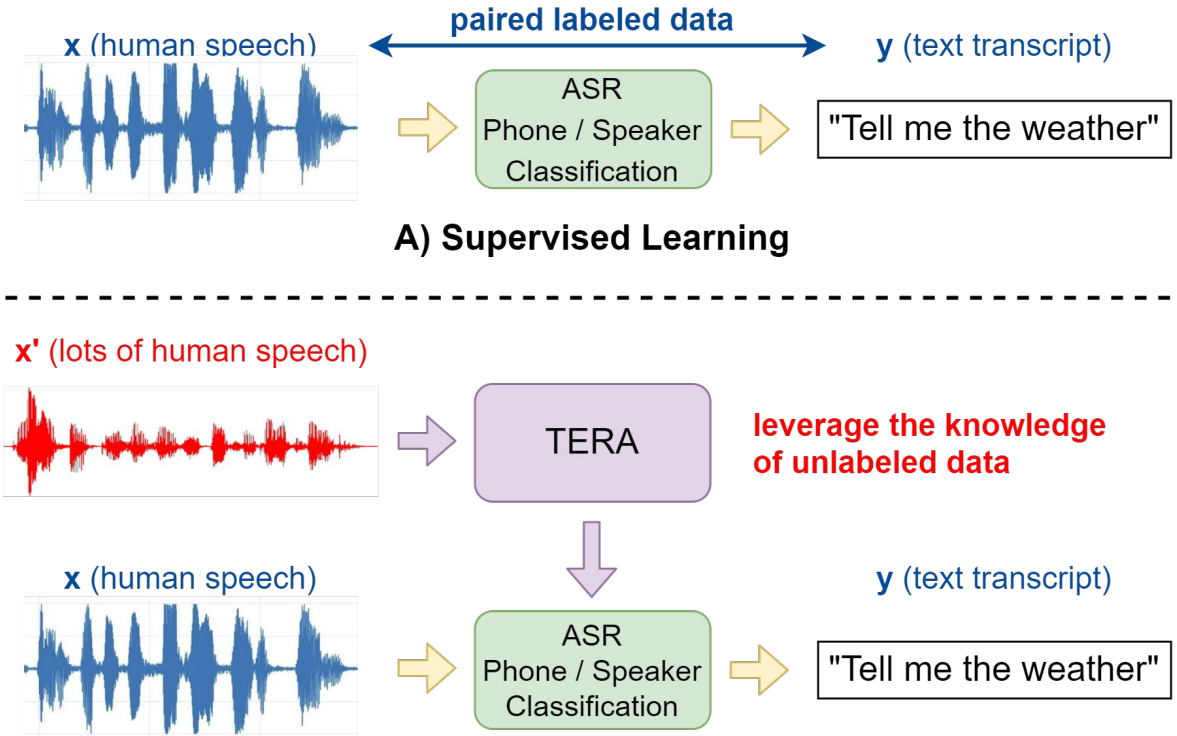
Multi-target Pre-training



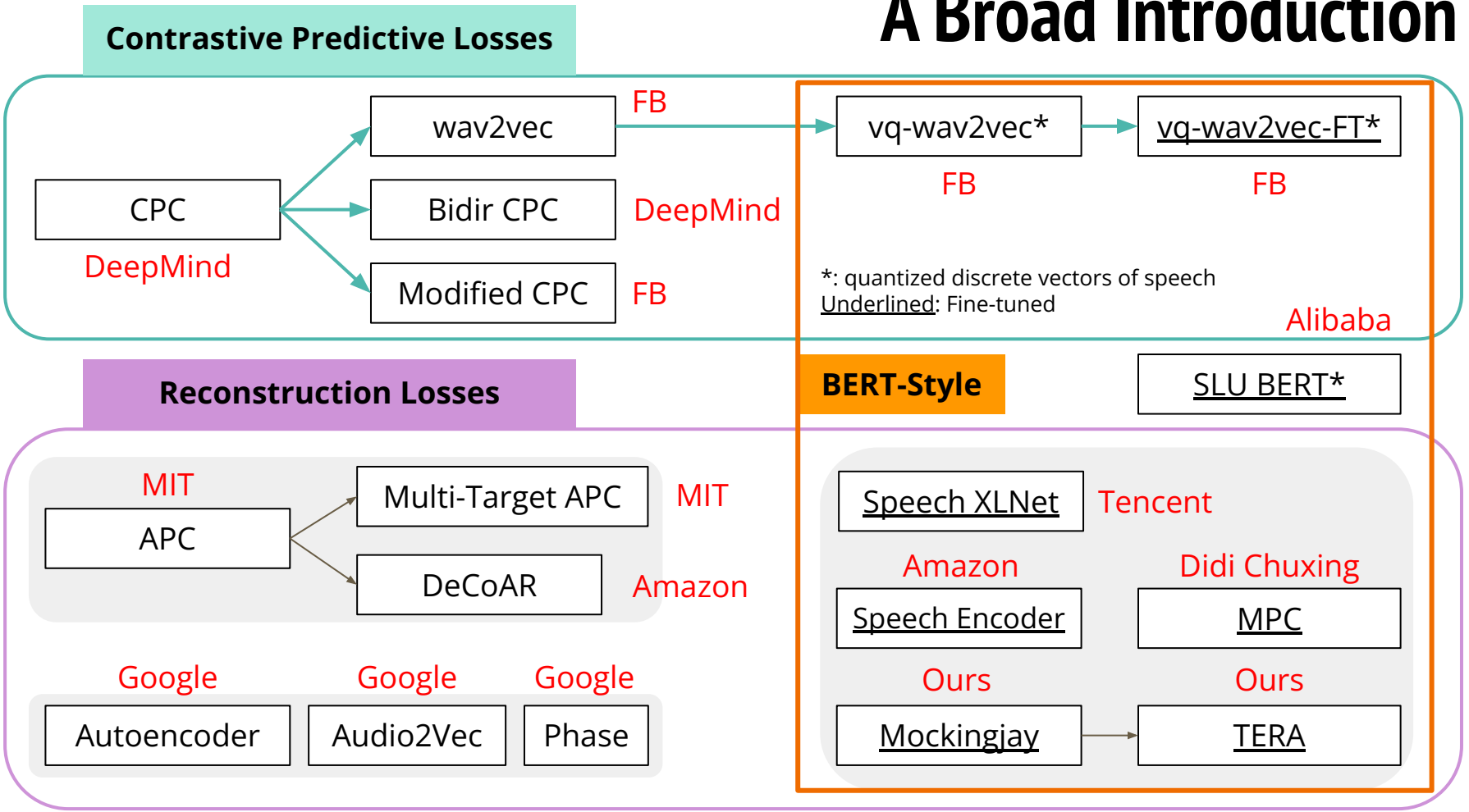
Recall: Self-Supervised Learning for Speech



