Self-Supervised Learning

For Speech

Andy T. Liu

2020/05/14

Outline

- Overview
- Examples:
  - SSL on VC: ZeroSpeech 2019
  - SSL on NLP: BERT
- Speech BERT: Mockingjay
- Current Works: 4 InterSpeech 2020 Submissions
- Related Works
- Future Work: IEEE Journal Submission - TERA

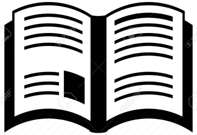
Overview

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.

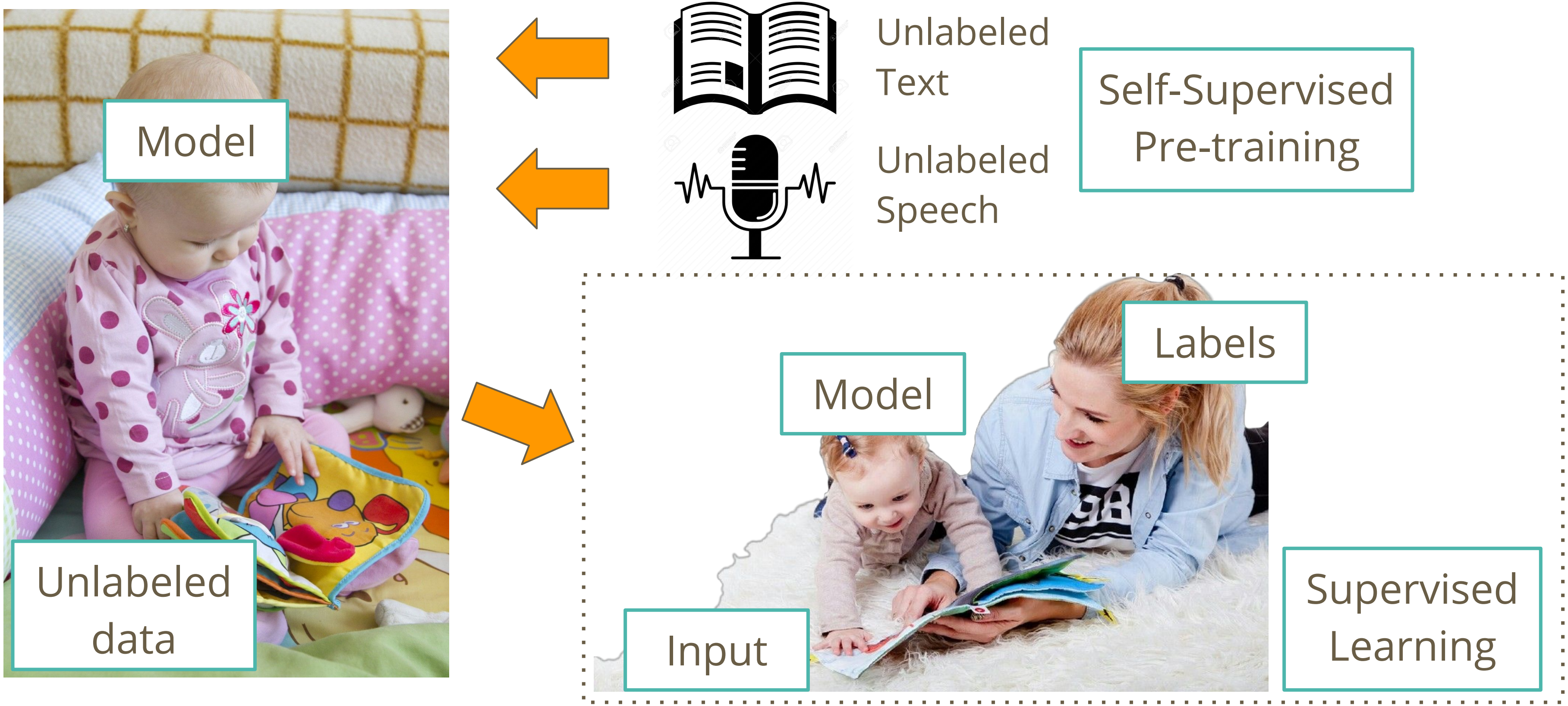
How do Infants Learn? Yes! Self-Supervised Learning

Unlabeled Text

Unlabeled Speech

Self-Supervised Pre-training

How do Infants Learn? Yes! Self-Supervised Learning



Self-Supervised Learning for Speech

paired labeled data

x (human speech)

ASR system

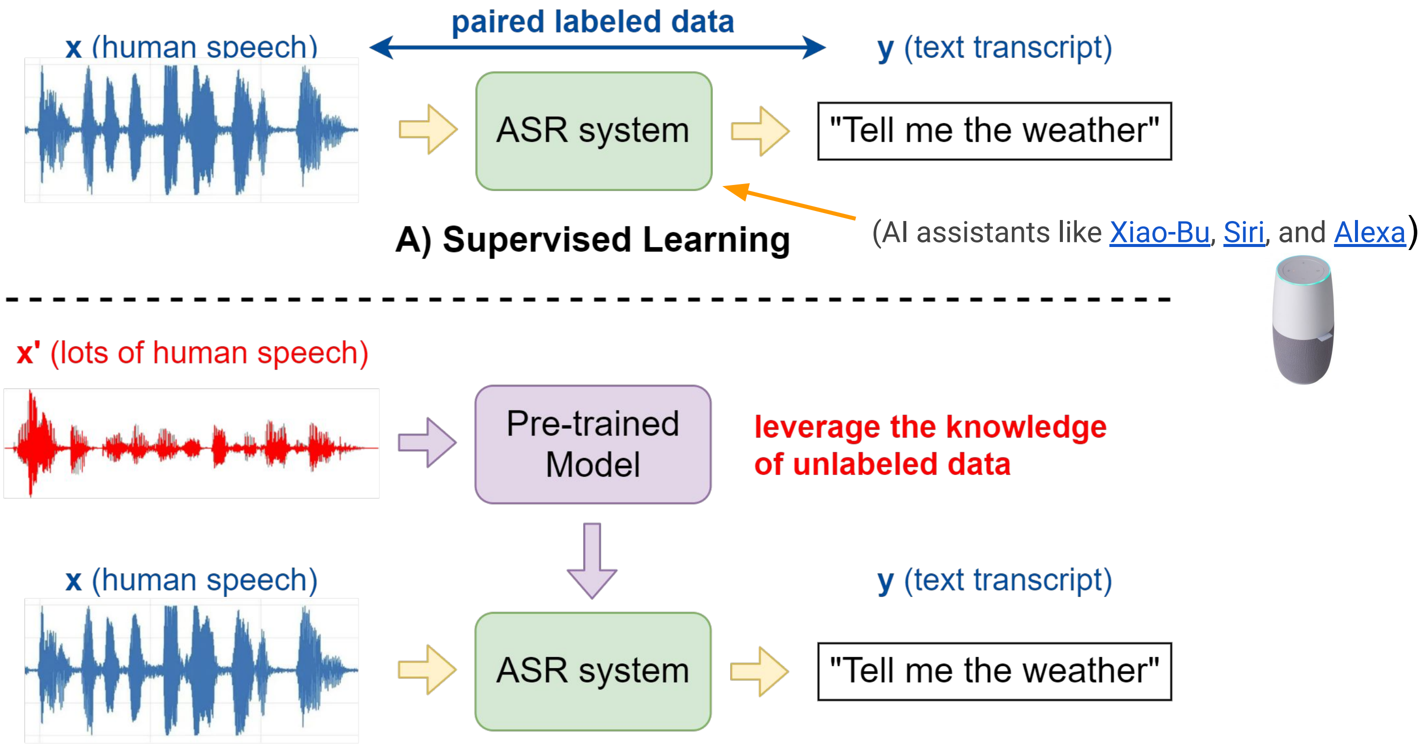
y (text transcript)

"Tell me the weather"

A) Supervised Learning

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

Self-Supervised Learning for Speech



Pre-trained models are evaluated on downstream tasks.

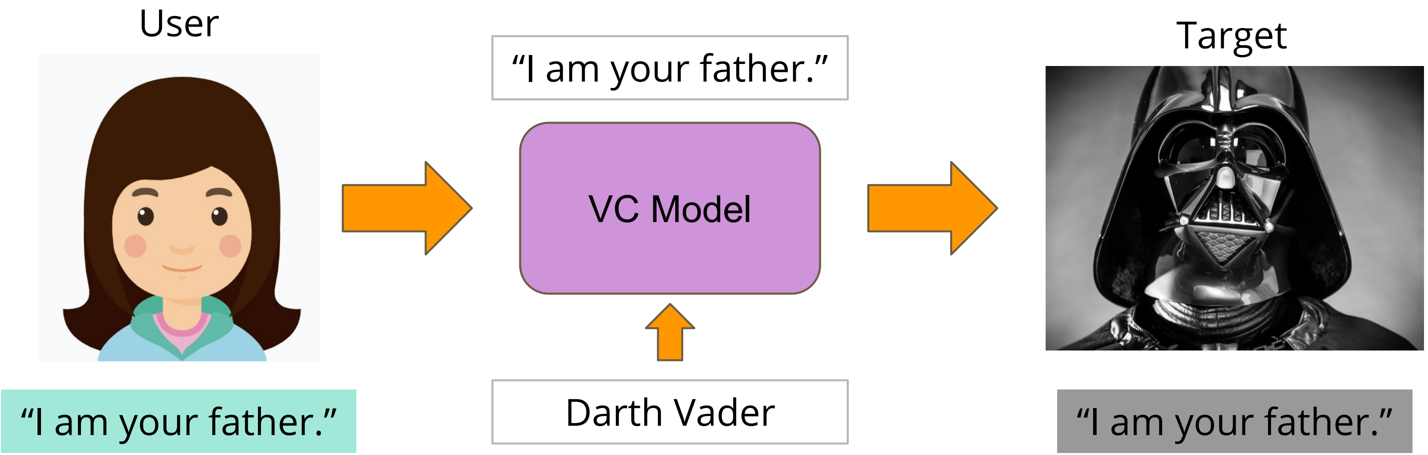
B) Self-Supervised Learning for Improving Supervised Systems

Examples

Voice Conversion and **BERT**

An example: Voice Conversion

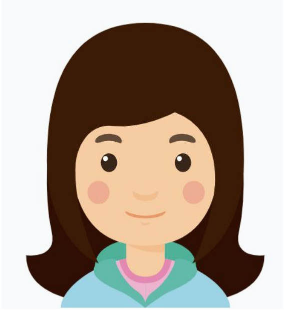
Given a source speech (User: I am your father), 
and target speaker's identity (Darth Vader), 
the converted speech should sound like B uttering A's content.



An example: Voice Conversion

Needs a lot of parallel data for supervised training: (, ), (, ), ...
What about *self-supervised learning*?

User



"I am your father."

"I am your father."

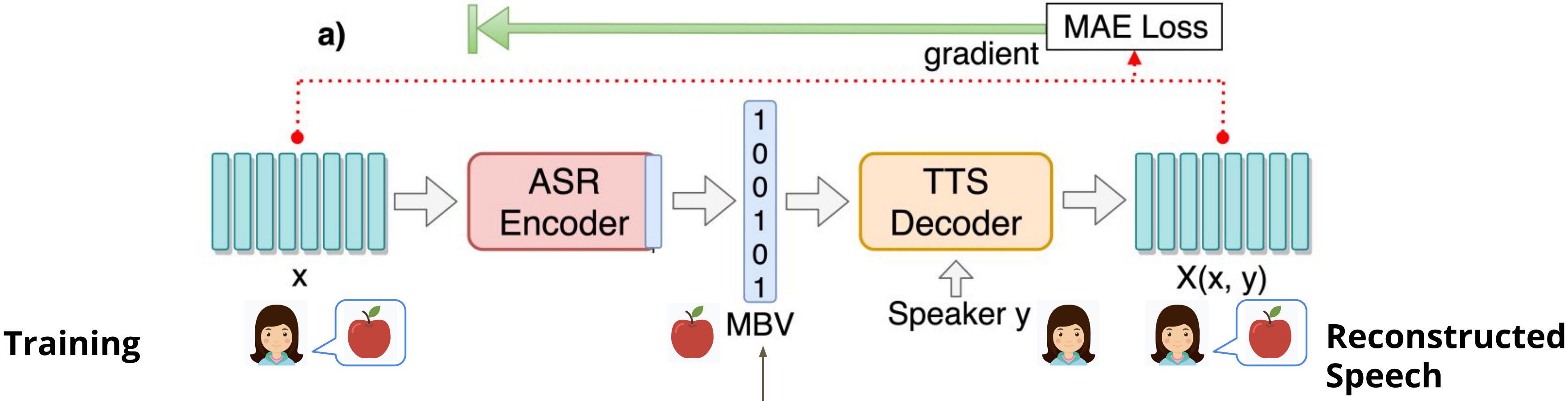
VC Model

Darth Vader

Target

"I am your father."

Voice Conversion (1/3) - Discrete linguistic units discovery



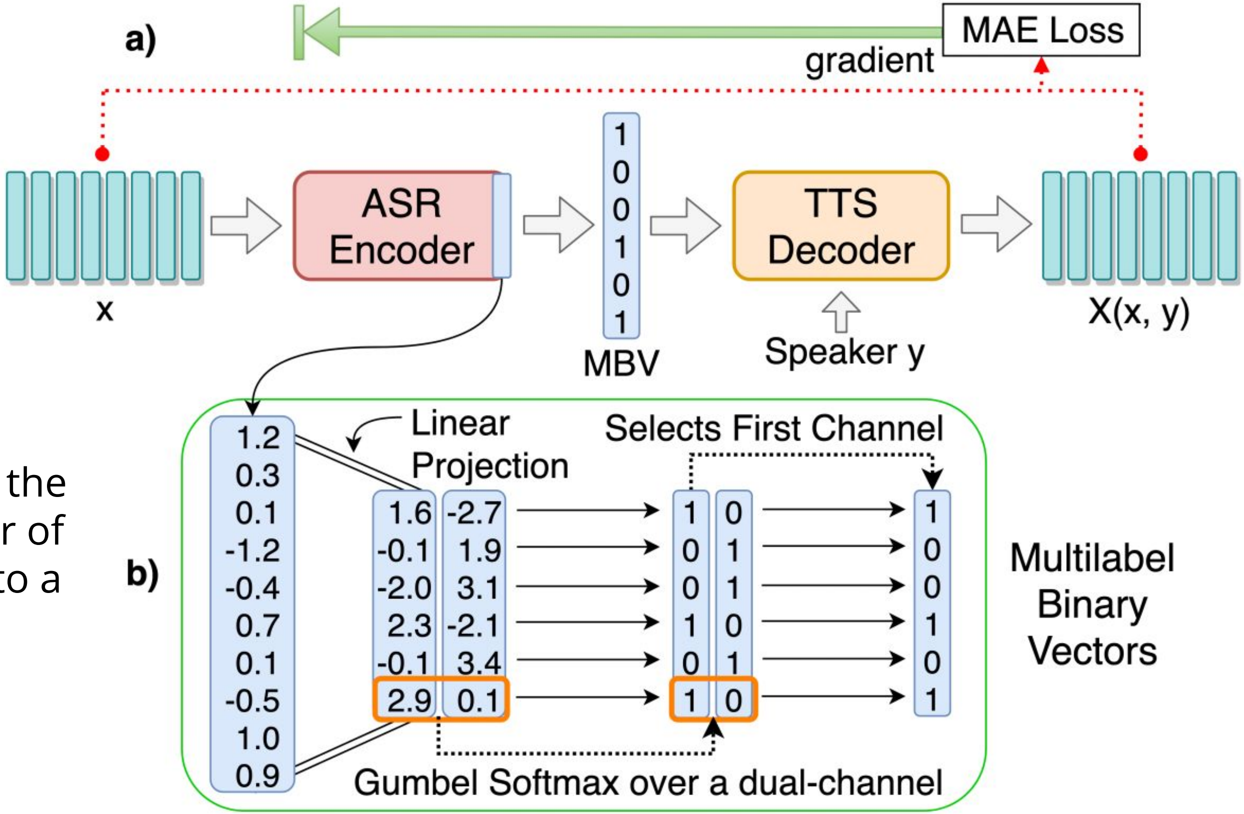
ASR-Encoder is trained to discover a set of common linguistic units given a variety of speakers

TTS-Decoder is trained to project the discovered units back to the designated speech

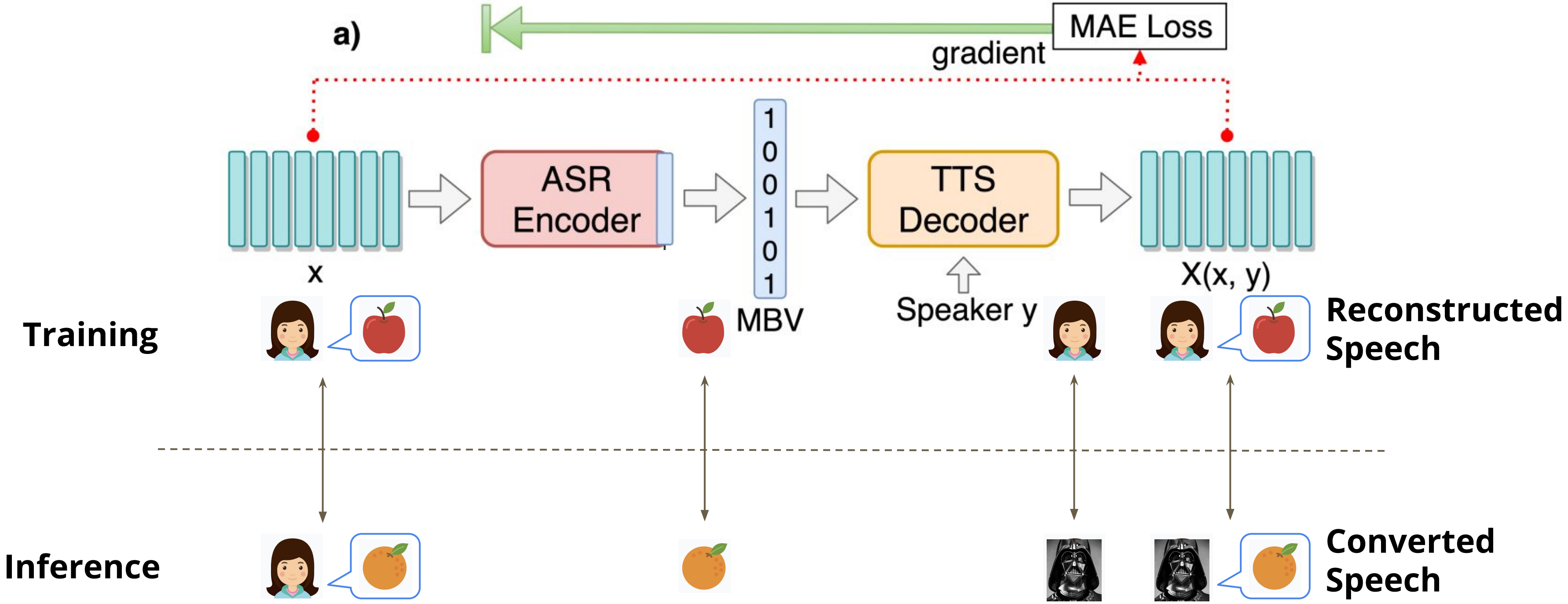
In this **self-supervised end-to-end** manner, **discrete linguistic units** are learned and **represented as multilabel binary vectors (MBVs)**.

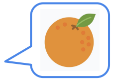
Voice Conversion (2/3) - MBV: vectors of zeros one ones

Linearly project the last hidden layer of ASR-Encoder into a $\mathbb{R}^{n\times 2}$ space.