Self-Supervised Learning: TERA

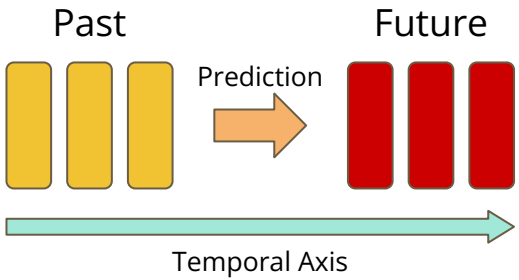


A Broad Introduction



Quick Summary

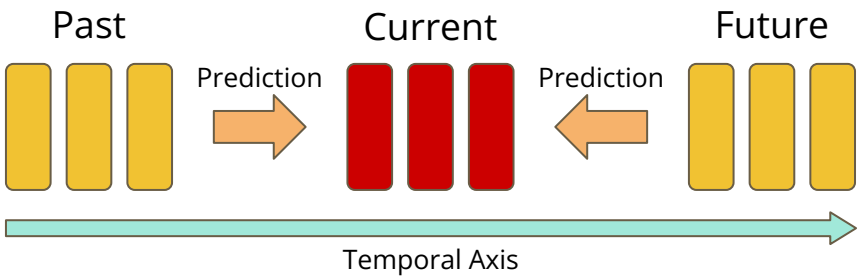
The design of auxiliary task fundamentally defines what the model learns!



CPC | wav2vec | Bidir CPC

APC | Multi-Target APC | Modified CPC

Autoencoder | Phase



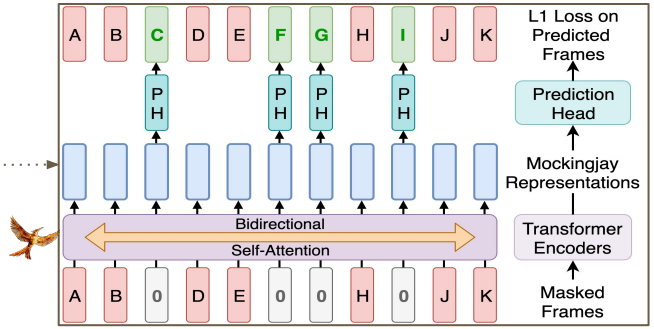
vq-wav2vec | vq-wav2vec-FT | SLU BERT

DeCoAR | Audio2Vec | Speech XLNet

Mockingjay | MPC | Speech Encoder

From here to beyond

- Mockingjay *(ICASSP 2020)*



Submitted to InterSpeech 2020 (5/15)

1. **Mockingjay for Adversarial Defence**
2. **Improving Mockingjay: Speech ALBERT**
3. **Understanding Self-Attention**

What else?

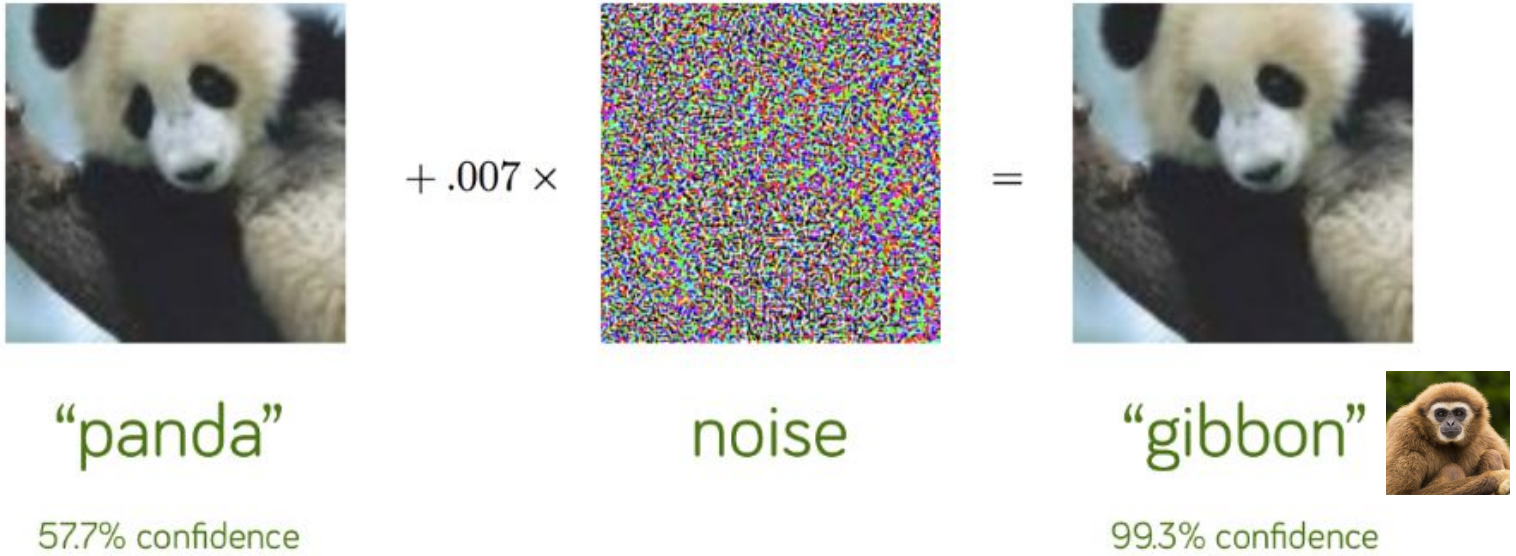
What else can we do with Mockingjay?

1. Adversarial Defense

Employ Mockingjay to protect models against adversarial attacks

1. Adversarial Defense

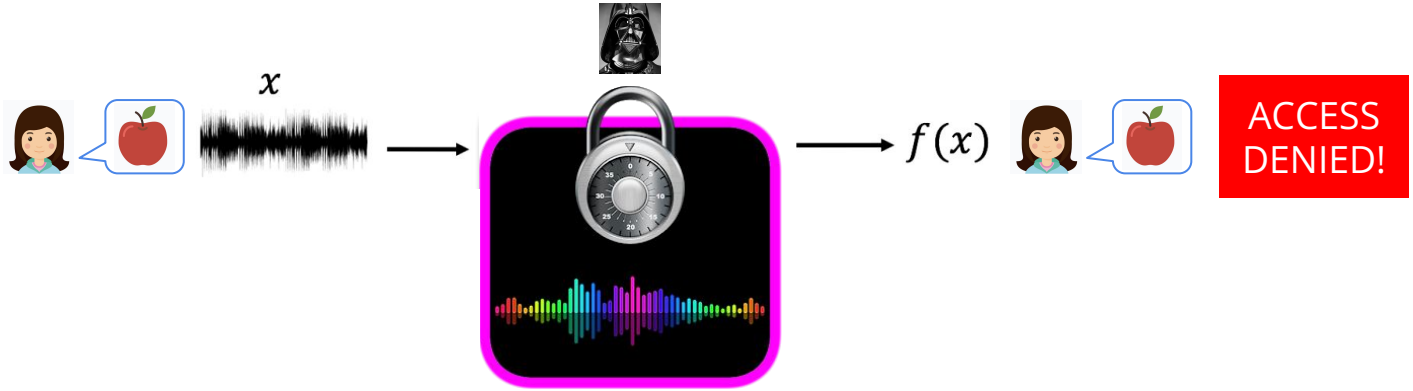
What is Adversarial Attack?



Hacking AI security systems: Face ID / Voice ID

1. Adversarial Defense

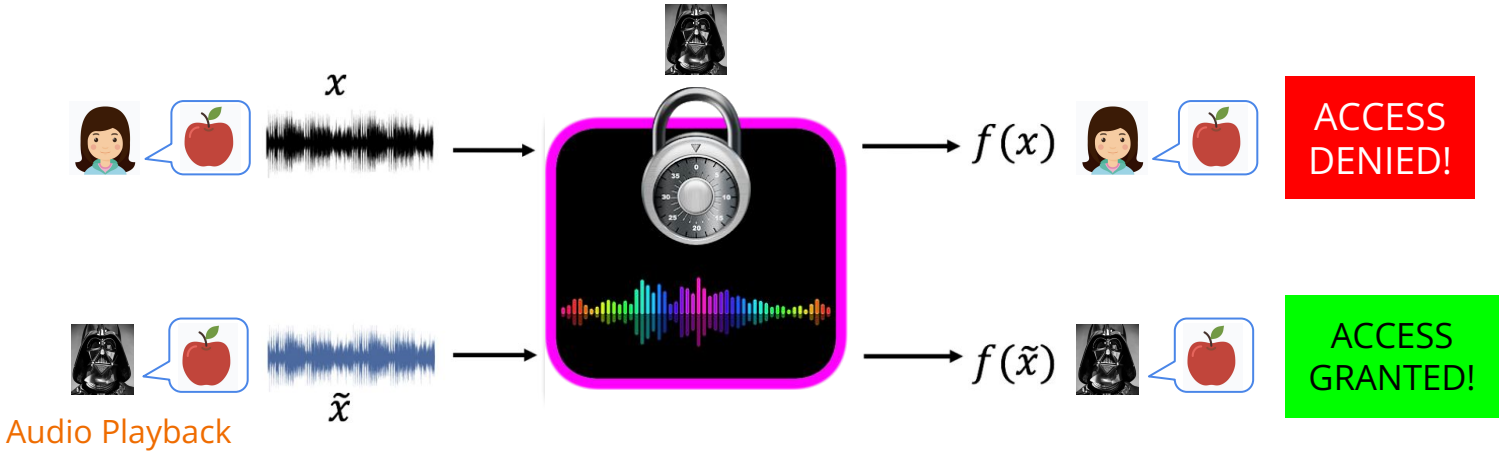
What is Adversarial Attack?