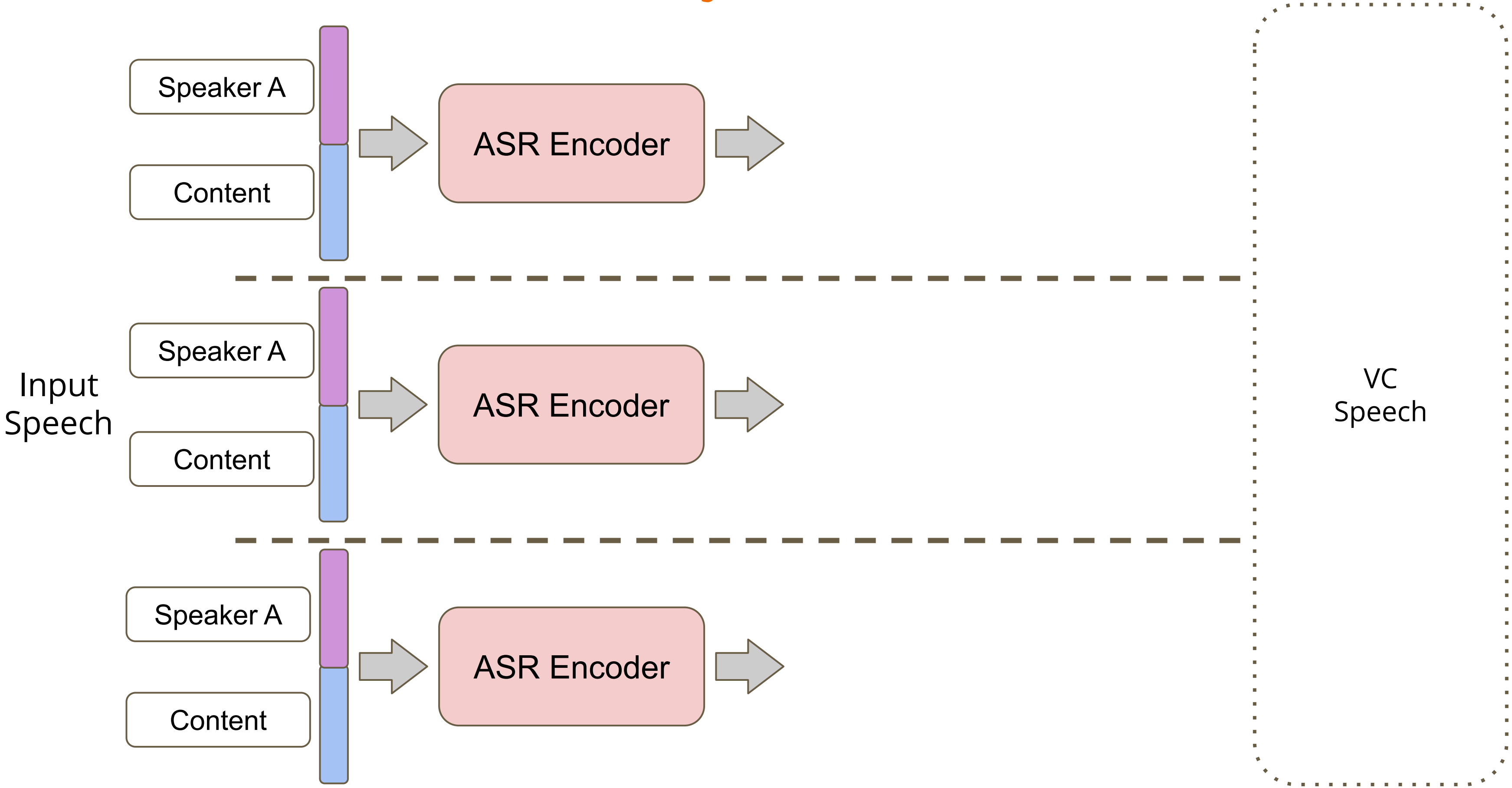


Voice Conversion (3/3) - VC using the ASR-TTS autoencoder

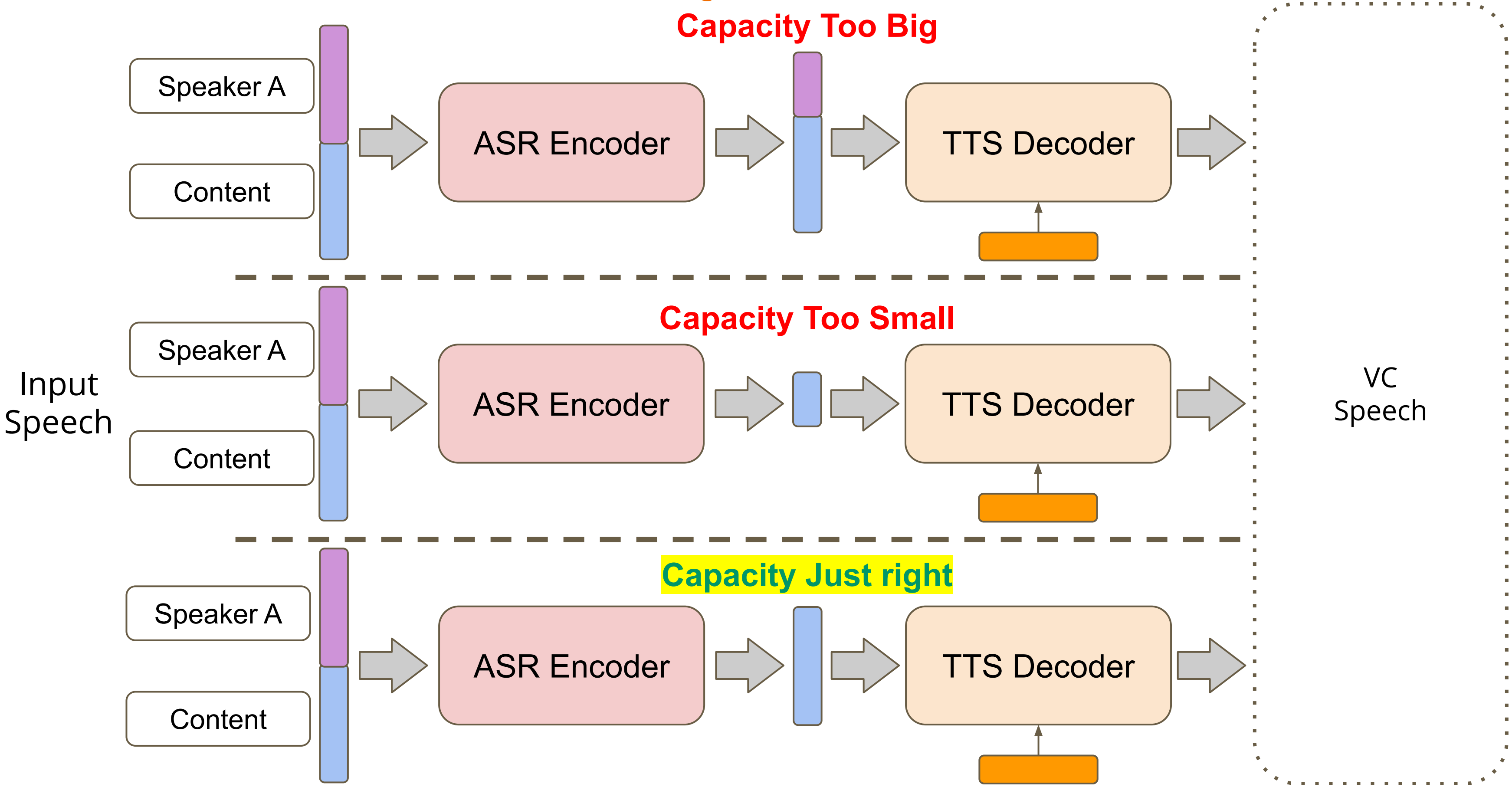


The voice converted speech would sound like  uttering  's content. 

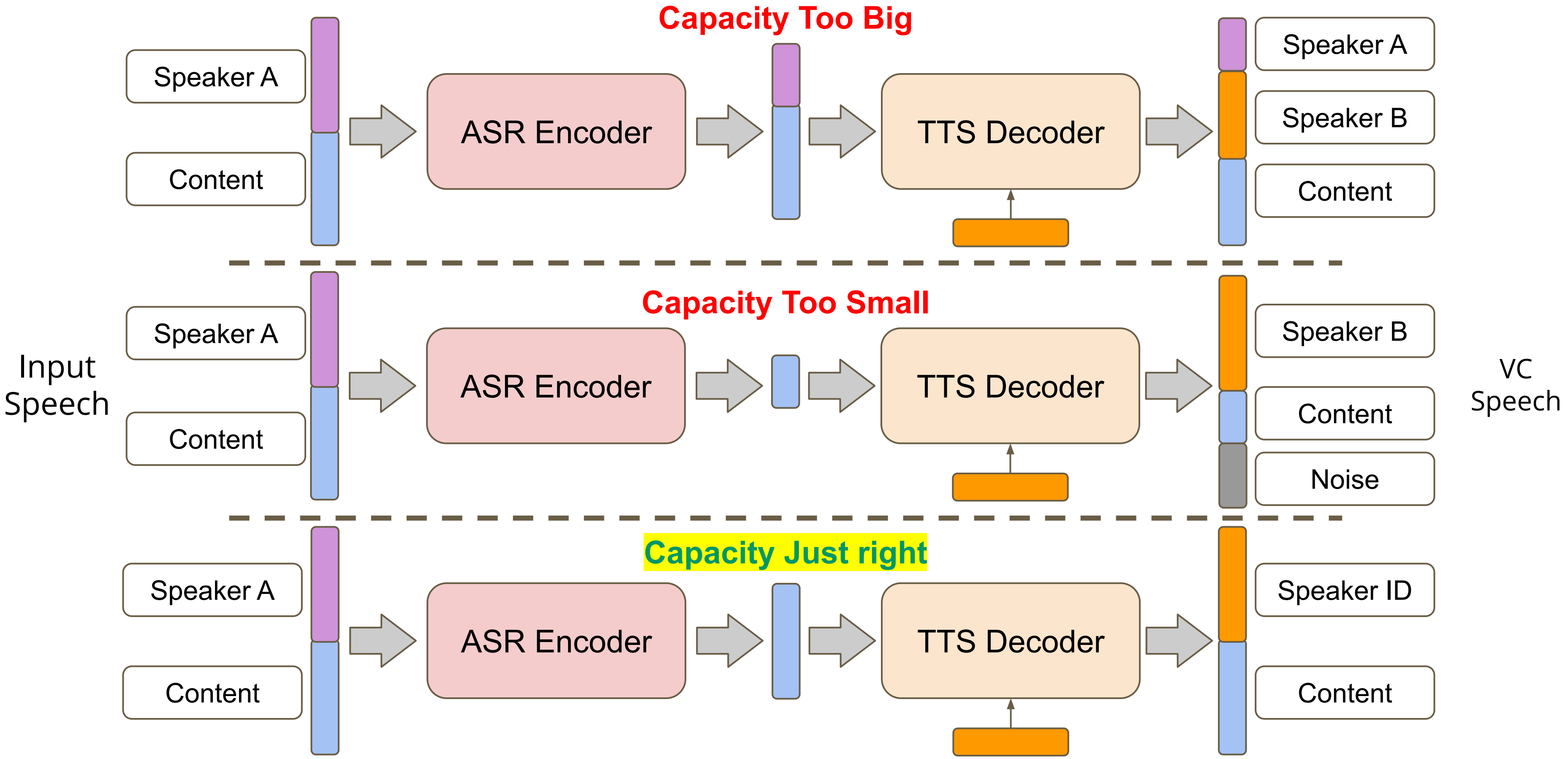
Voice Conversion - Why does it work?



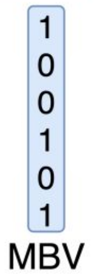
Voice Conversion - Why does it work?



Voice Conversion - Why does it work?



Voice Conversion: ZeroSpeech 2019

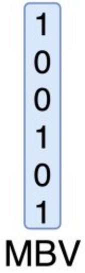


https://zerospeech.com/2019/results.html

- Global Competition
- How good are the learned vector?

Voice Conversion: ZeroSpeech 2019

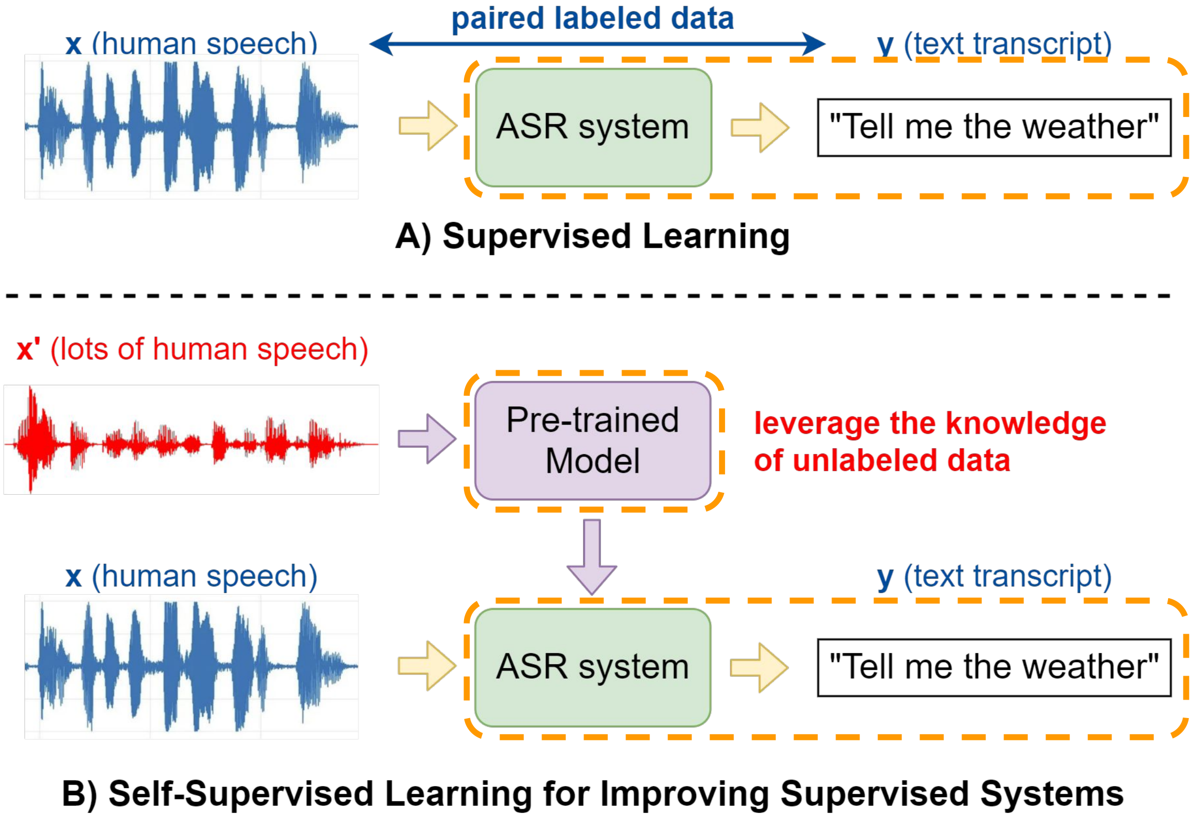
https://zerospeech.com/2019/results.html



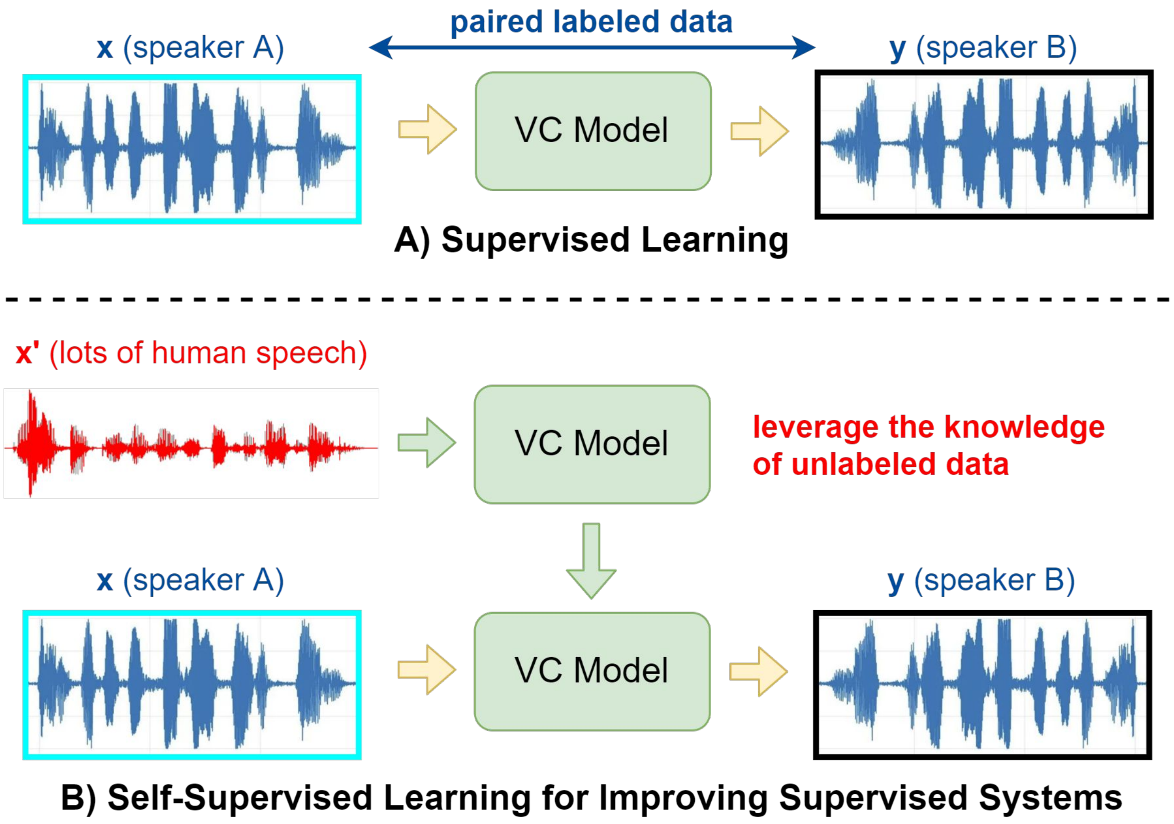
- Global Competition
- How good are the learned vector?
- We achieved 2nd place in terms of learned vectors, while achieving better VC quality than the 1st place.
- Published in InterSpeech 2019 as first author (oral presentation).

#	Authors	Surprise language				
		MOS	CER	Similarity	ABX	Bitrate
17	Gok *et al.*	1.46	0.86	3.03	27.26	29.46
21	**Topline**	3.92	0.28	3.95	16.09	35.2
6	Liu *et al.*	1.69	0.81	1.97	44.25	43.95
18	Liu *et al.*	1.27	0.86	1.96	43.42	43.95
9	Kumar *et al.*	1.44	0.89	3.02	45.64	44.07
10	Kumar *et al.*	1.82	0.86	3.3	40.17	46.07
12	Kamper *et al.*	1.94	0.58	1.95	26.49	69.22
15	Rallabandi *et al.*	1.89	0.71	3.02	28.41	71.42
1	**Baseline**	2.07	0.62	3.41	27.46	74.55
3	Pandia *et al.*	2.02	0.48	3.21	20.77	94.15
2	Pandia *et al.*	2.53	0.43	3.58	23.56	115.43
16	Yusuf *et al.*	1.84	0.8	2.84	24.16	121.03
7	Kamper *et al.*	1.96	0.6	1.76	19.76	139.54
13	Cho *et al.*	1.23	0.85	1.28	12.05	143.76
14	Cho *et al.*	1.53	0.78	1.33	10.39	144.63
19	Tjandra *et al.*	3.25	0.35	2.67	17.8	151.77
5	Feng *et al.*	1.67	0.66	2.6	16.87	299.21
20	Tjandra *et al.*	3.2	0.21	2.3	13.98	362.99
11	Feng *et al.*	1.28	0.74	2.01	10.64	470.23
4	Horizon Robotics	2.89	0.36	1.43	24.92	842.46
8	Horizon Robotics	3.55	0.32	1.34	24.92	842.46

Recall: Self-Supervised Learning for Speech



Self-Supervised Learning: VC



https://rajpurkar.github.io/SQuAD-explorer/

An example: Machine QA

SQuAD2.0

The Stanford Question Answering Dataset

Machine reading comprehension (MRC) is an AI challenge that requires machine to determine the correct answers to questions based on a given passage.

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
4 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.115	92.580
5 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
6 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
6 Feb 25, 2020	Albert_Verifier_AA_Net (ensemble) *QIANXIN*	89.743	92.180

https://rajpurkar.github.io/SQuAD-explorer/

An example: Machine QA

SQuAD2.0

The Stanford Question Answering Dataset

Machine reading comprehension (MRC) is an AI challenge that requires machine to determine the correct answers to questions based on a given passage.

Humans are outperformed by machine!

ALBERT (BERT)

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

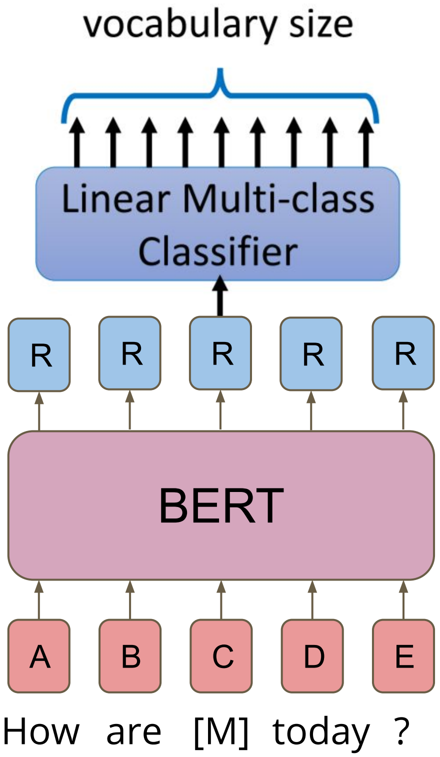
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
4 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.115	92.580
5 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
6 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
6 Feb 25, 2020	Albert_Verifier_AA_Net (ensemble) *QIANXIN*	89.743	92.180

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:

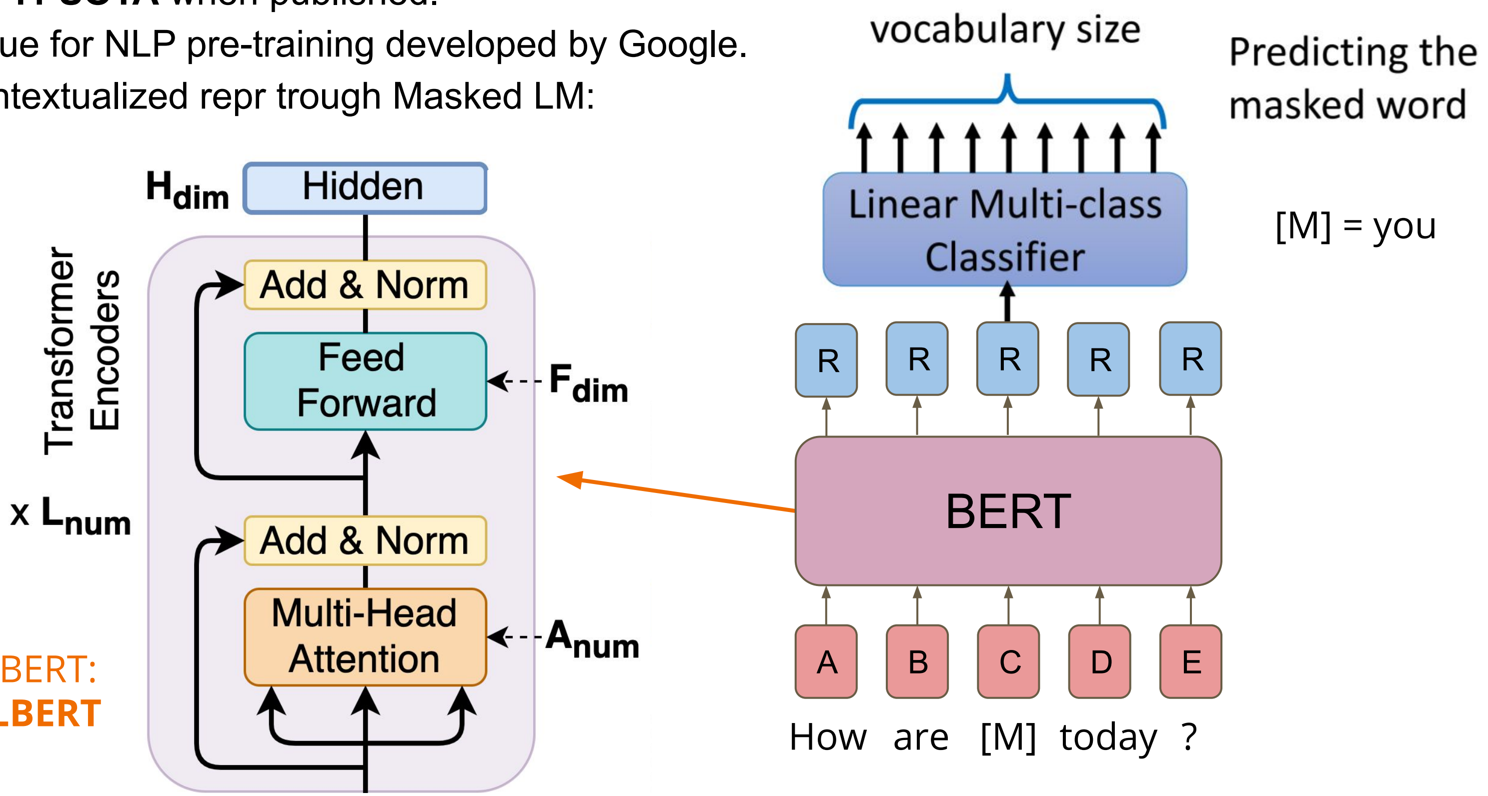


Predicting the masked word

[M] = you

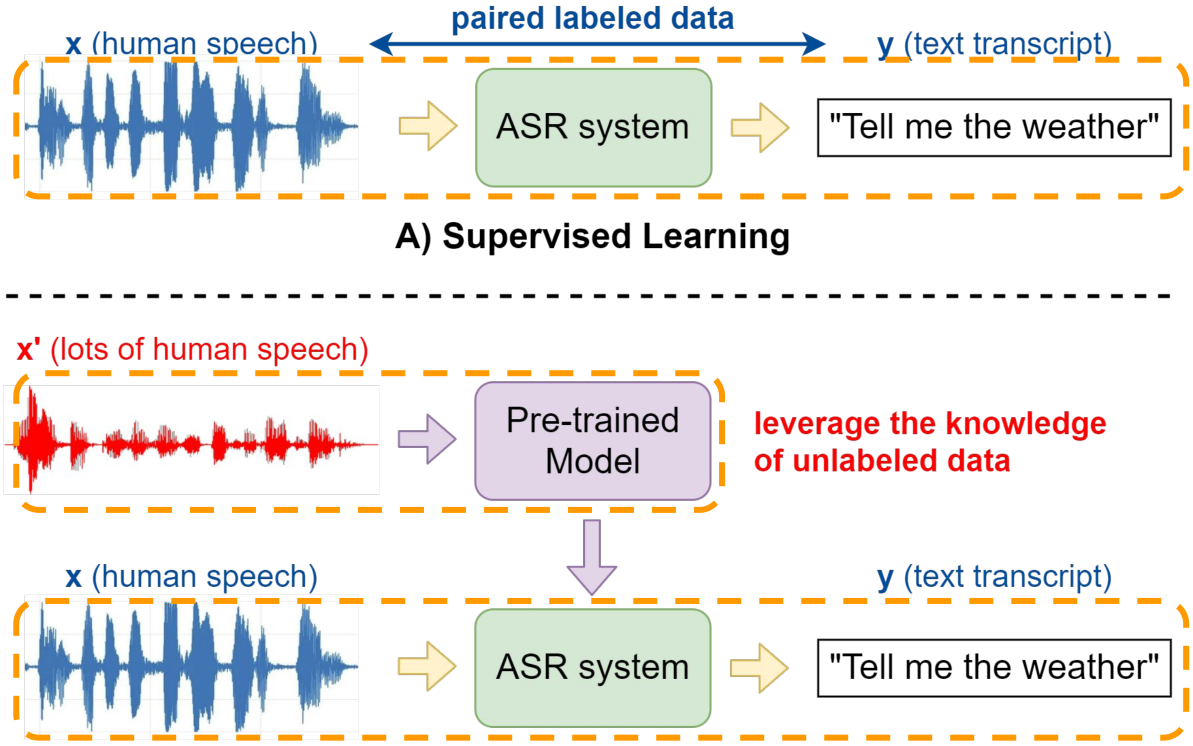
BERT (Bidirectional Encoder Representations from Transformers)

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



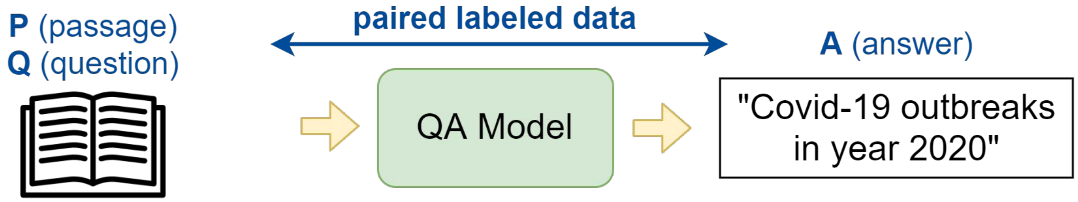
Share layers of BERT:
A Lite BERT = **ALBERT**

Recall: Self-Supervised Learning for Speech

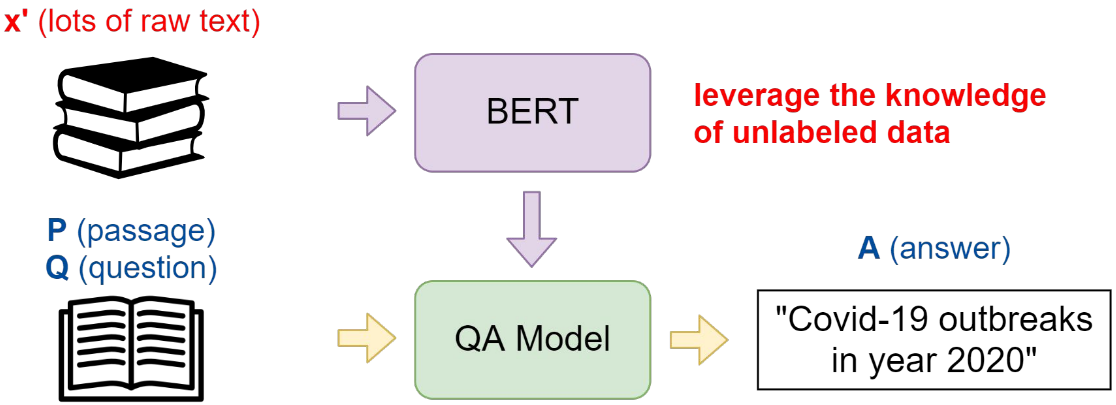


B) Self-Supervised Learning for Improving Supervised Systems

Self-Supervised Learning: BERT



A) Supervised Learning



B) Self-Supervised Learning for Improving Supervised Systems

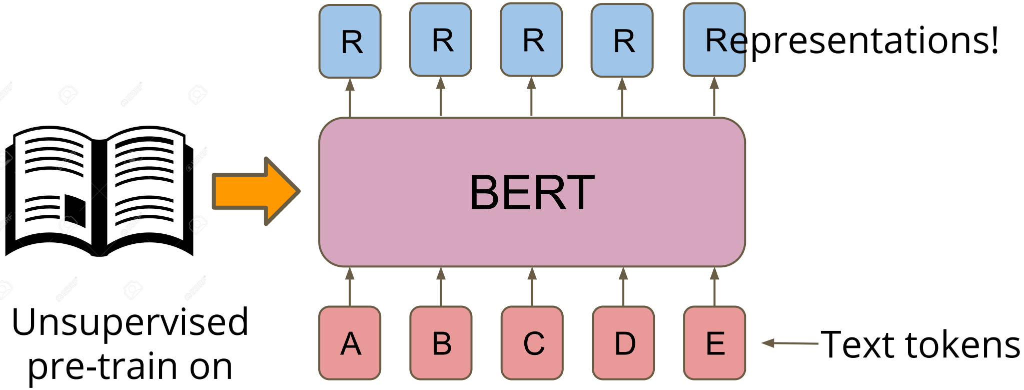
Mockingjay

From BERT to Speech BERT

From BERT to Speech BERT

NLP BERT:

Language Representation Learning



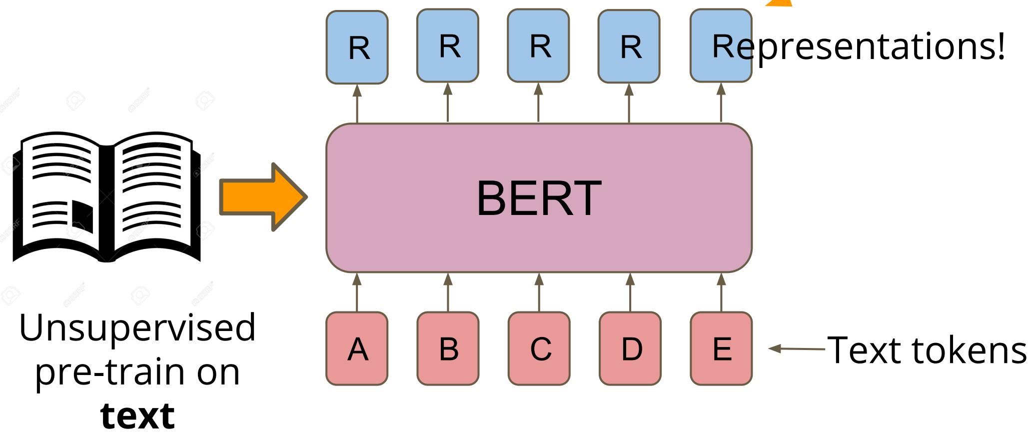
Unsupervised
pre-train on
text

From BERT to Speech BERT

NLP BERT:

Language Representation Learning

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



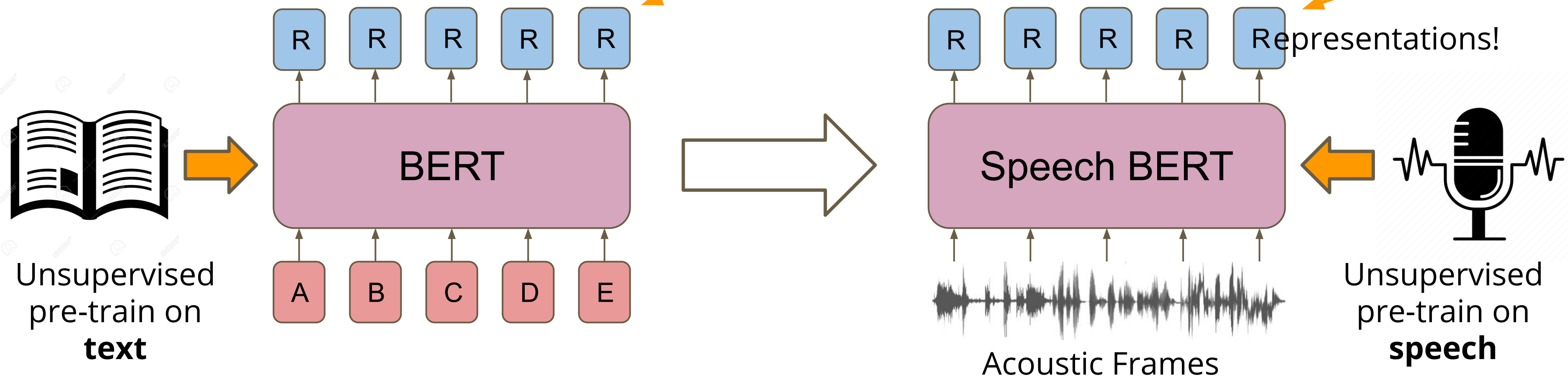
From BERT to Speech BERT

NLP BERT:
Language Representation Learning

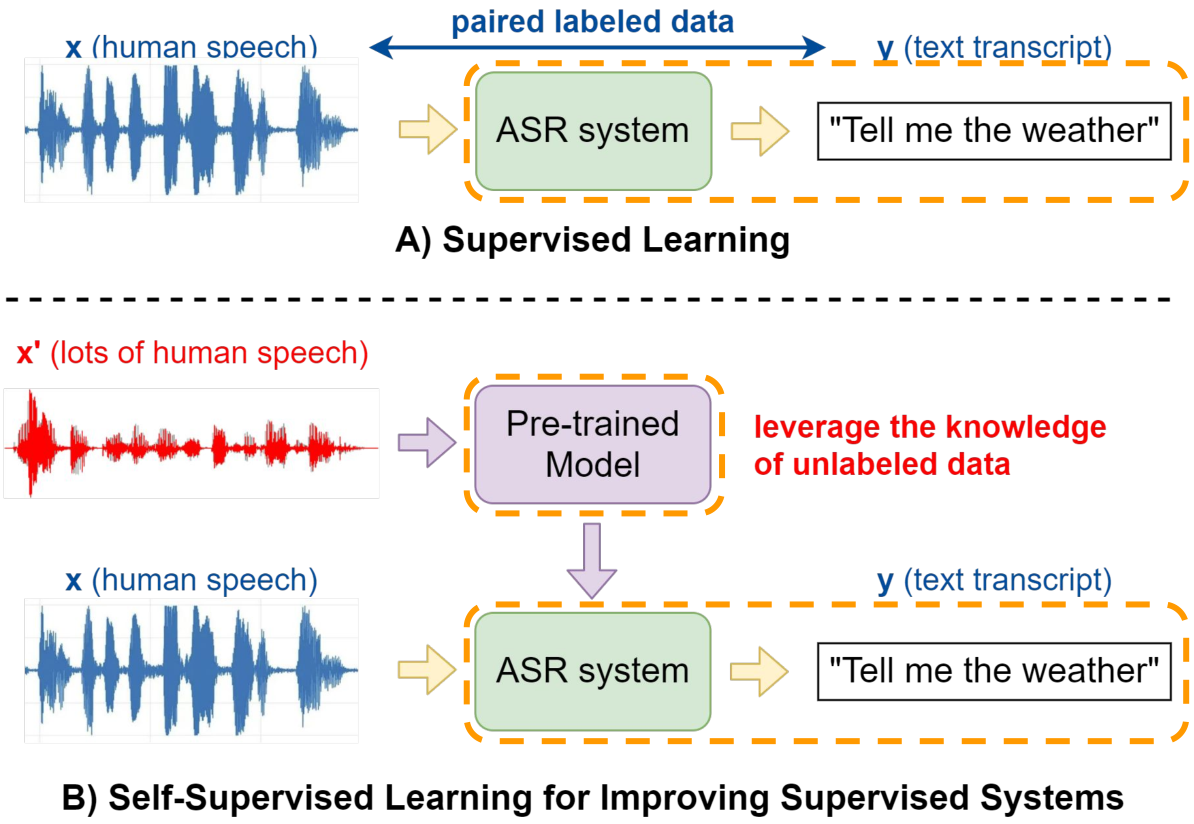
Usage:
Extracts features for downstream **NLP** models (can also be fine-tuned)

Speech BERT:
Speech Representation Learning

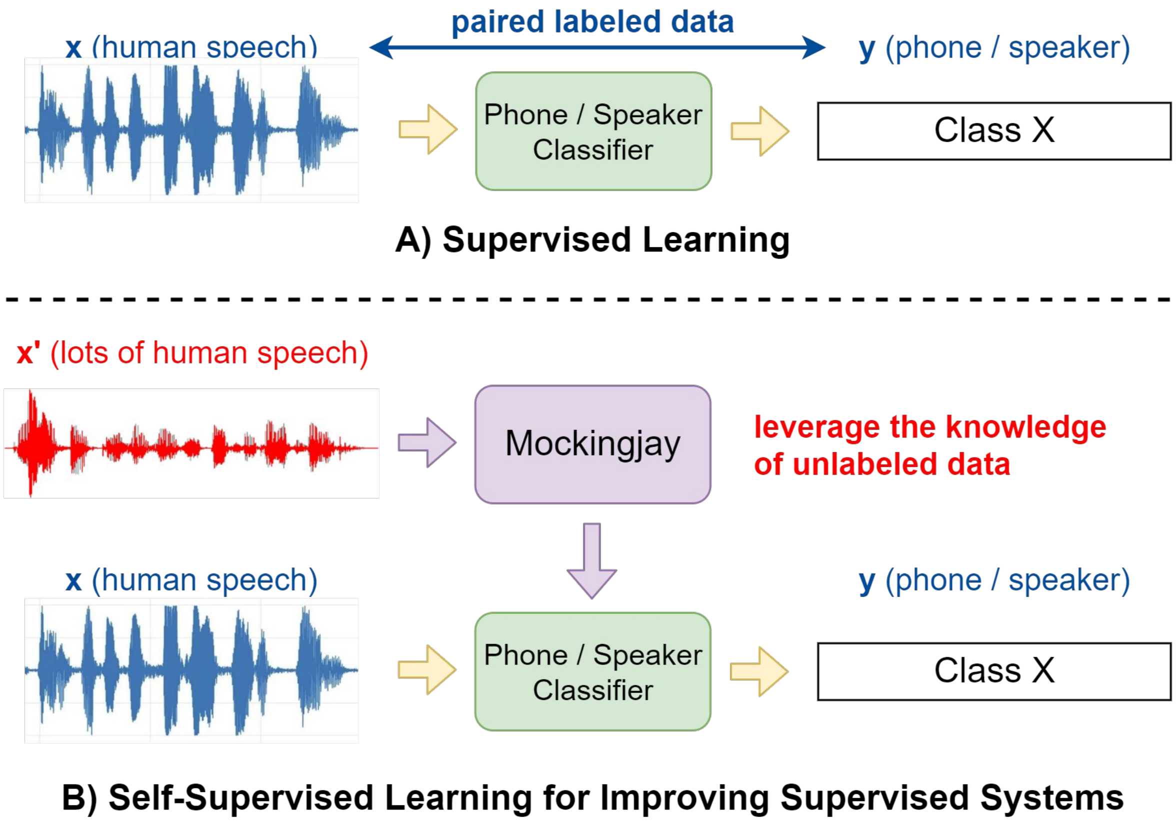
Usage:
Extracts features for downstream **SLP** models (can also be fine-tuned)