Hacking AI security systems: Face ID / Voice ID

1. Adversarial Defense

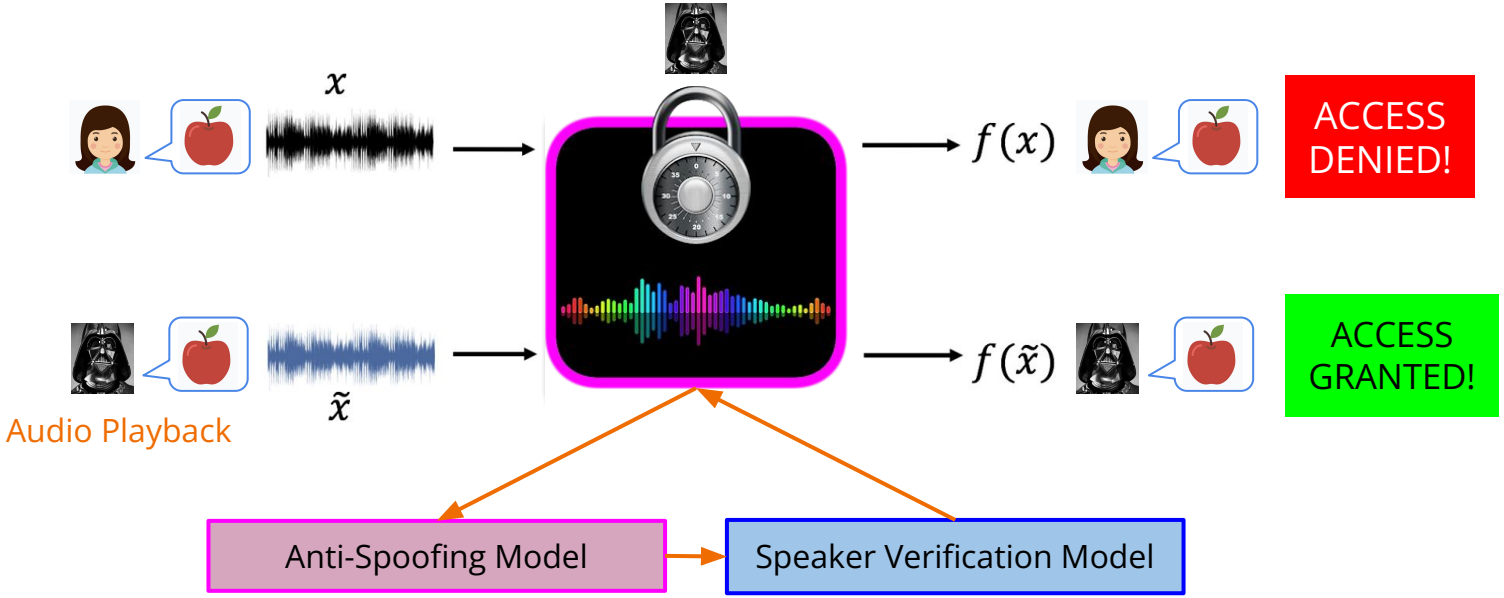
What is Adversarial Attack?