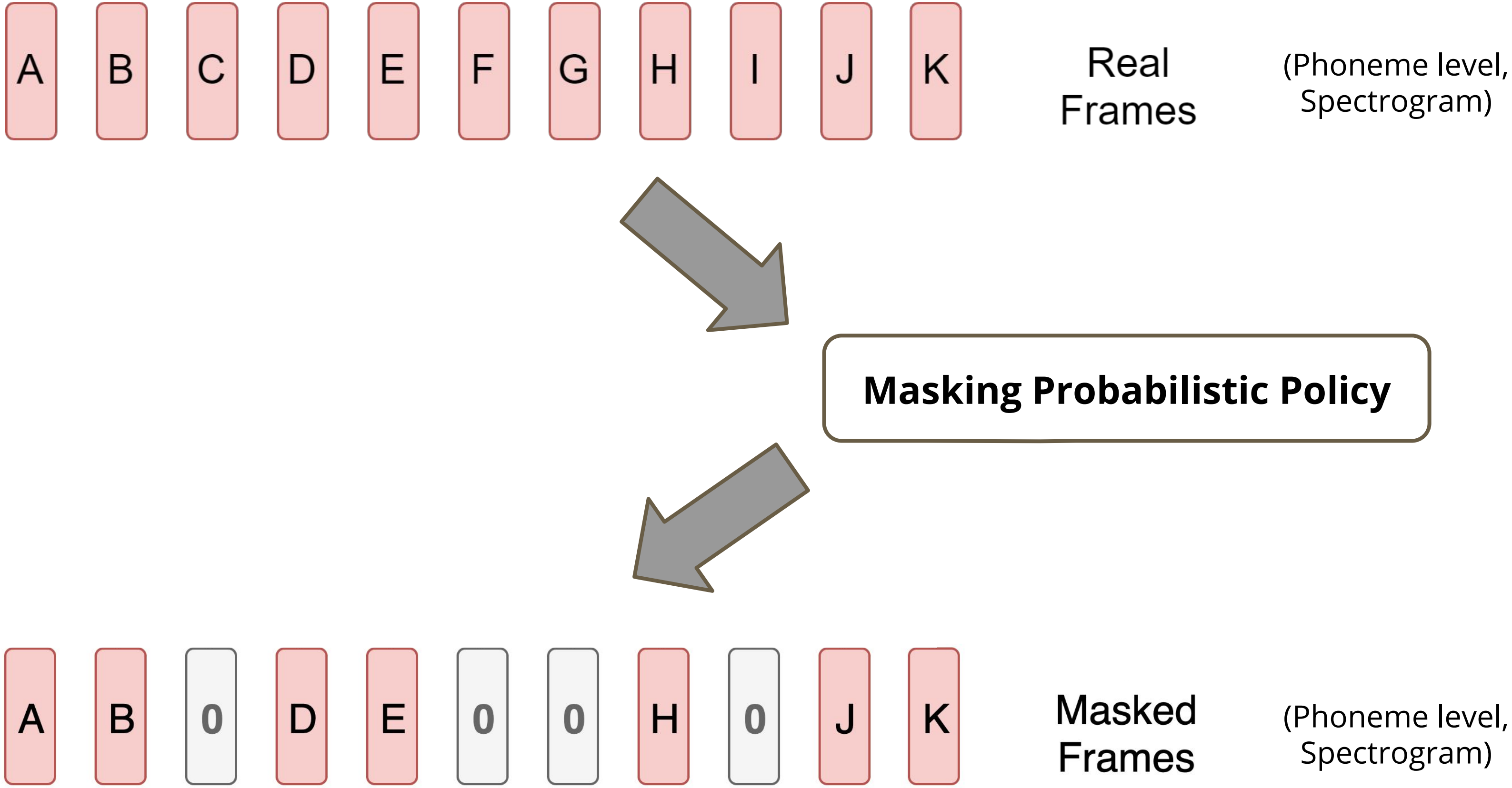
Recall: Self-Supervised Learning for Speech



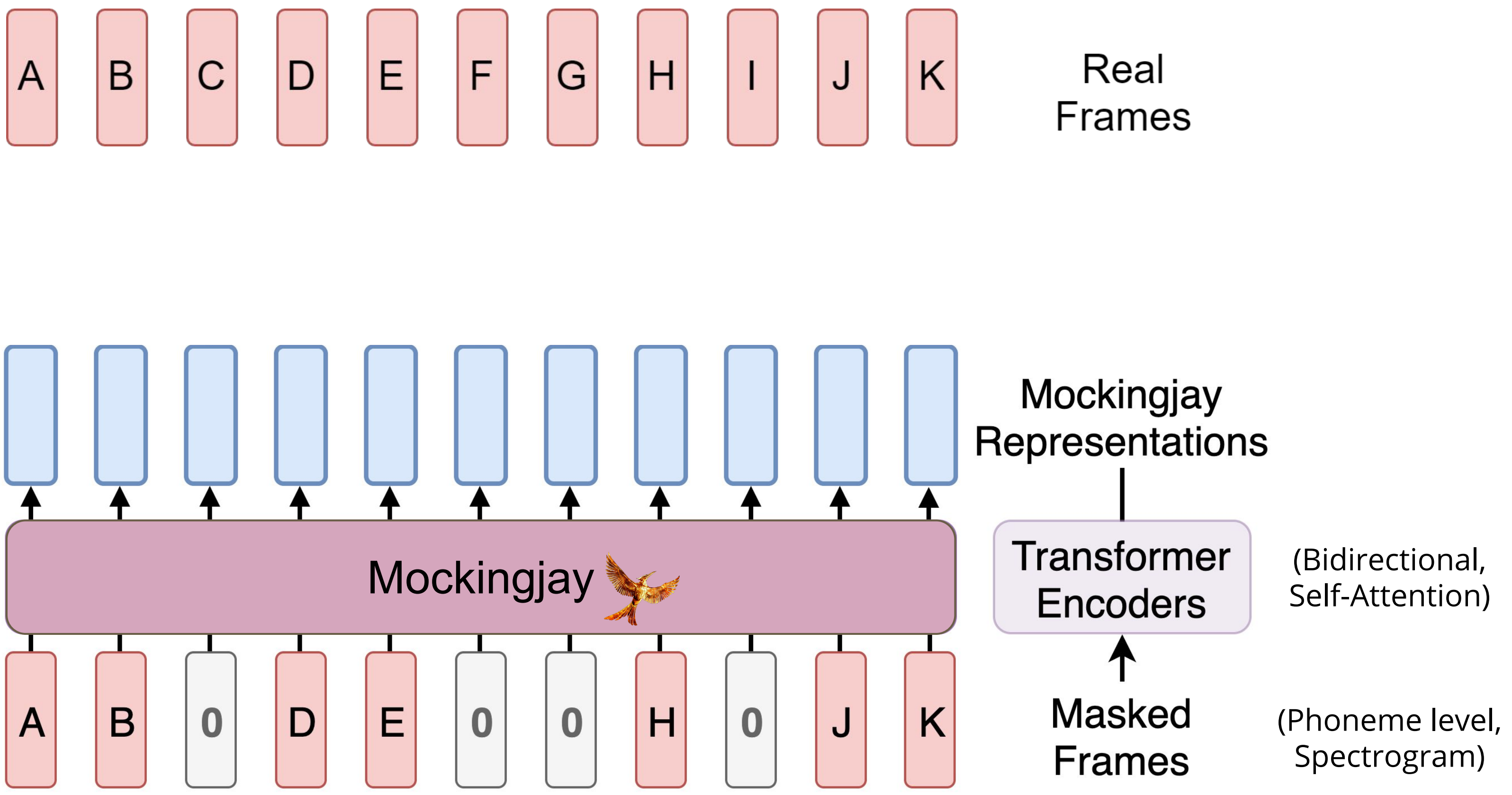
Self-Supervised Learning: Mockingjay



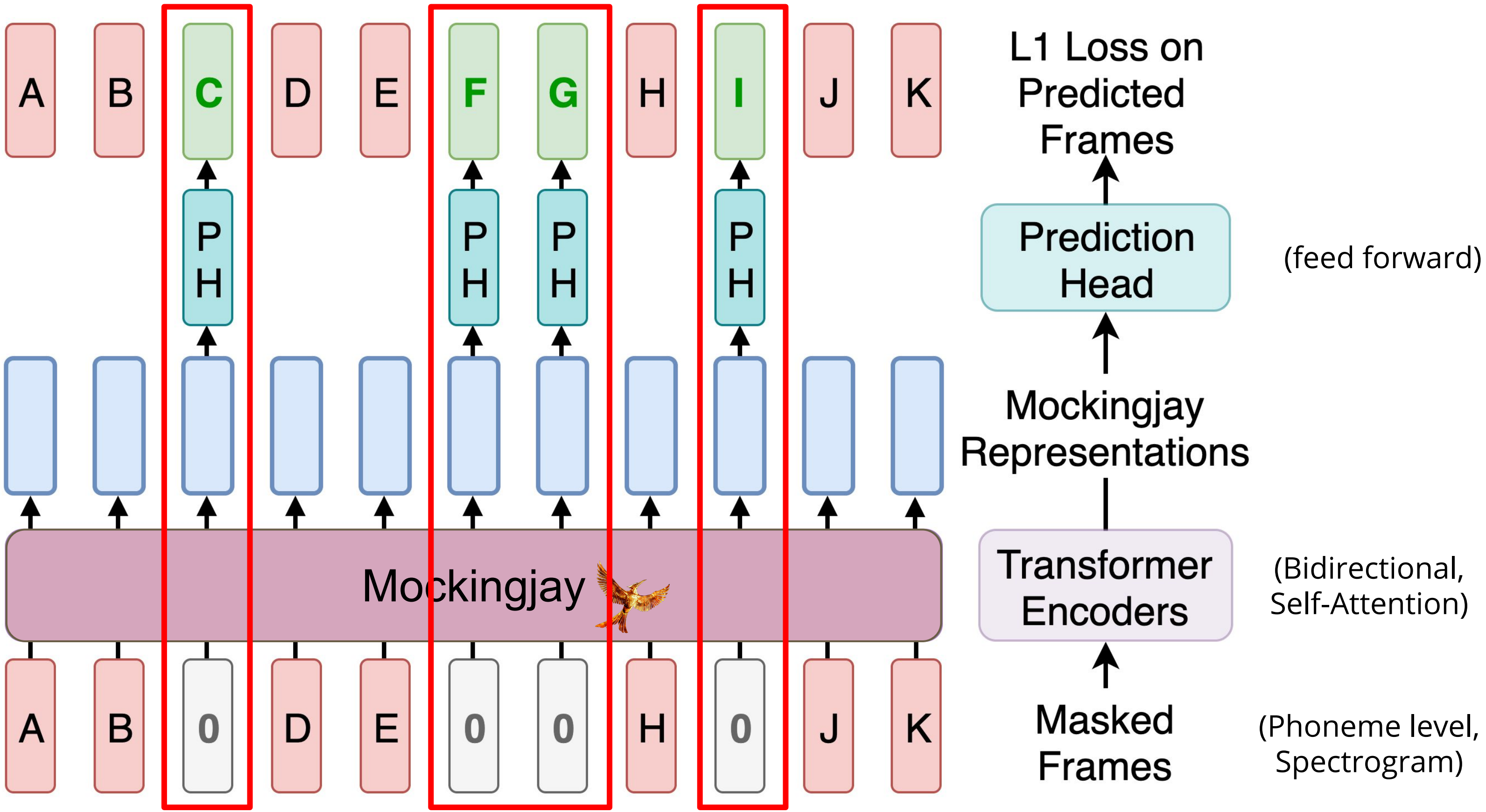
Pre-Training Task: Masked Acoustic Model



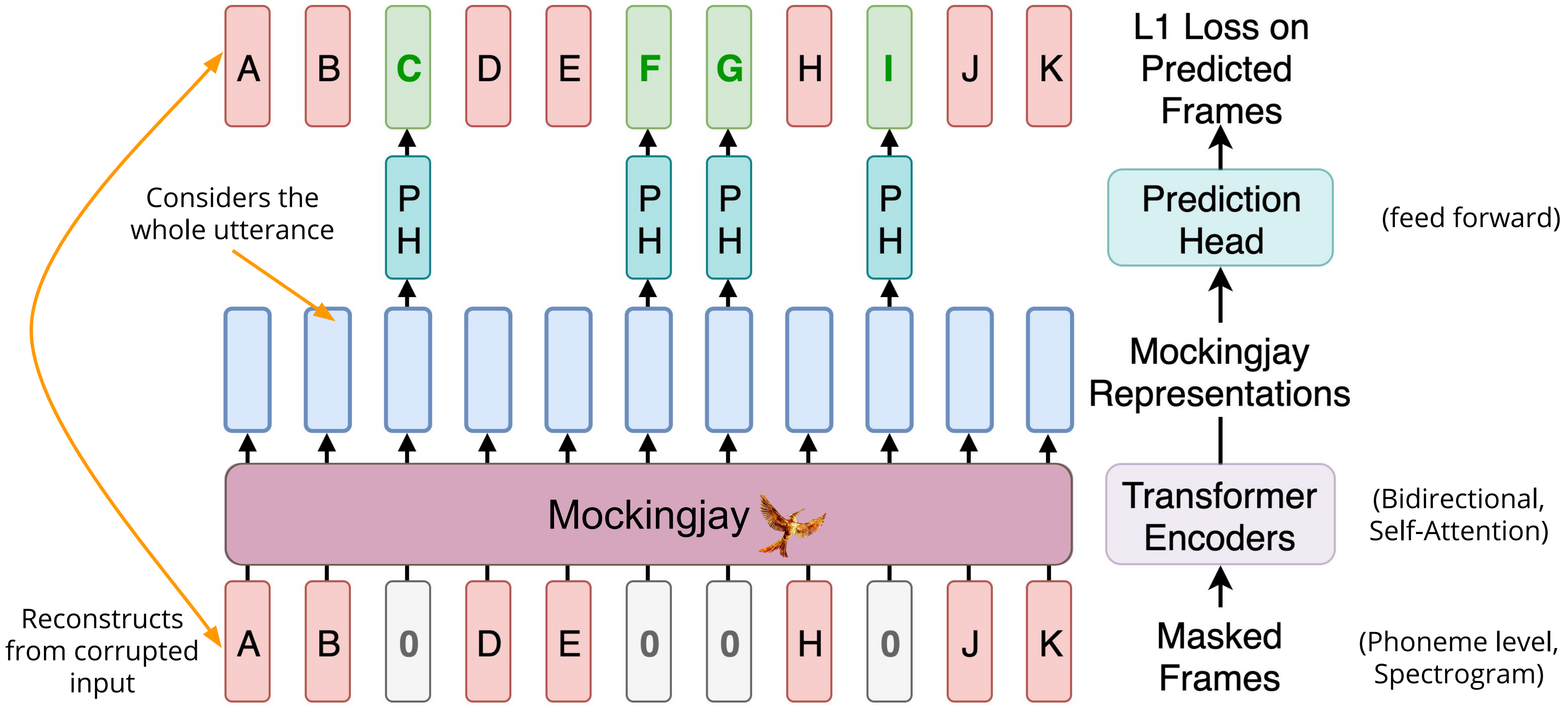
Pre-Training Task: Masked Acoustic Model



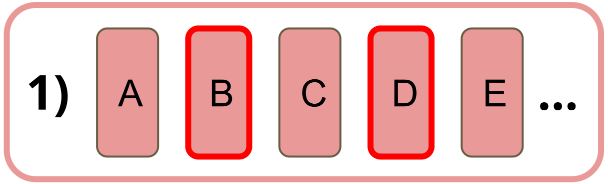
Pre-Training Task: Masked Acoustic Model



Pre-Training Task: Masked Acoustic Model

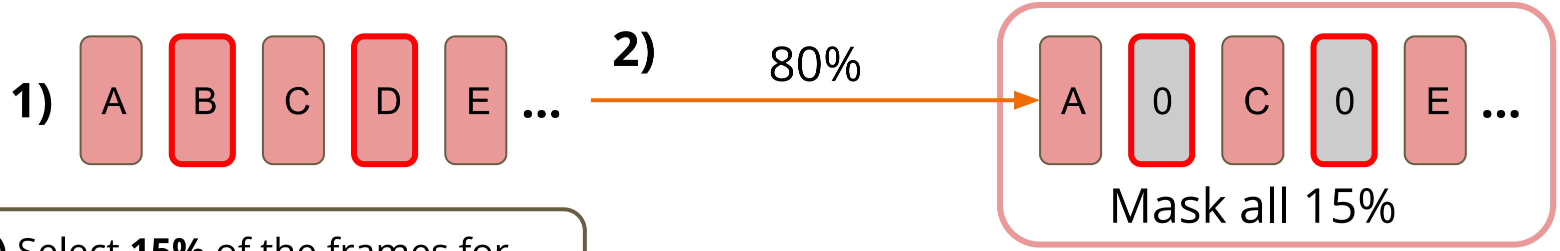


Probabilistic Policy for Masking Frames



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

Probabilistic Policy for Masking Frames

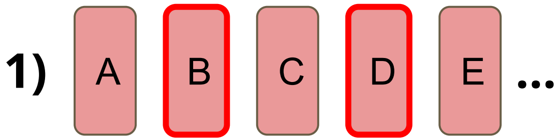


1) Select **15%** of the frames for prediction (highlighted in green).

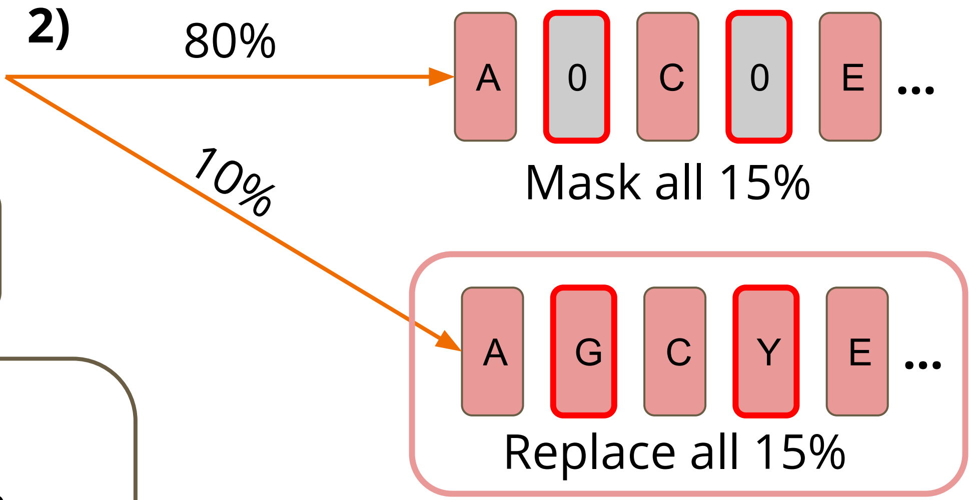
2) For all selected frames:

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

Probabilistic Policy for Masking Frames



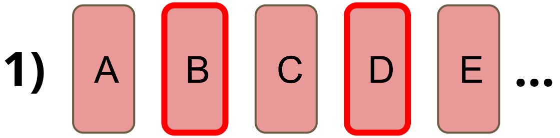
1) Select **15%** of the frames for prediction (highlighted in green).