Hacking AI security systems: Face ID / Voice ID

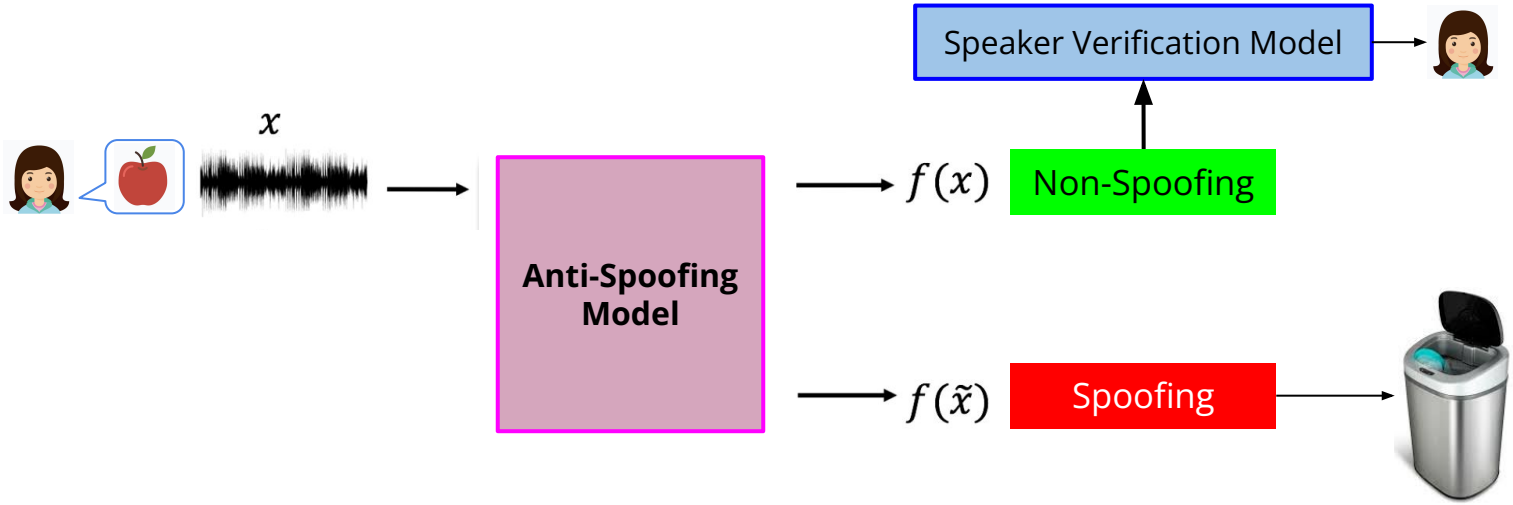
1. Adversarial Defense

Anti-Spoofing Model



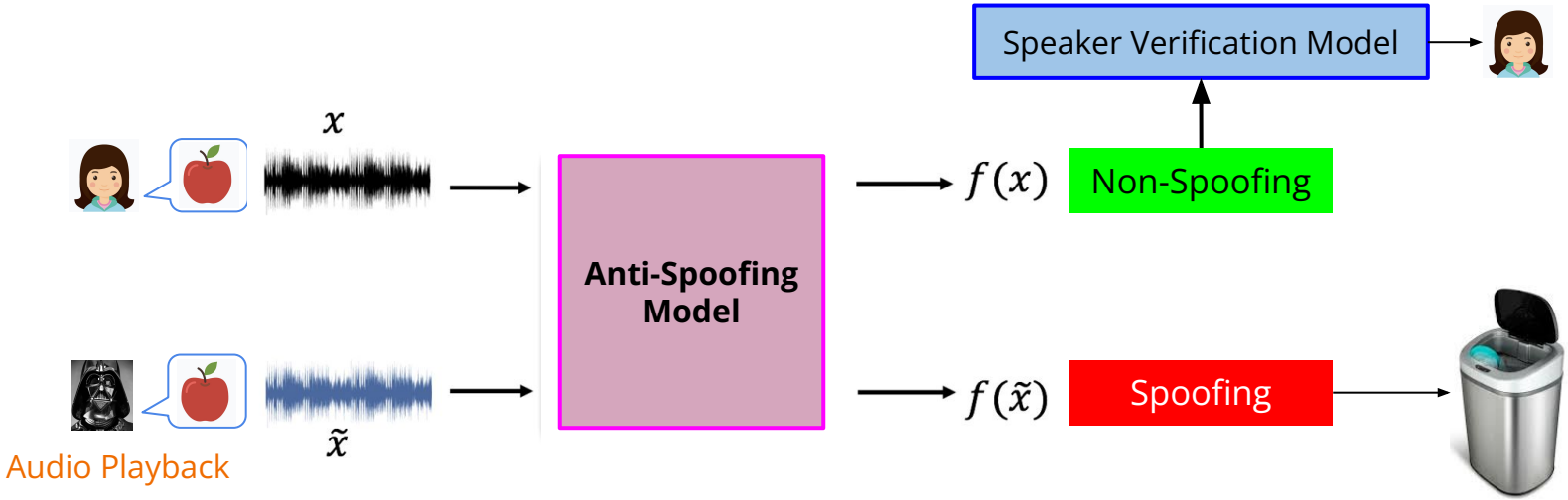
1. Adversarial Defense

Anti-Spoofing Model