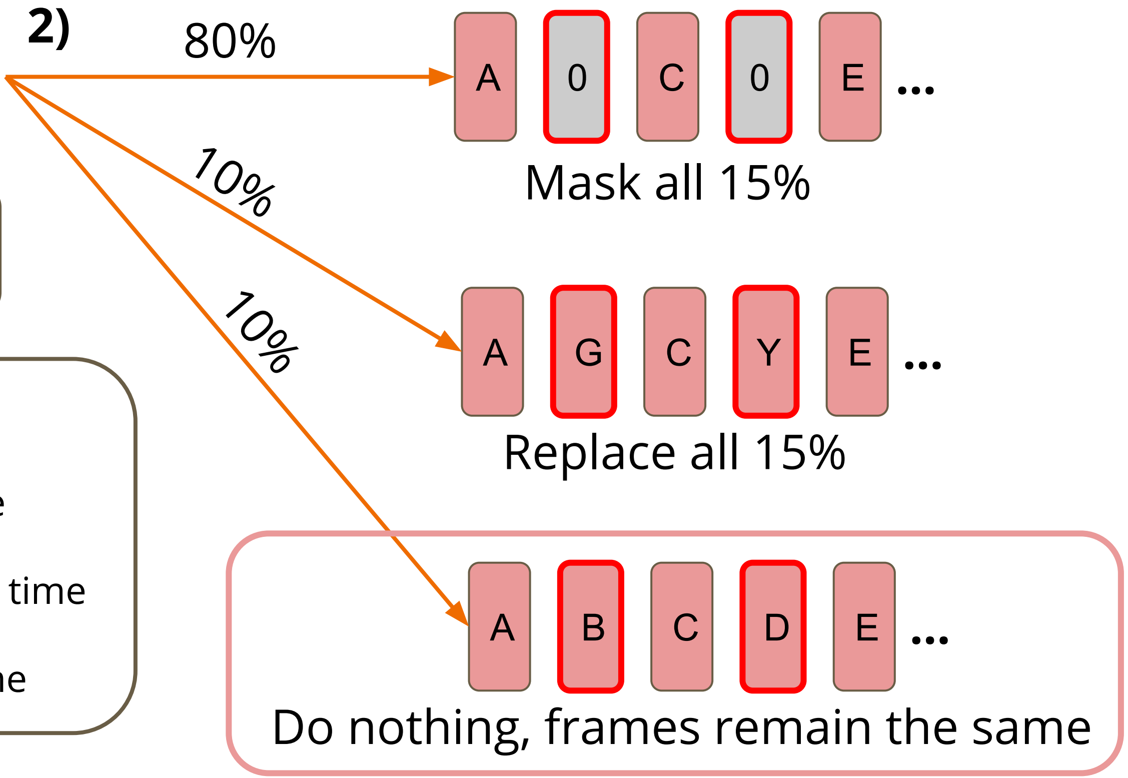
2) For all selected frames:

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

Probabilistic Policy for Masking Frames



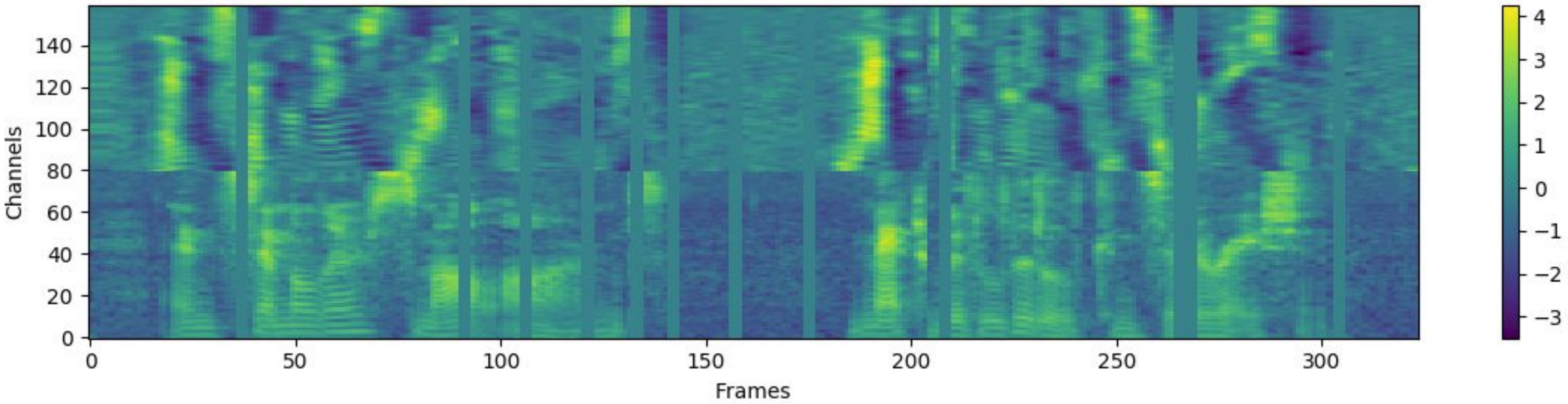
1) Select **15%** of the frames for prediction (highlighted in green).



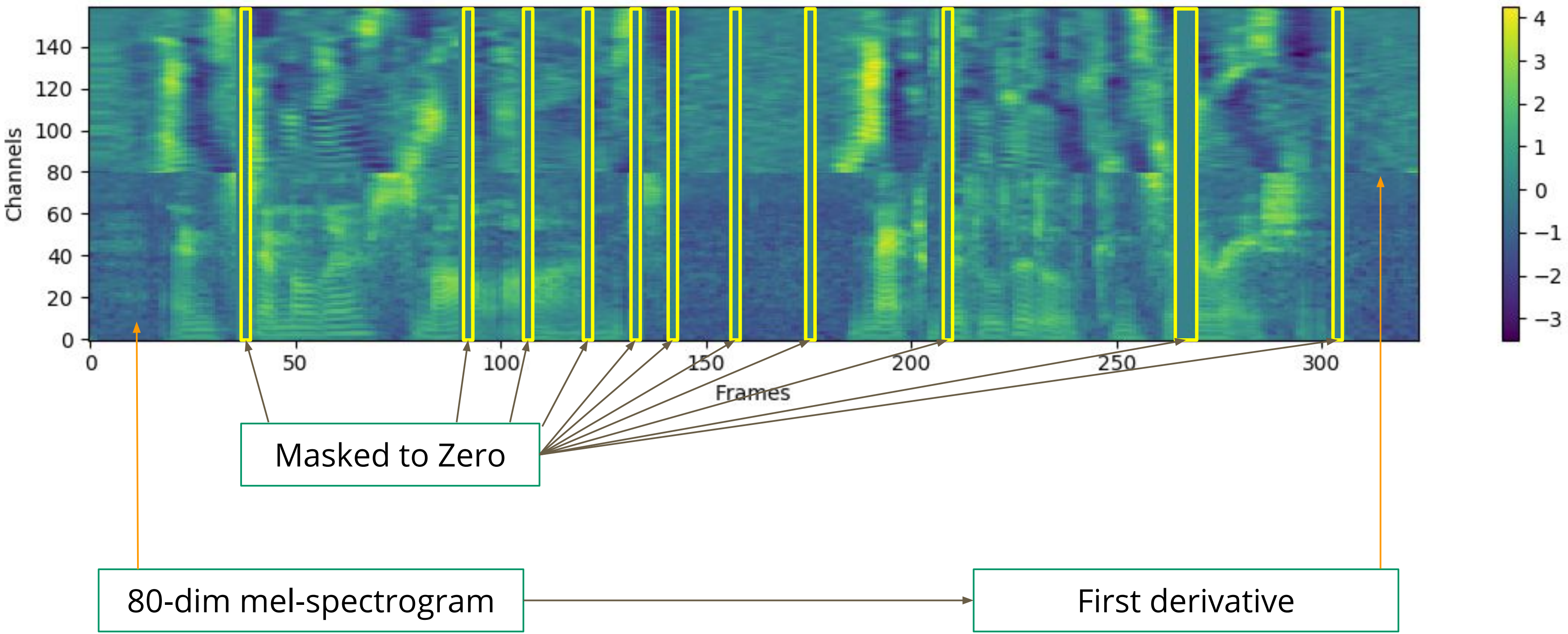
2) For all selected frames:

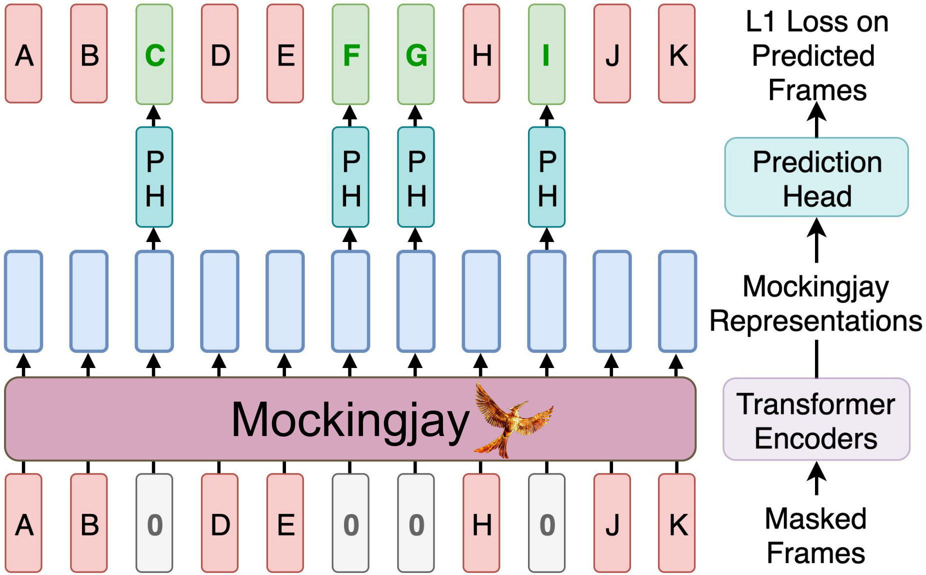
- mask to zero **80%** of the time
- replace randomly **10%** of the time
- leave untouch **10%** of the time

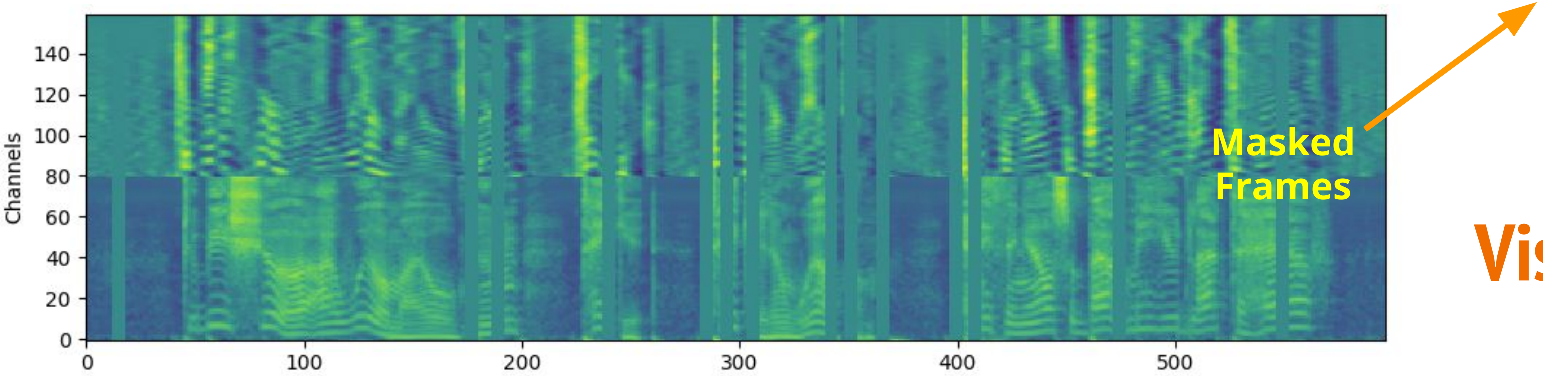
Input Feature: Masked Spectrogram



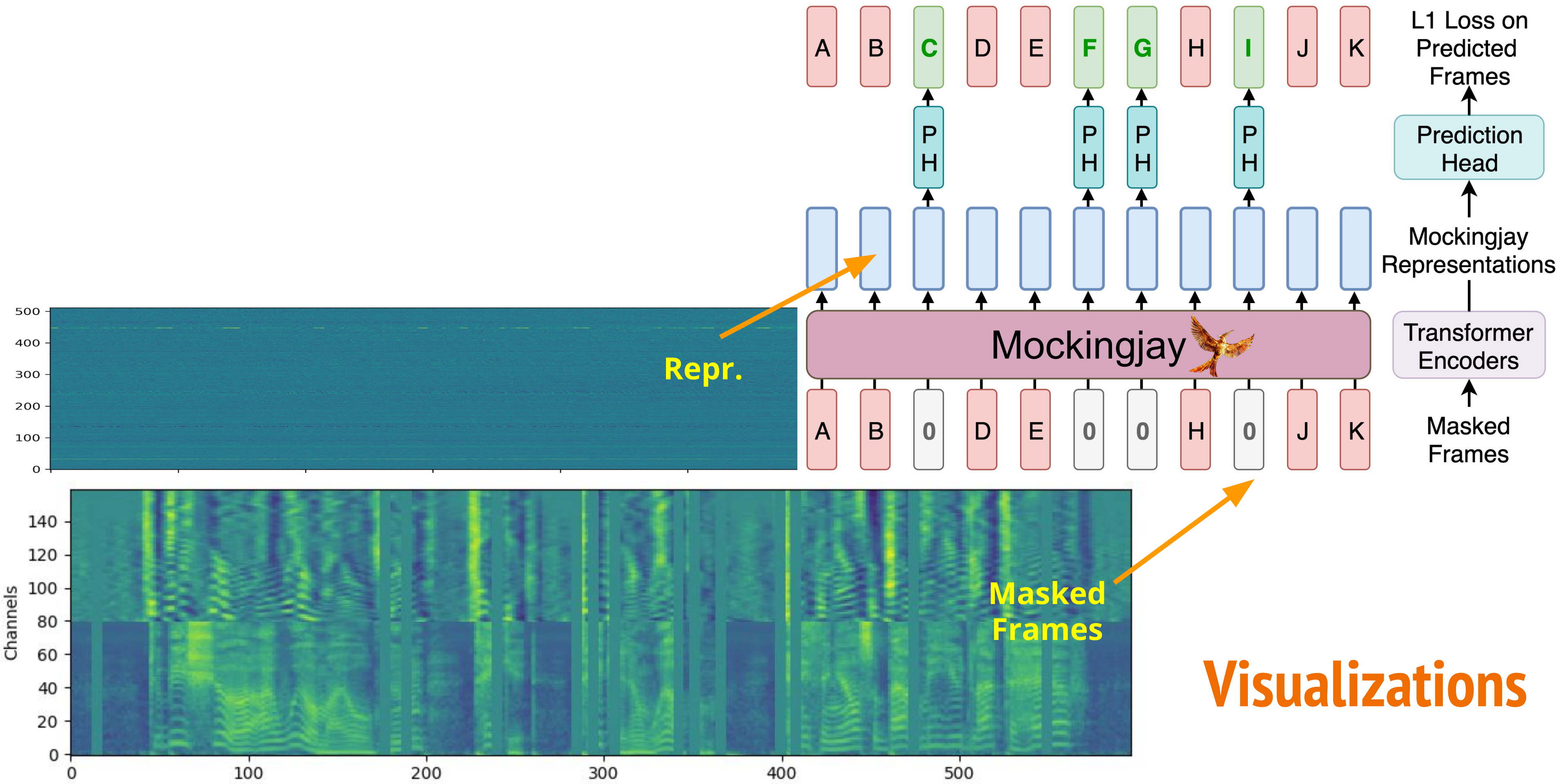
Input Feature: Masked Spectrogram



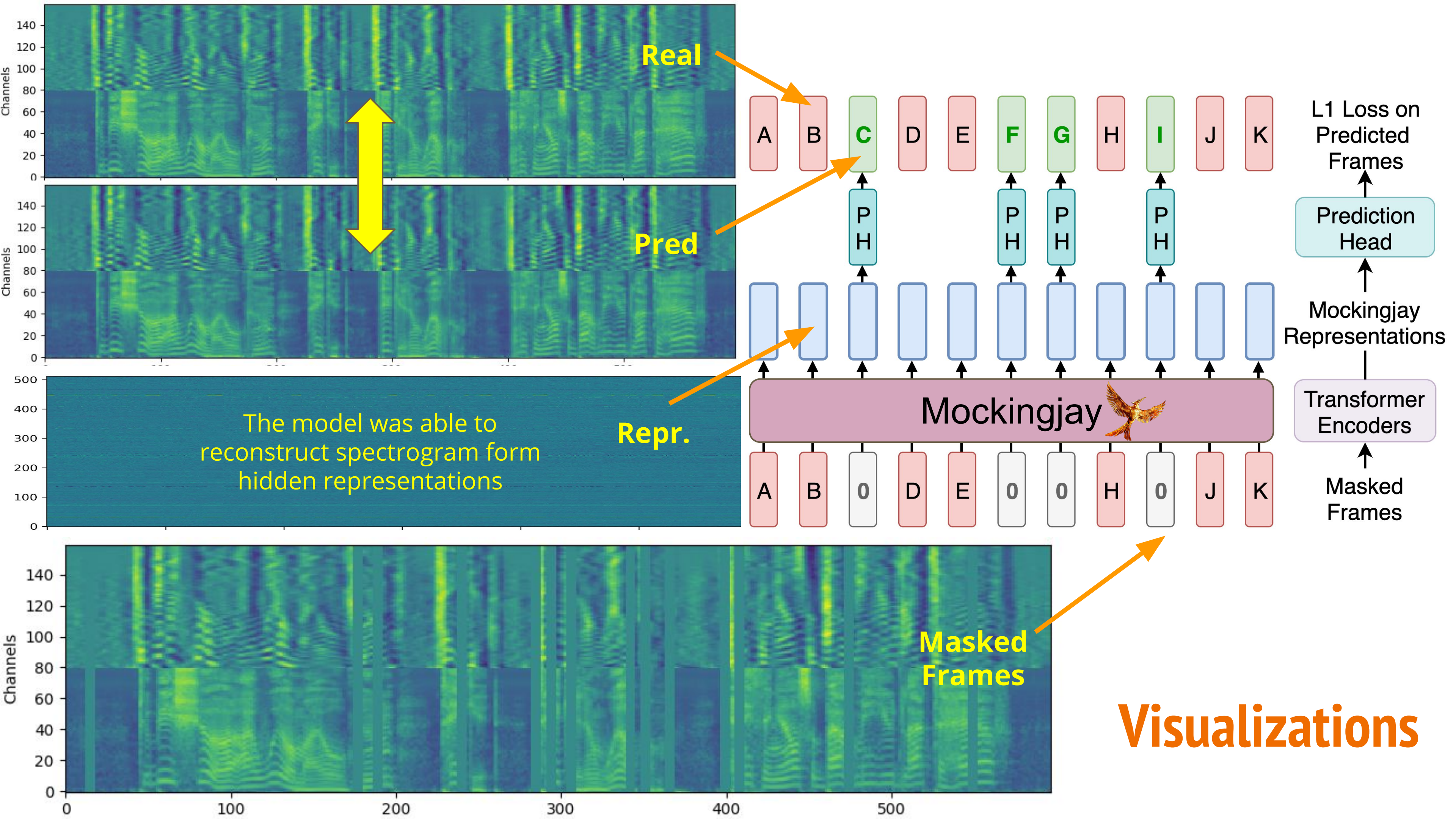




Visualizations



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

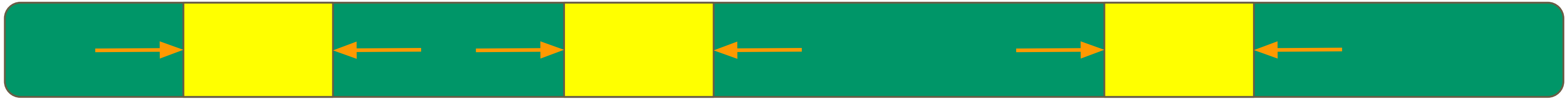


Real
Pred
Repr.
The model was able to reconstruct spectrogram form hidden representations
Masked Frames
A B C D E F G H I J K
P H
L1 Loss on Predicted Frames
Prediction Head
Mockingjay Representations
Mockingjay
Transformer Encoders
A B 0 D E 0 0 H 0 J K
Masked Frames
Channels
Visualizations

Migrating from text to speech

Acoustic Features: long and locally smooth in nature,

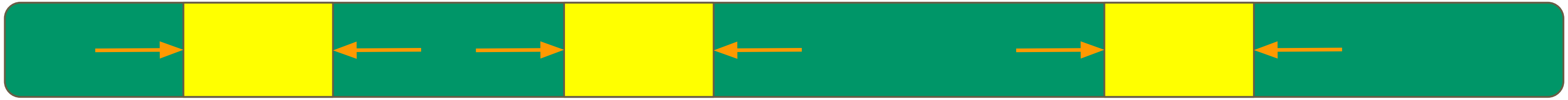
need to 1) shorten the sequence and 2) mask over a longer span



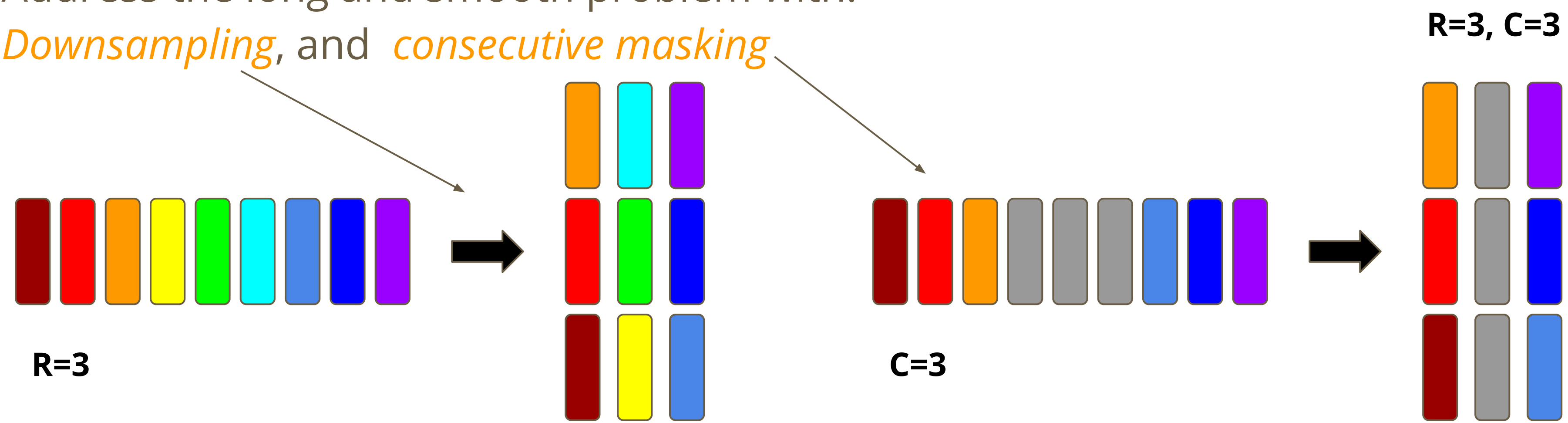
Migrating from text to speech

Acoustic Features: long and locally smooth in nature,

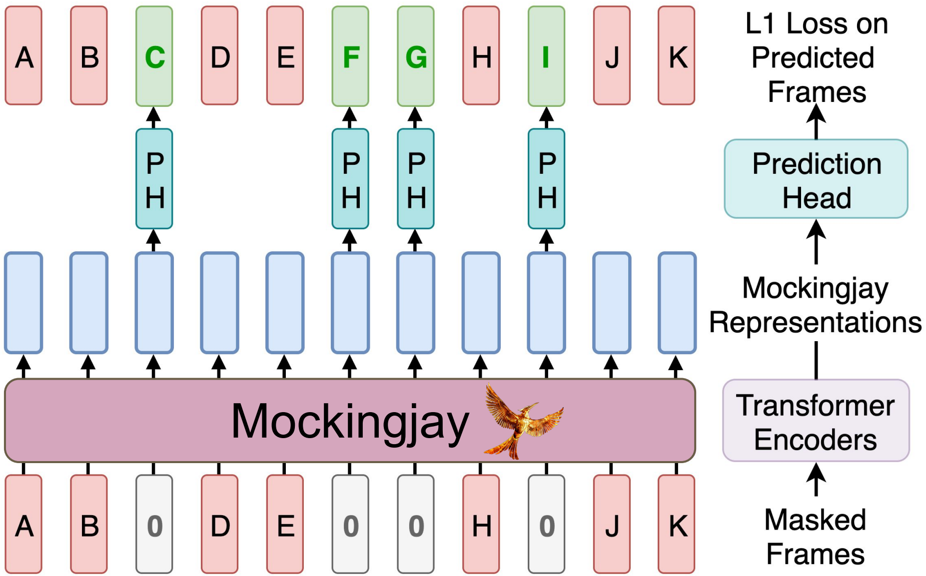
need to <u>1) shorten the sequence</u> and <u>2) mask over a longer span</u>



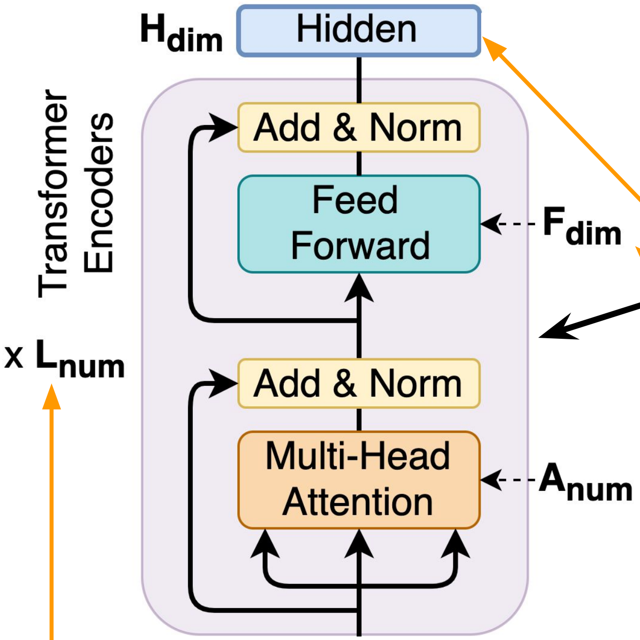
Address the long and smooth problem with:
Downsampling, and *consecutive masking*

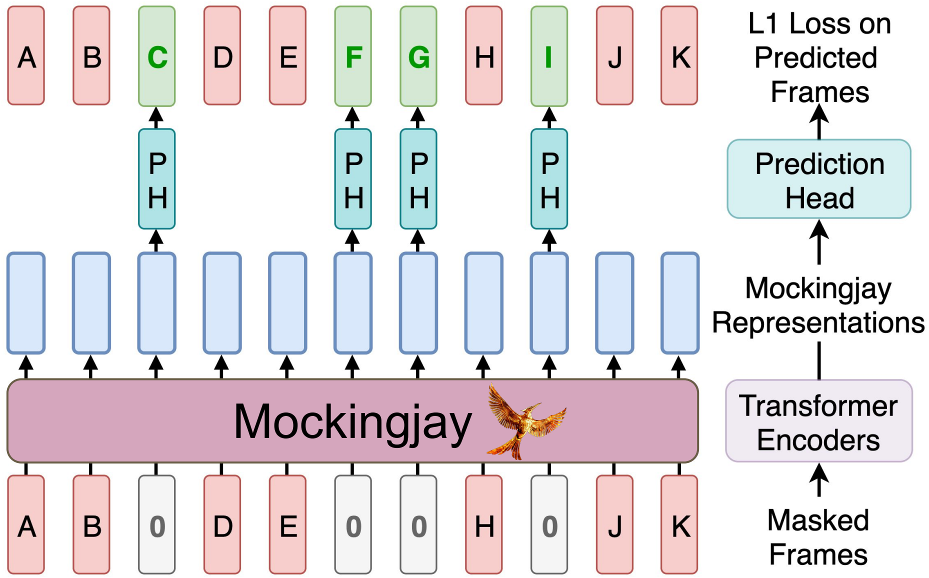


Model Architecture



Model Architecture





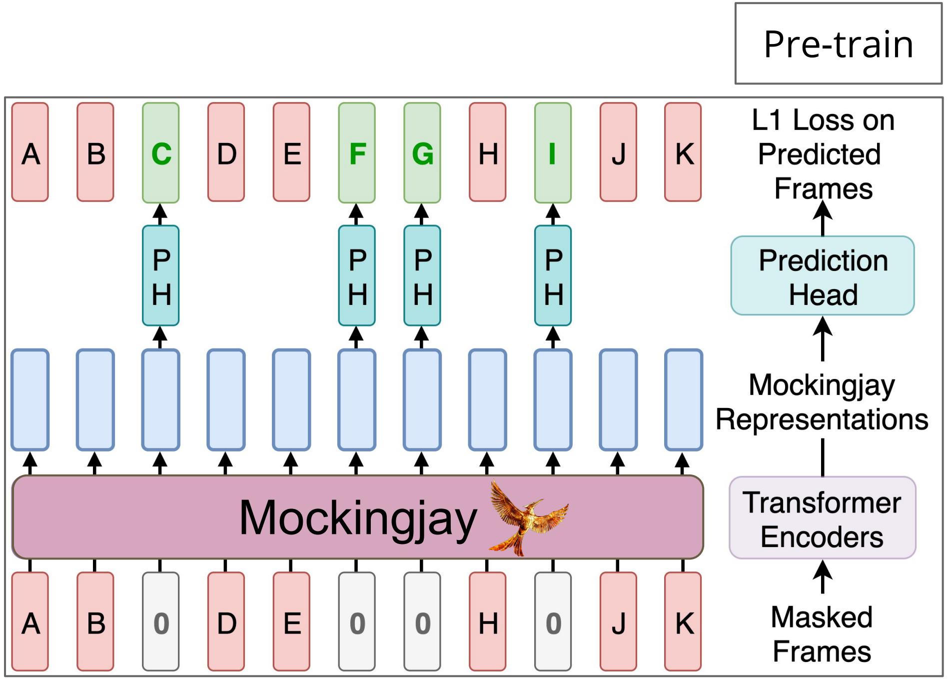
- H_dim = **768**
- F_dim = **3072**
- A_num = **12**

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

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

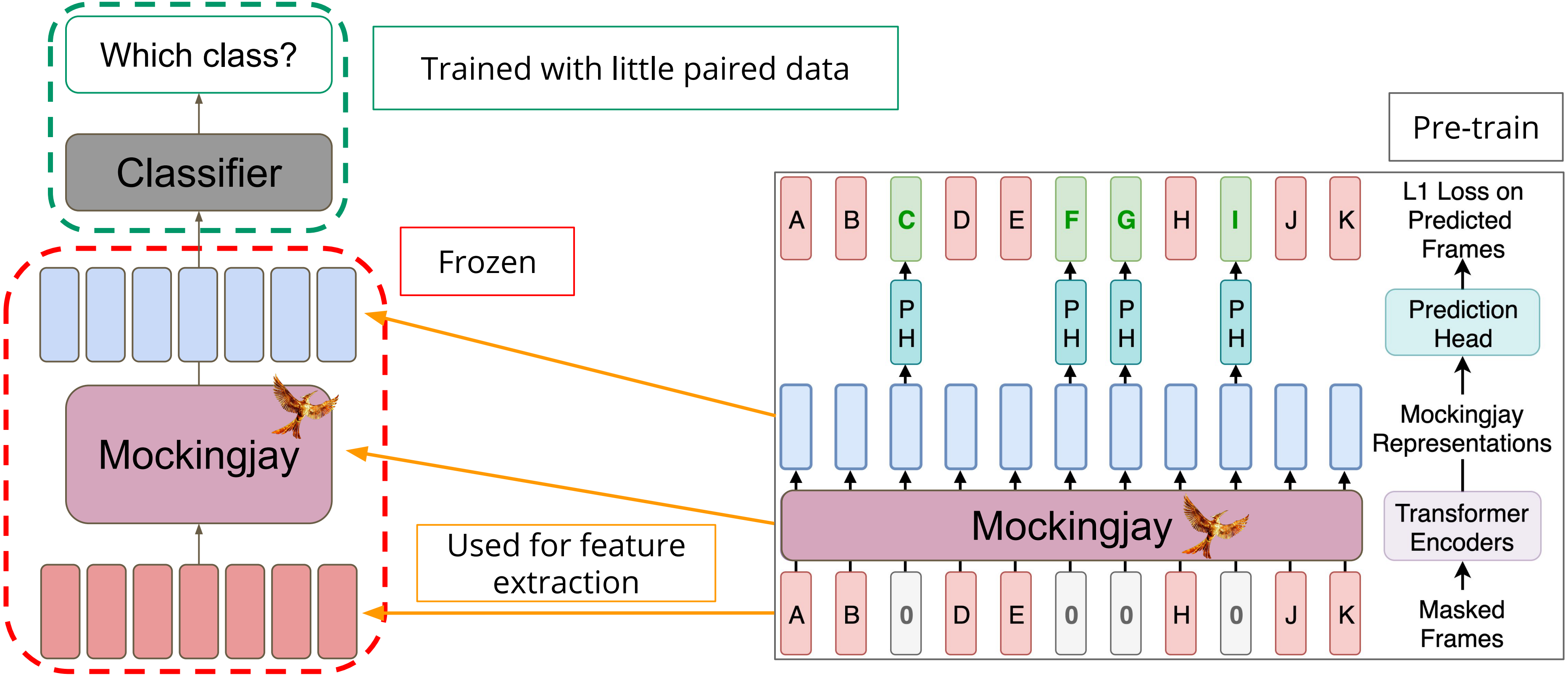
Incorporating with Downstream Tasks

1) Feature Extraction



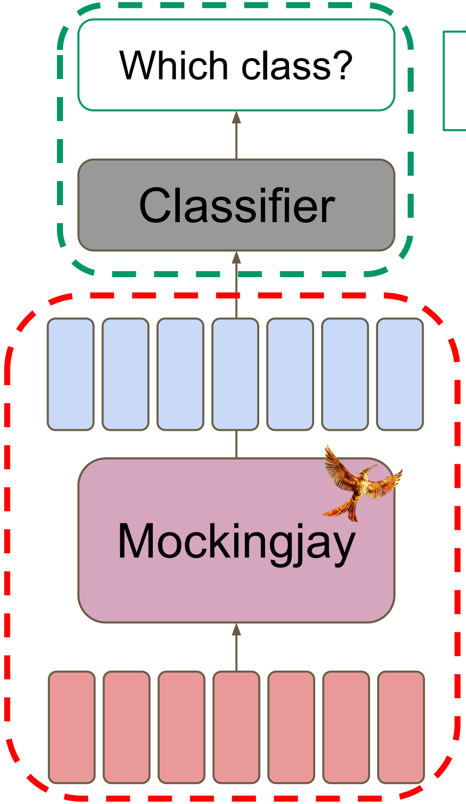
Incorporating with Downstream Tasks

1) Feature Extraction



Incorporating with Downstream Tasks

2) Weighted Sum from All Layers (WS)

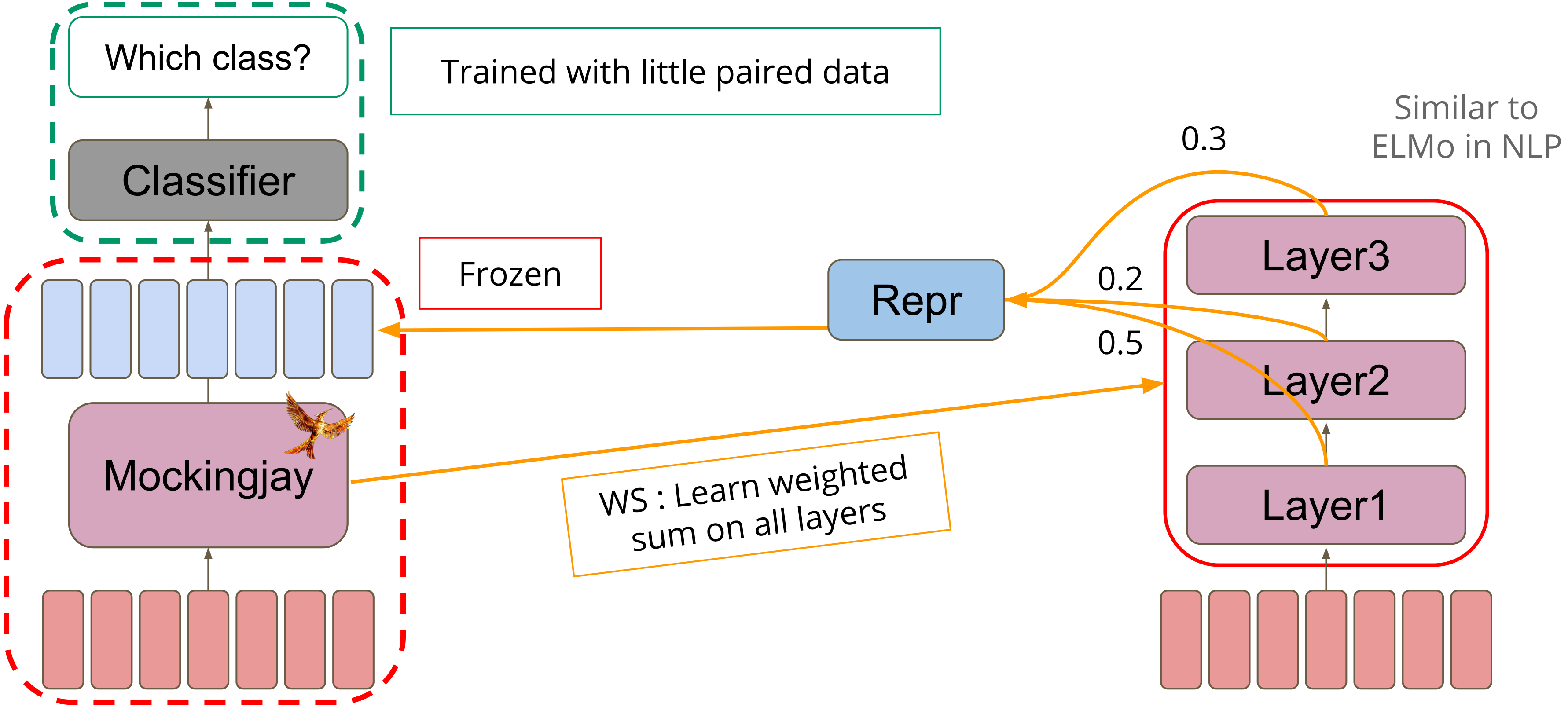


Trained with little paired data

Frozen

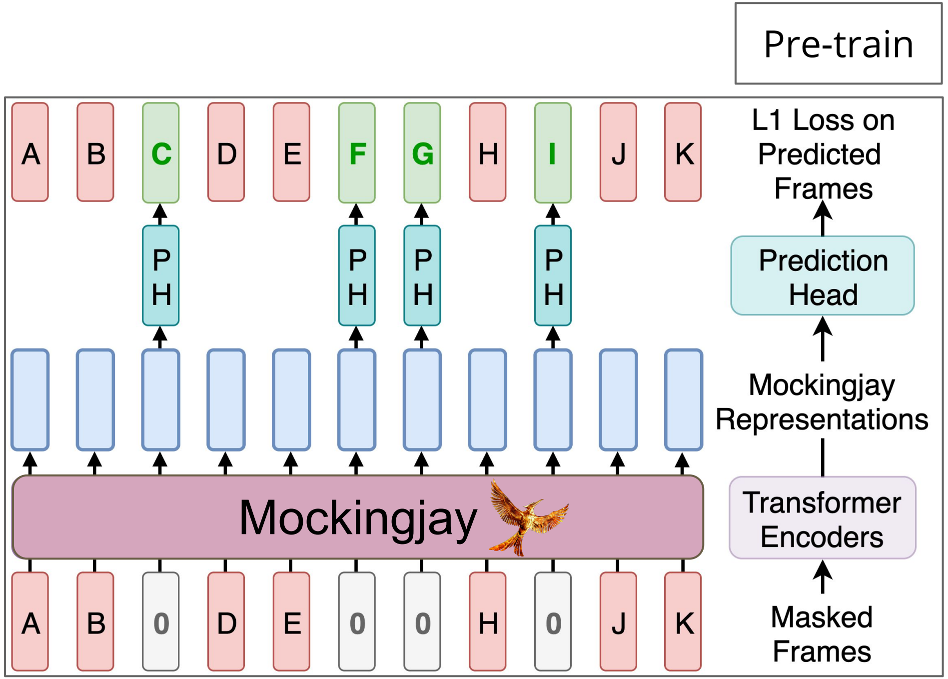
Incorporating with Downstream Tasks

2) Weighted Sum from All Layers (WS)