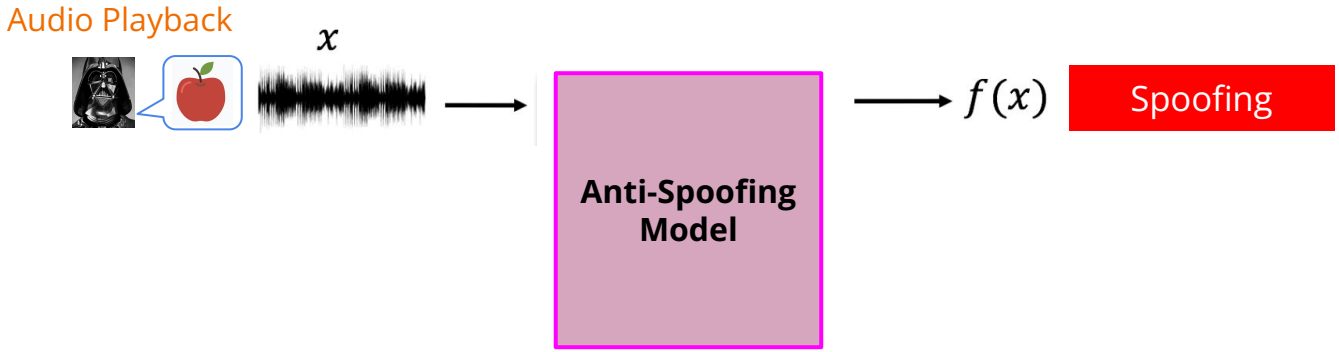
1. Adversarial Defense

Anti-Spoofing Model



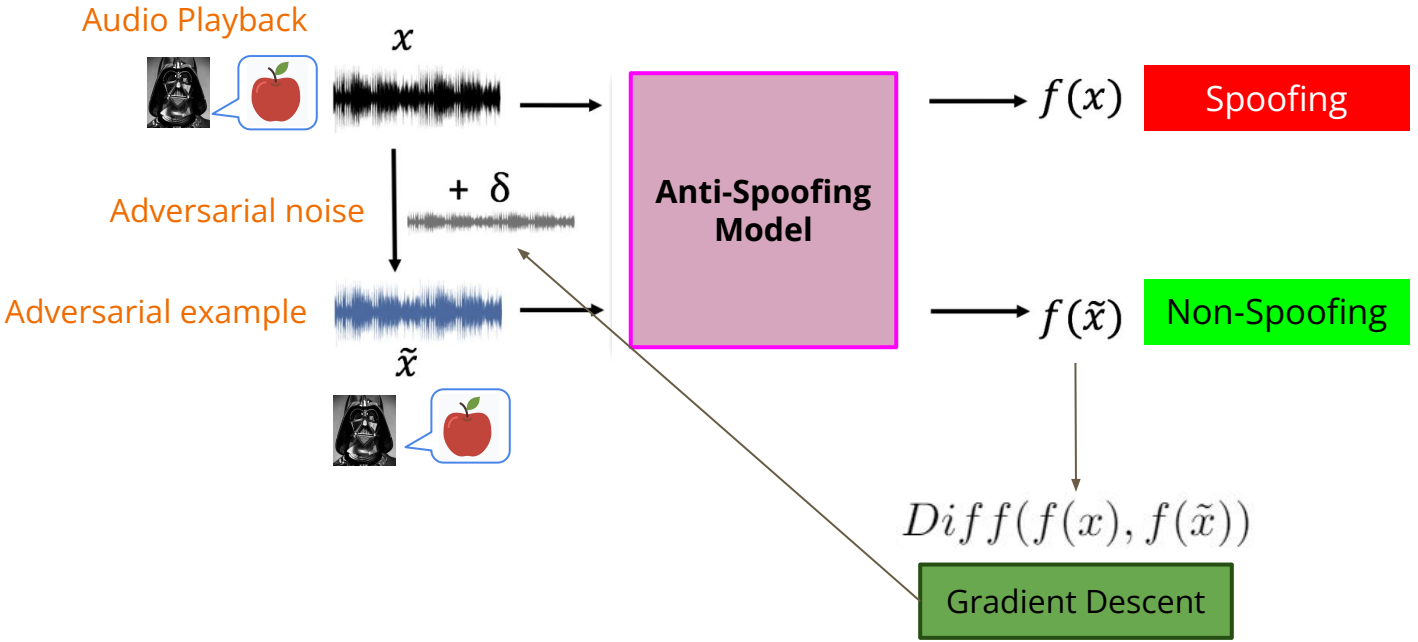
1. Adversarial Defense

How to Attack?



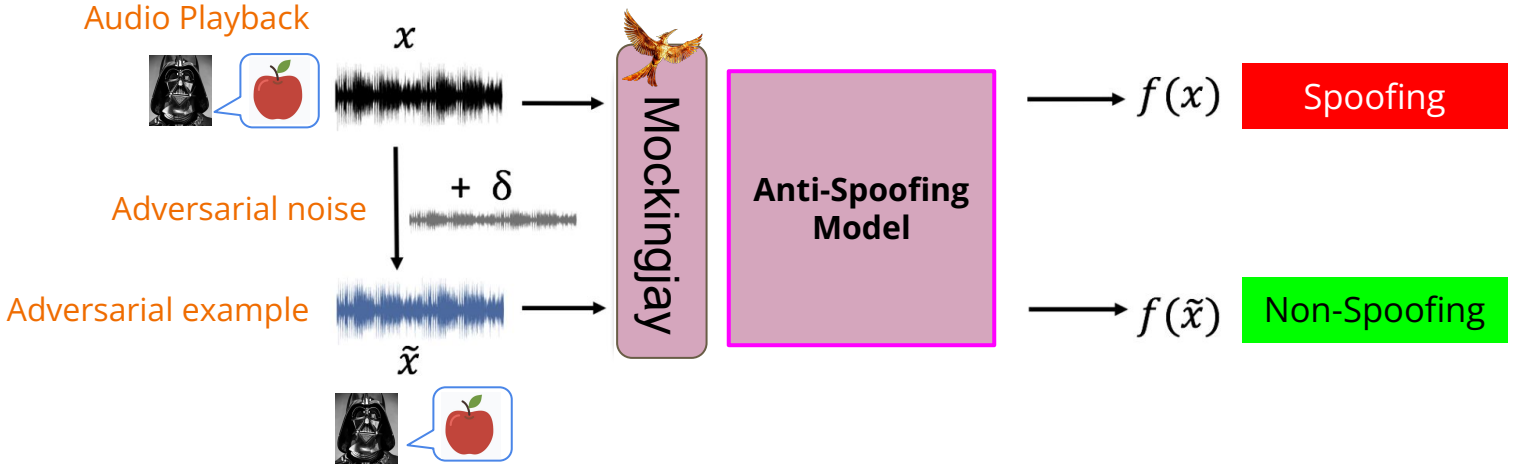
1. Adversarial Defense

How to Attack?