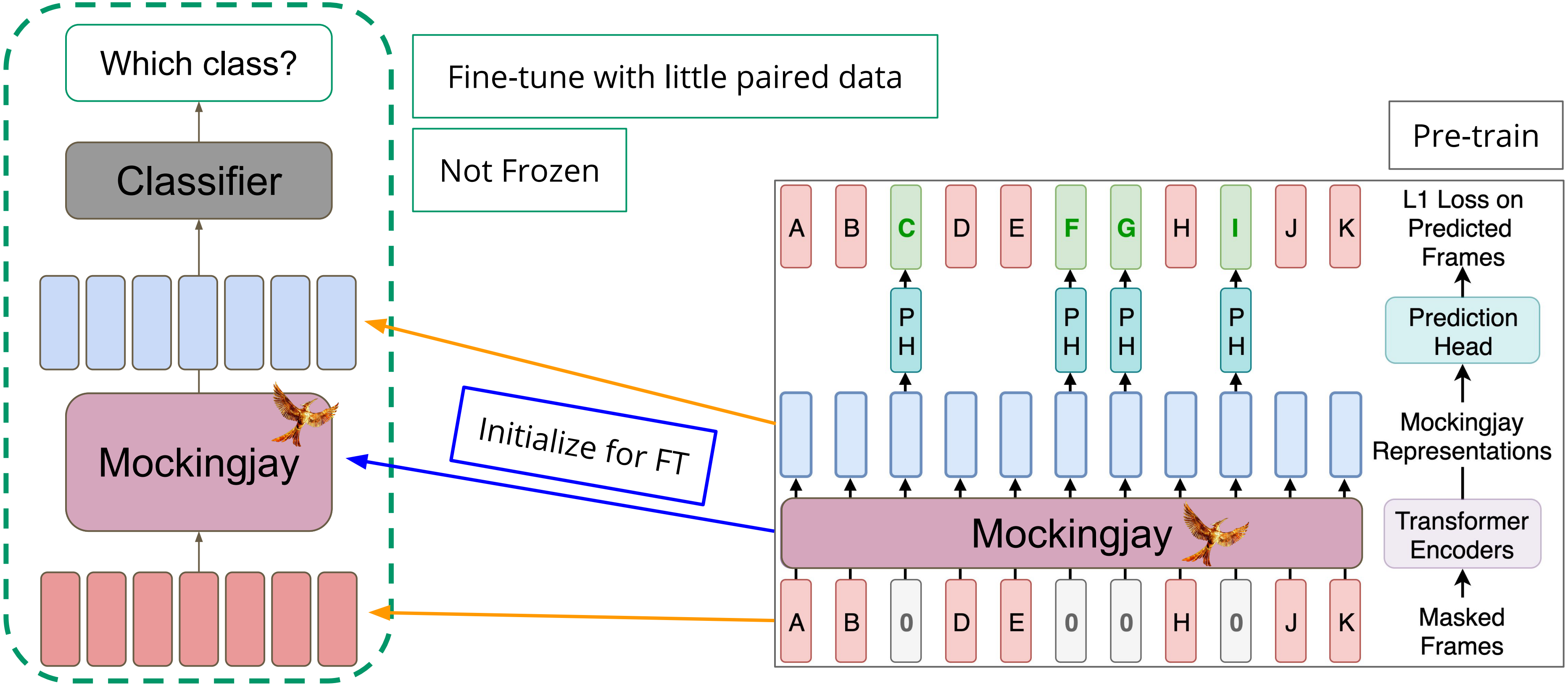
Incorporating with Downstream Tasks

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**

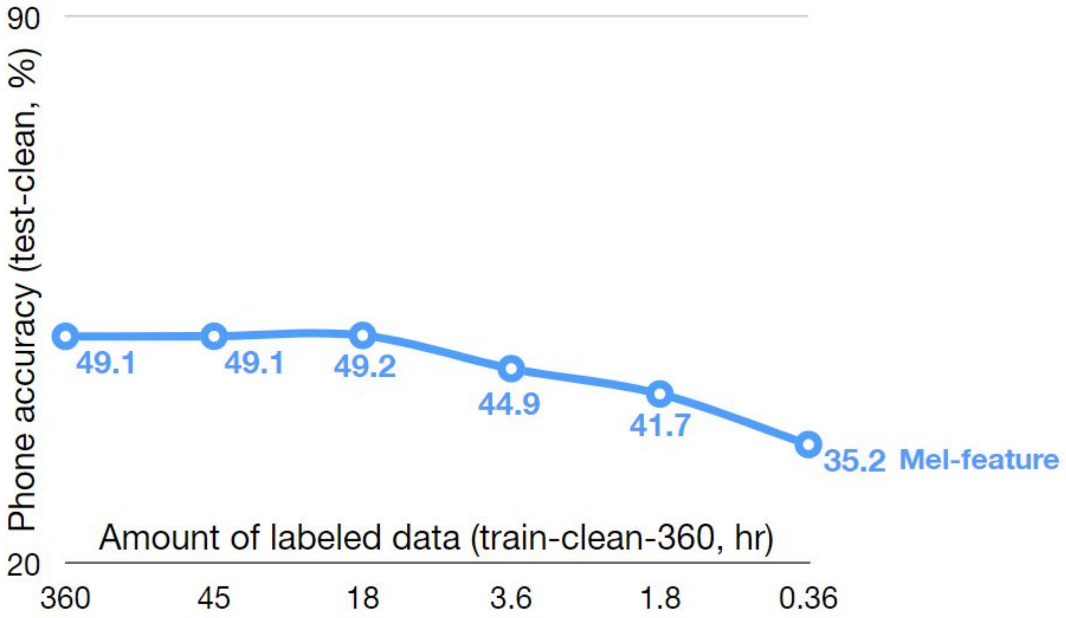
Consistent results over all three tasks:
Mel < BASE < LARGE

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**

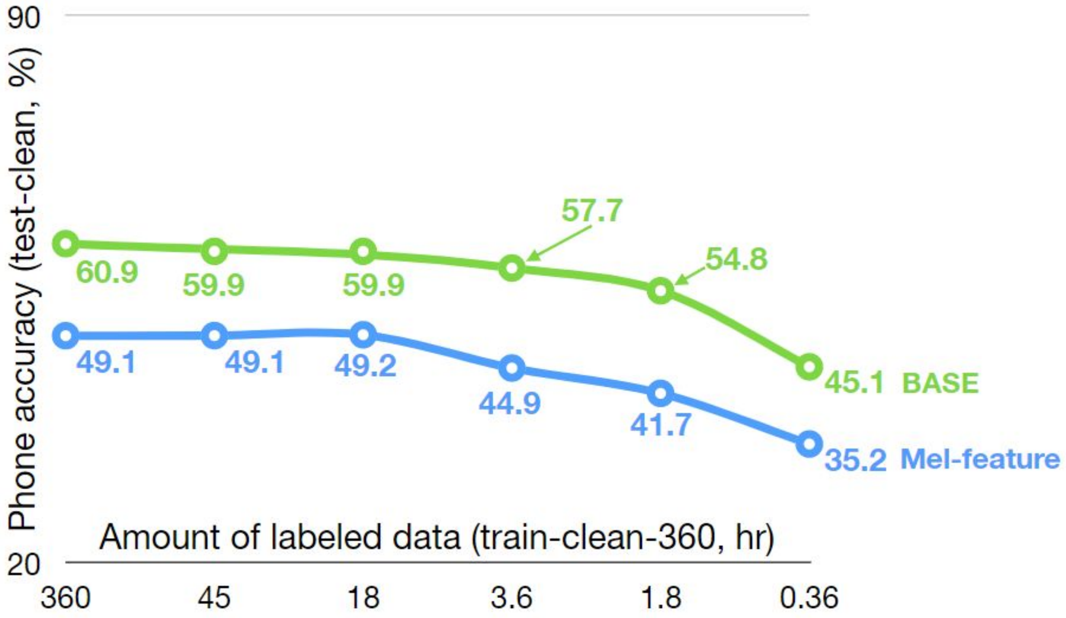
Consistent results over all three tasks:
LARGE < LARGE-WS

Low-Resource Experiments - 1/6



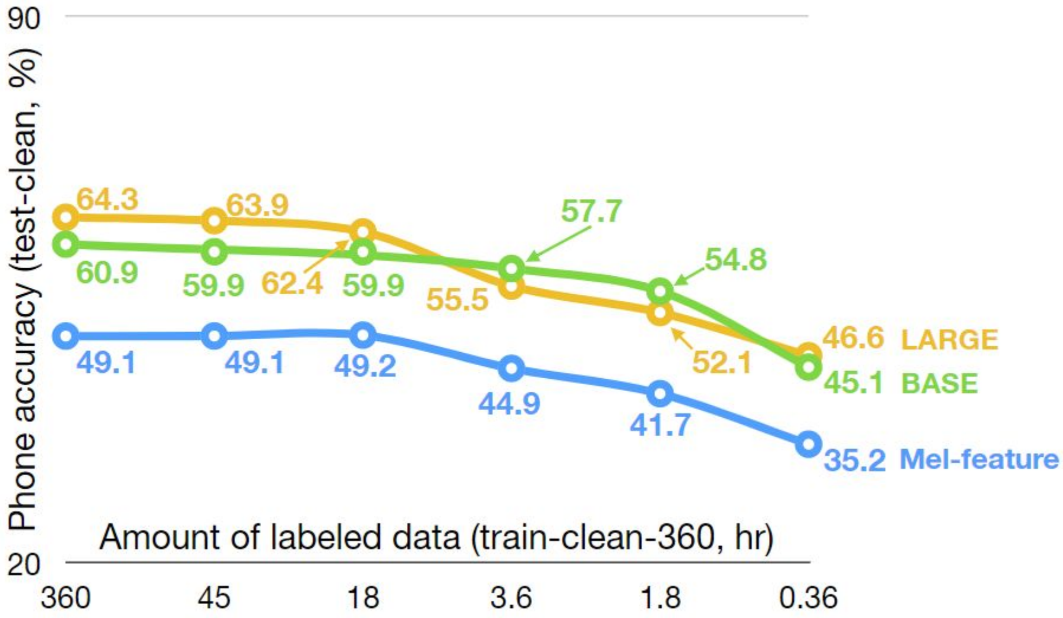
We demonstrate how pre-training on speech can improve supervised training in low resource scenarios, we train with reduced amount of labels.

Low-Resource Experiments - 2/6



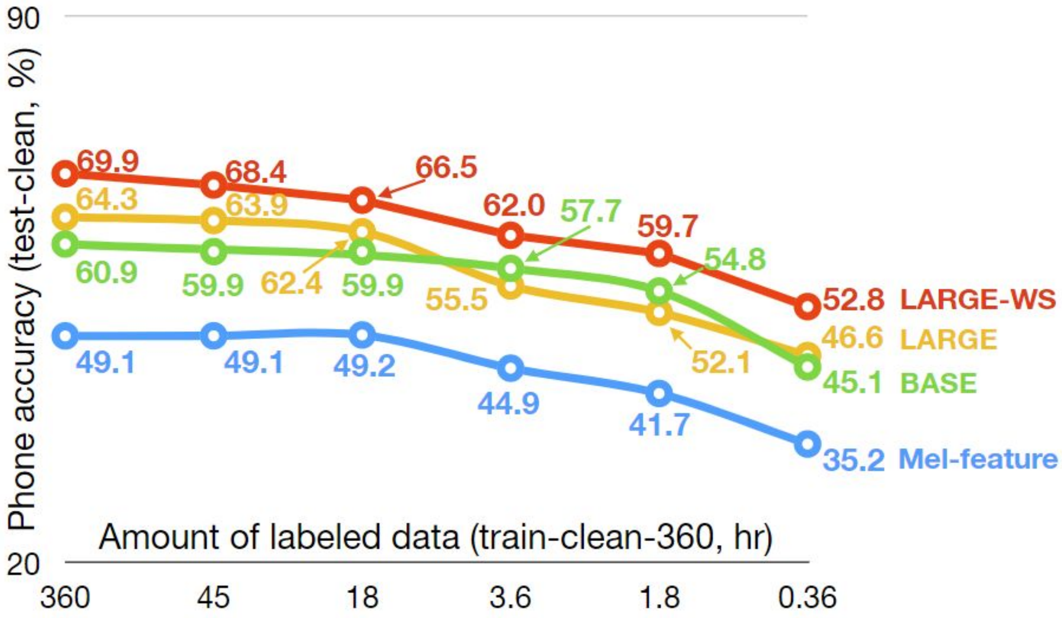
Mel < BASE

Low-Resource Experiments - 3/6



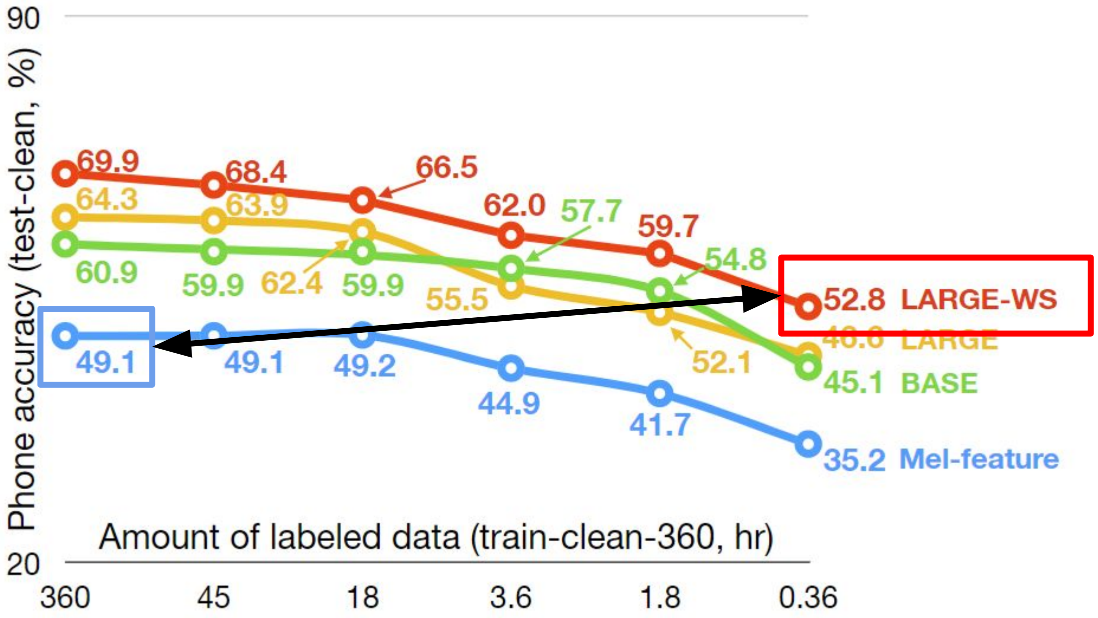
Mel < BASE < LARGE

Low-Resource Experiments - 4/6



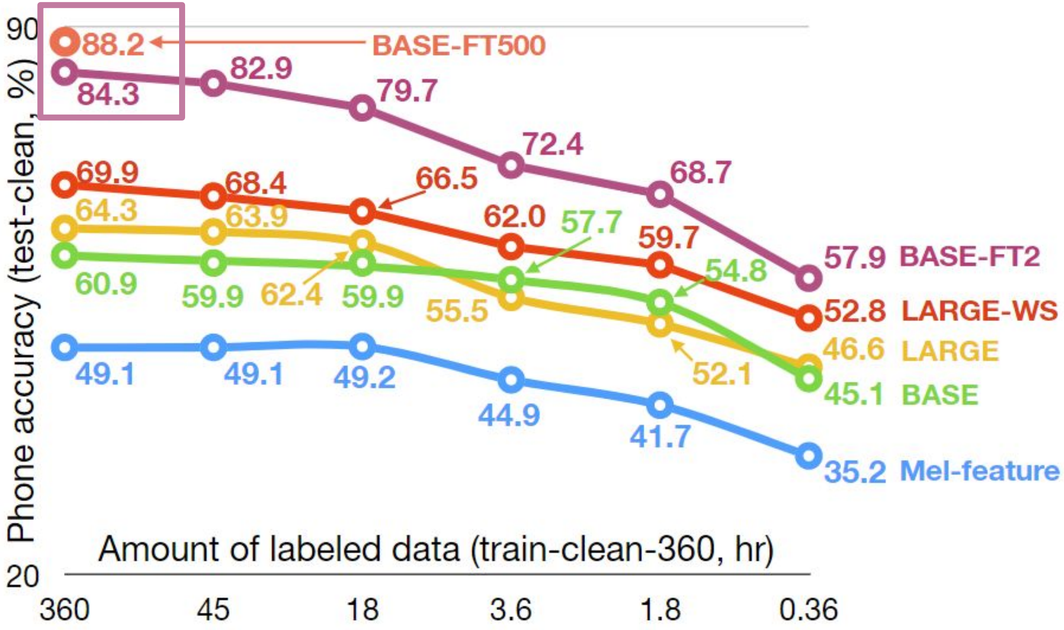
LARGE < LARGE-WS
with an avg 5.75% improvement

Low-Resource Experiments - 4/6



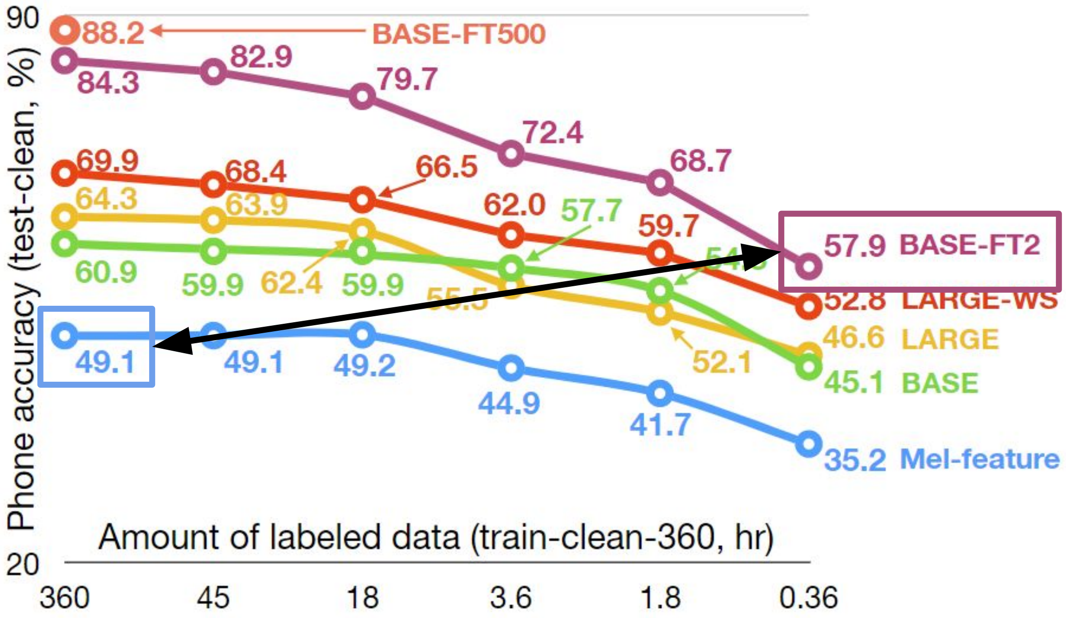
With 0.1% of labels,
LARGE-WS (52.8%) outperformed Mel (49.1%) that uses all 100% hours of labeled data.