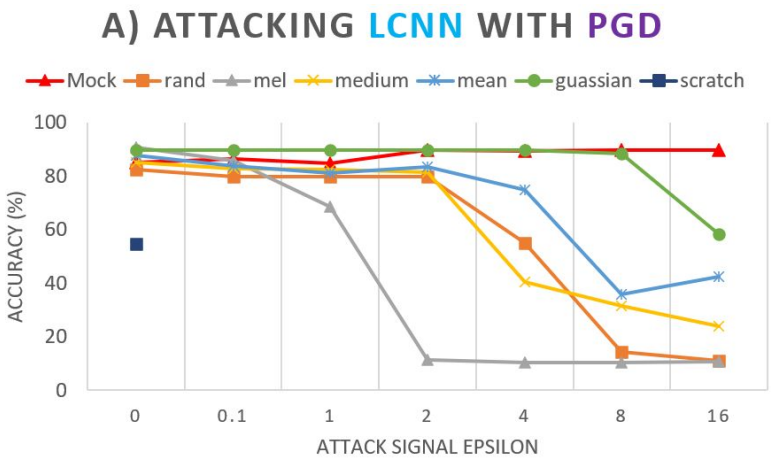


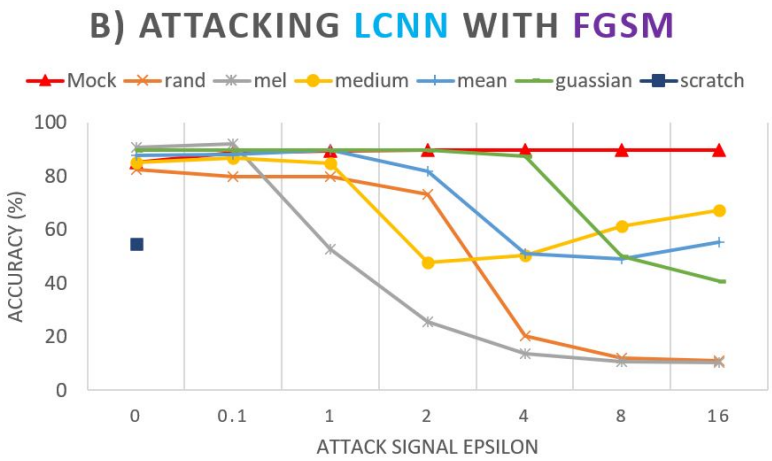
1. Adversarial Defense

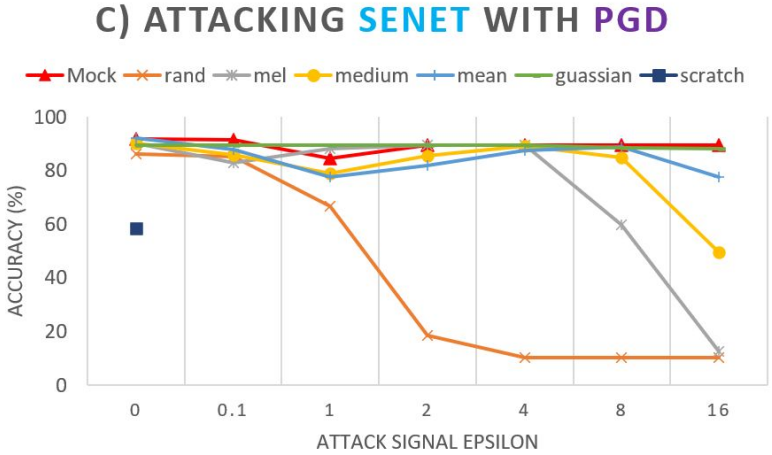
Employing Mockingjay

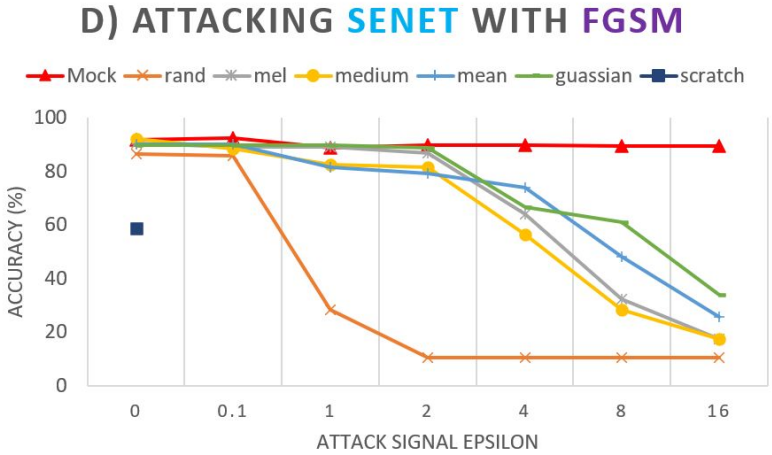


1. Adversarial Defense - Experiments

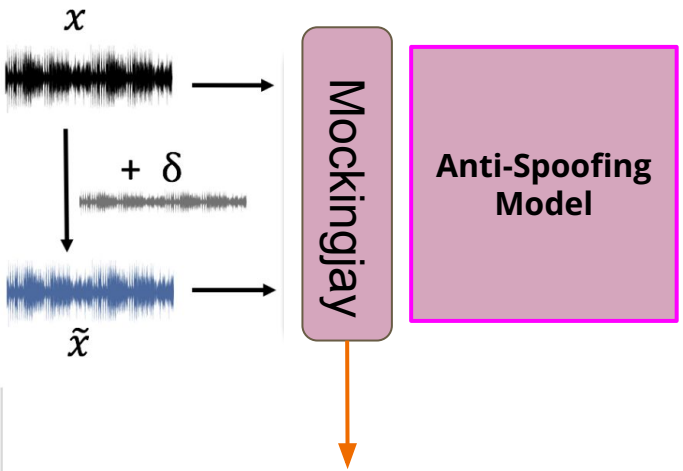


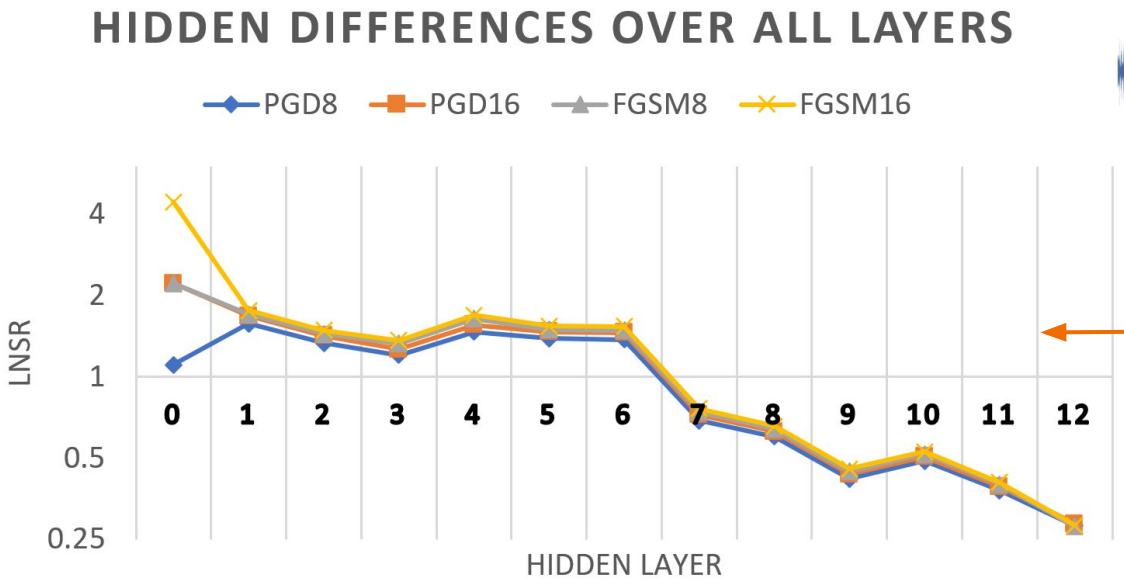






1. Adversarial Defense - Experiments





Intuition:
LNSR- Measure the amount of adversarial signal through the layers

$$LNSR_i = \sum_{n=1}^{N} \frac{\|\hat{h}_i^n - h_i^n\|_2}{\|h_i^n\|_2}$$

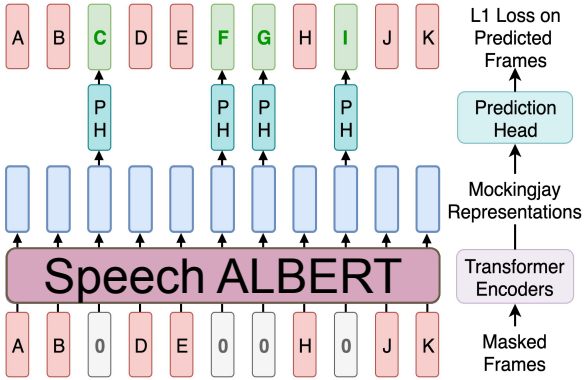
What else?

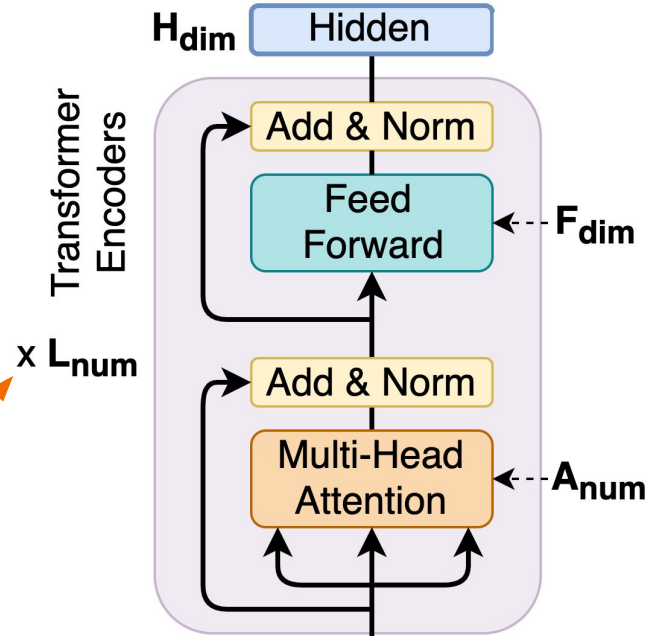
What else can we do with Mockingjay?

Talk 2: Audio ALBERT

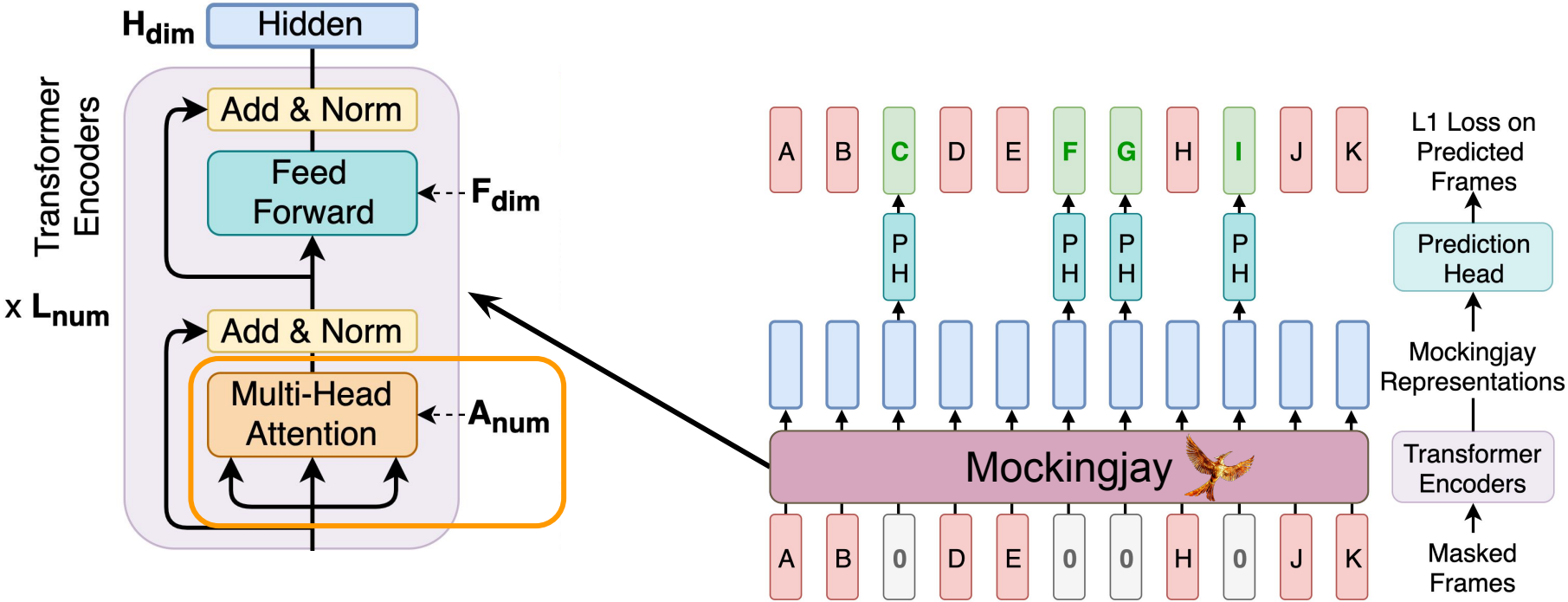
Recall that:
Audio BERT >>> Mockingjay

Now we also have:
ALBERT >>> Audio ALBERT





Talk 3: Understanding Self-Attention



Conclusion

Self-supervised learning,
a brand new topic with lots of ideas that we can work on!

Thank You

Q&A