Low-Resource Experiments - 5/6



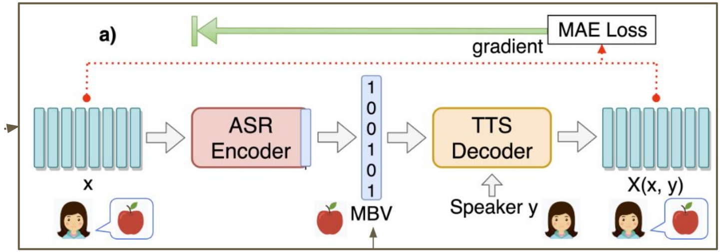
All < BASE-FT2

Low-Resource Experiments - 6/6

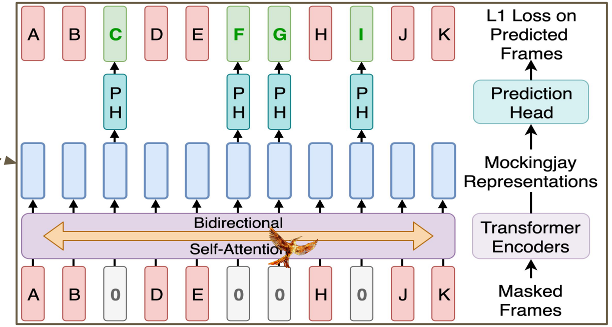


With 0.1% of labels,
BASE-FT2 (57.9%) outperformed Mel (49.1%) that uses all 100% hours of labeled data.

From here to beyond



- SSL on VC *(Interspeech 2019, first author Oral)*
- SSL on Mockingjay *(ICASSP 2020, first author Oral)*



Submitting to InterSpeech 2020 (5/15)

1. **Mockingjay for Adversarial Defence** *(2nd Author)*
2. **How Does Self-Supervised Models learn?** *(2nd Author)*
3. **Improving Mockingjay: Speech ALBERT** *(Advising)*
4. **Robust Neural Vocoding for Speech Generation** *(3rd Author)*
   Train WaveNet, WaveRNN, FFTNet, Parallel WaveGAN alternately on five different datasets.

Current Works

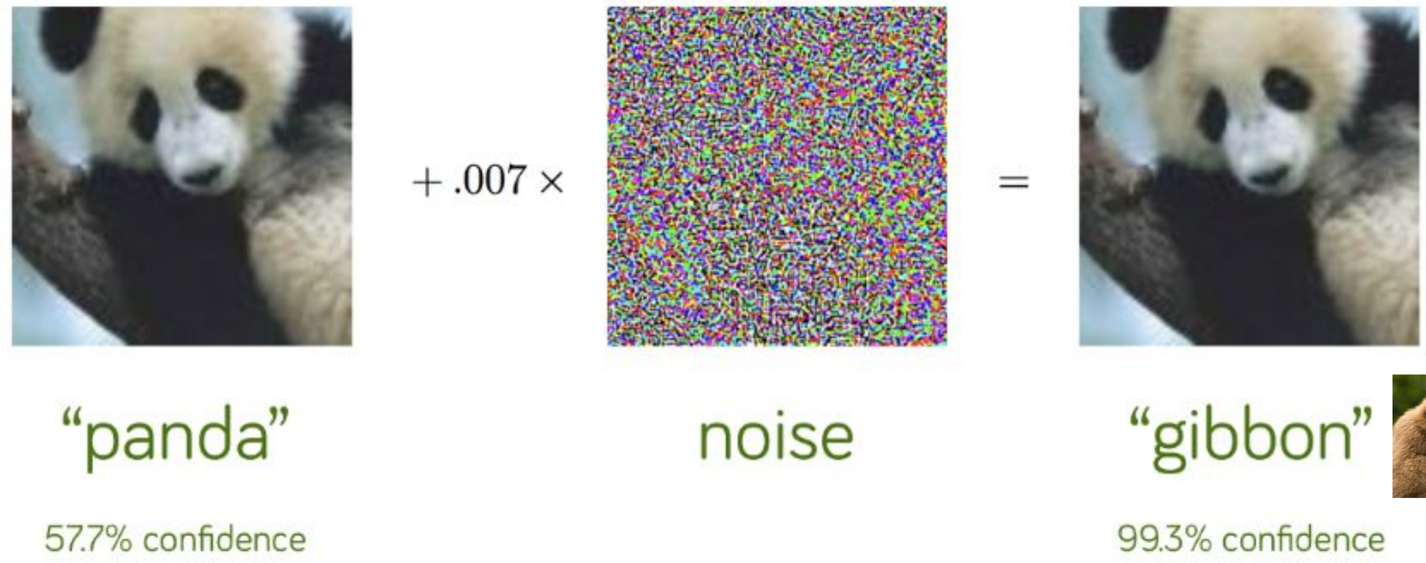
What else can we do with Mockingjay?

1. Adversarial Defense

Employ Mockingjay to protect models against adversarial attacks

1. Adversarial Defense

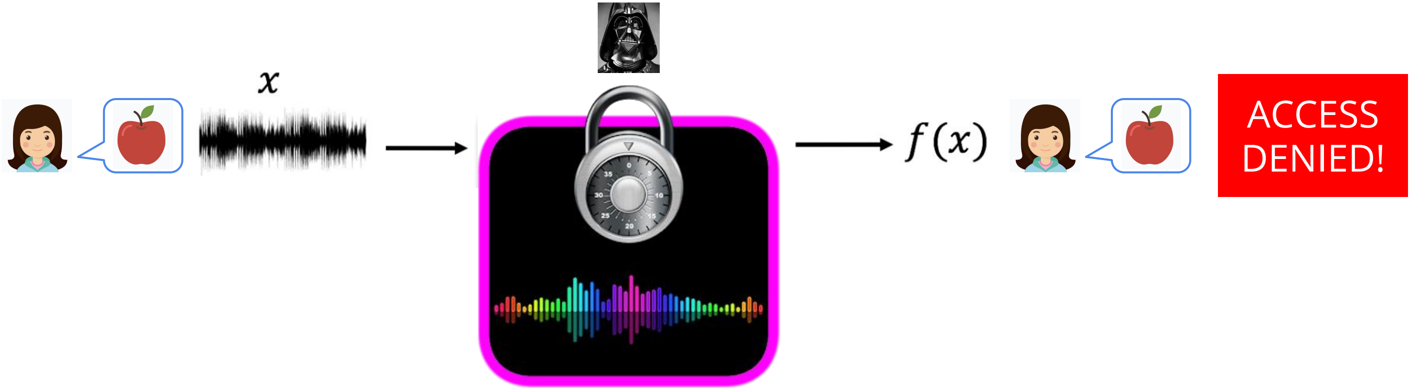
What is Adversarial Attack?



Hacking AI security systems: Face ID / Voice ID

1. Adversarial Defense

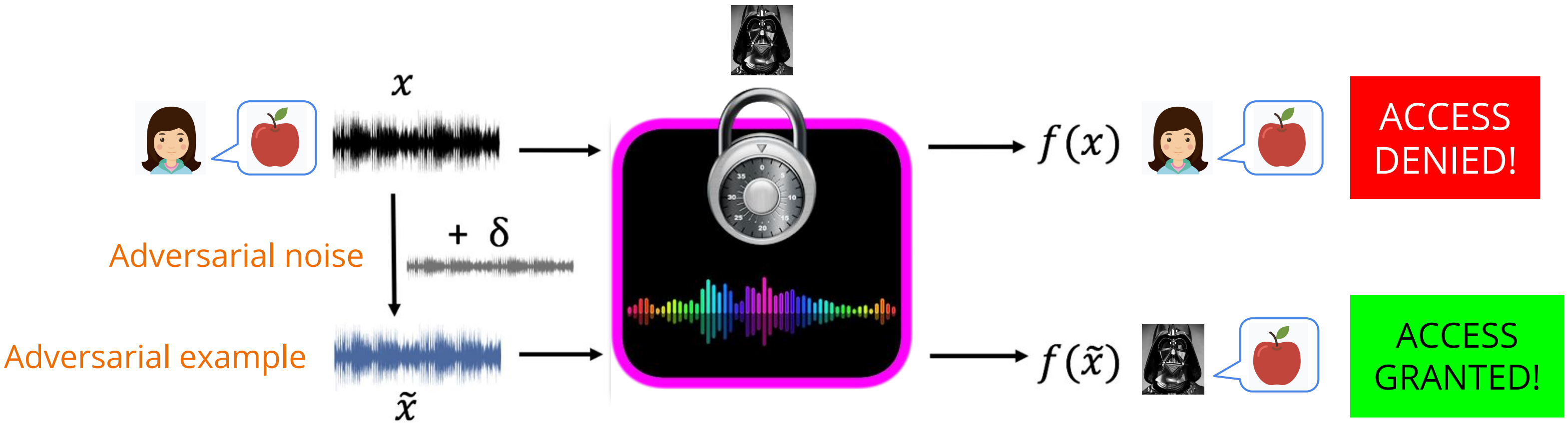
What is Adversarial Attack?



Hacking AI security systems: Face ID / Voice ID

1. Adversarial Defense

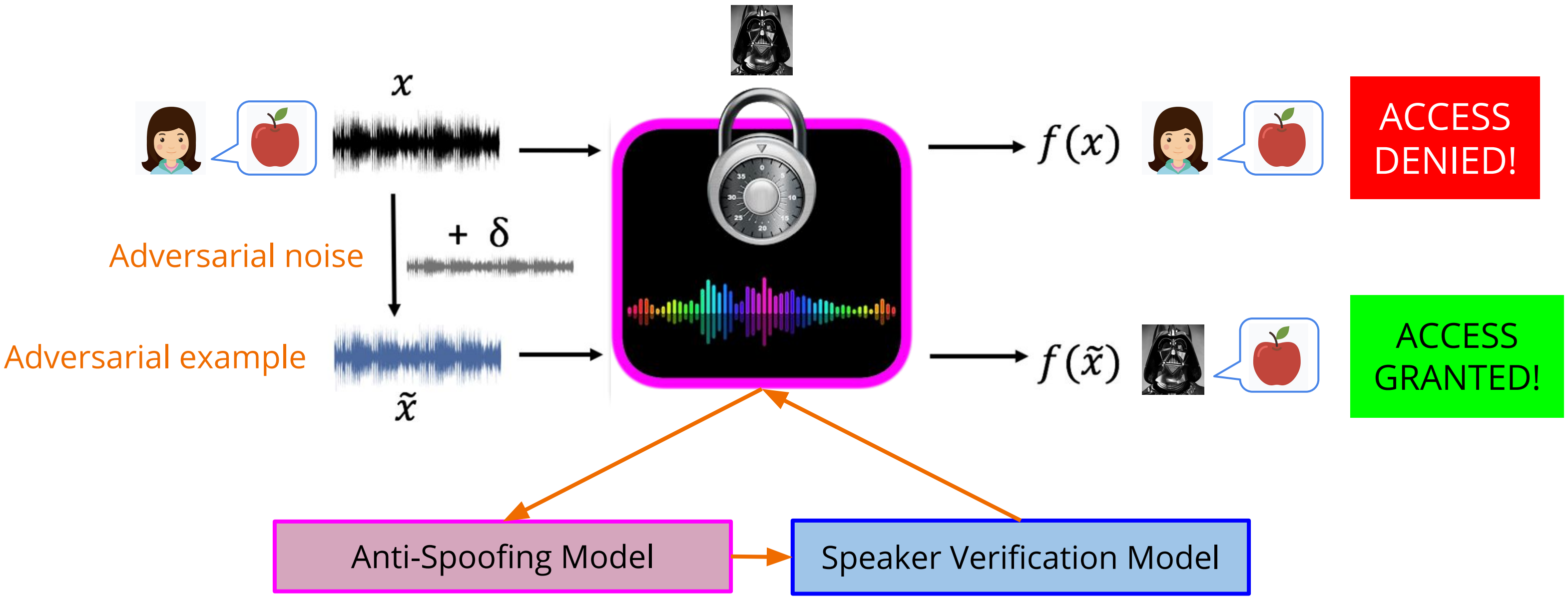
What is Adversarial Attack?



Hacking AI security systems: Face ID / Voice ID

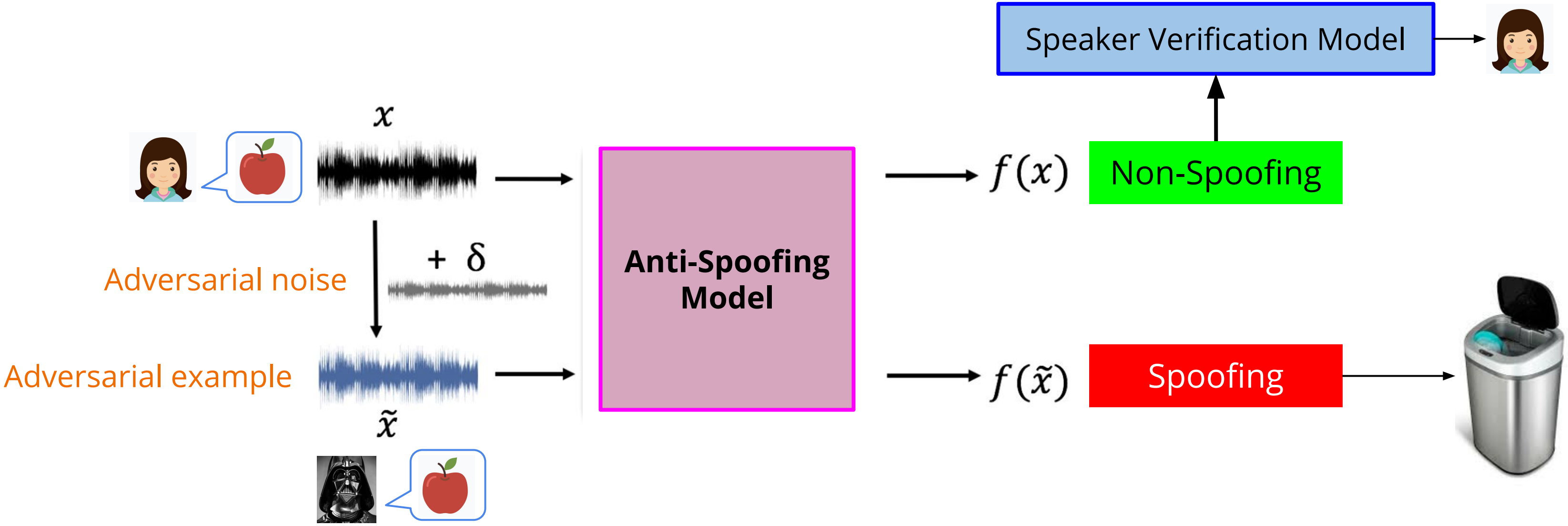
1. Adversarial Defense

What is Adversarial Attack?



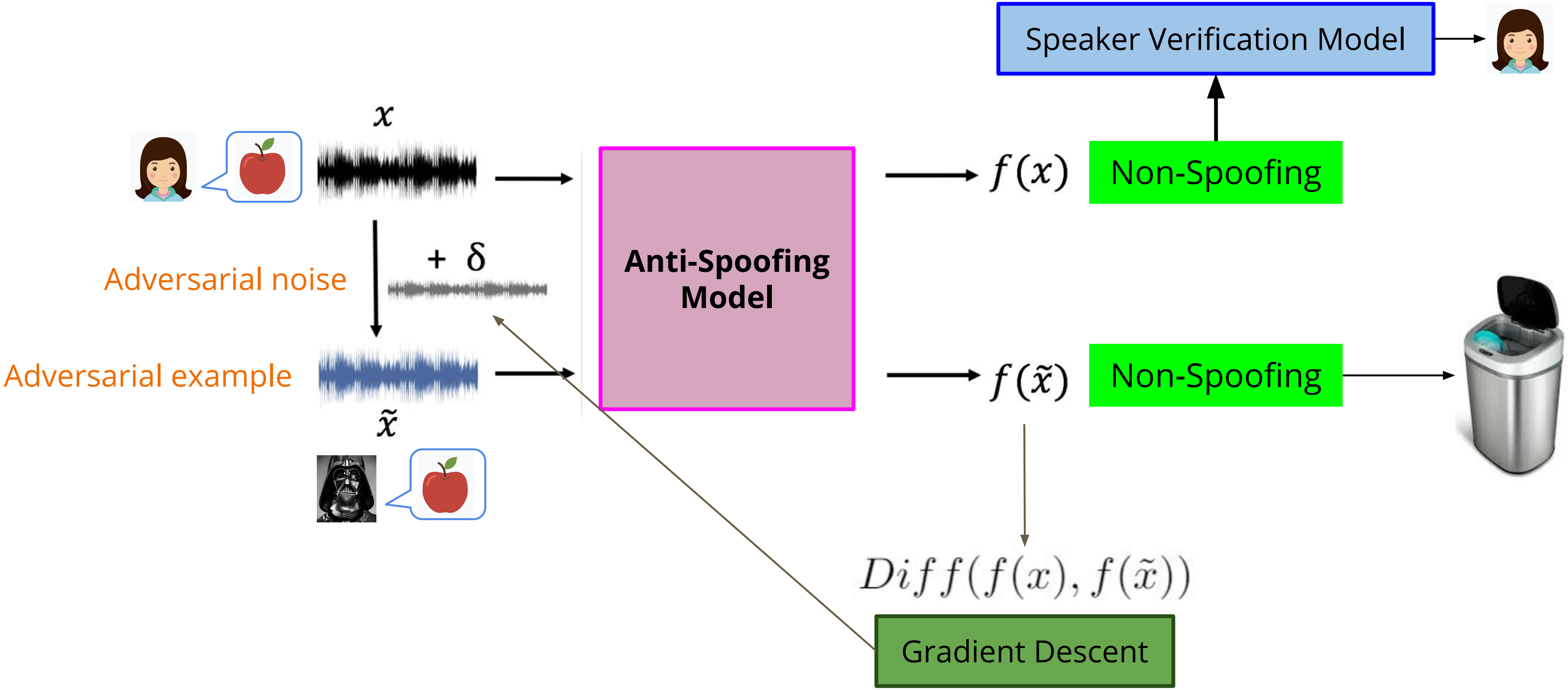
1. Adversarial Defense

How to Attack?



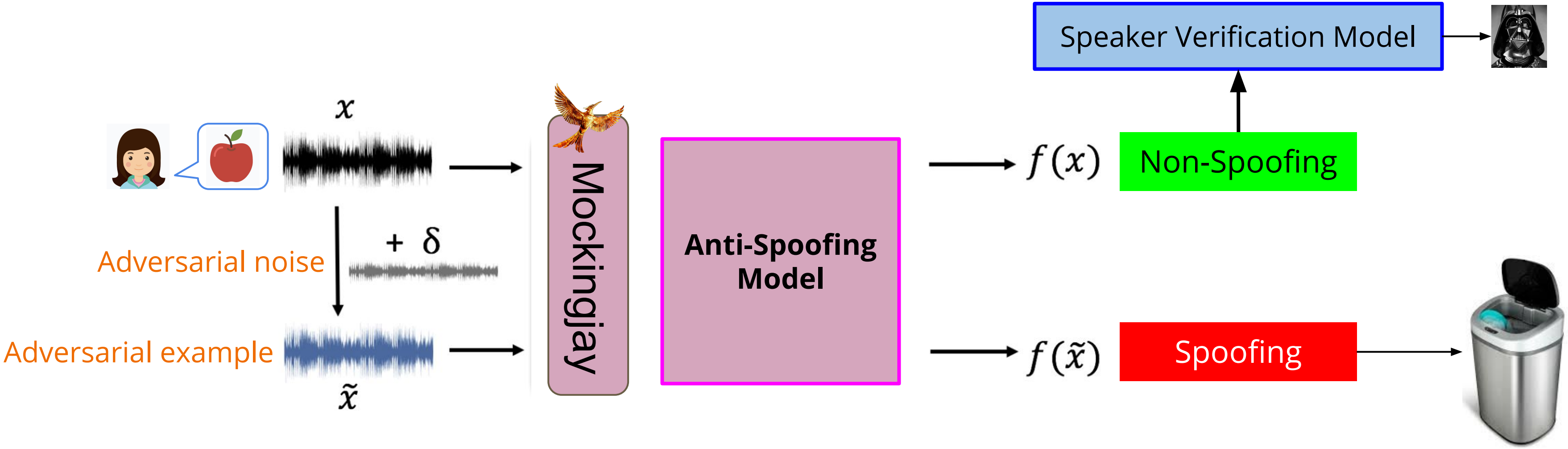
1. Adversarial Defense

How to Attack?

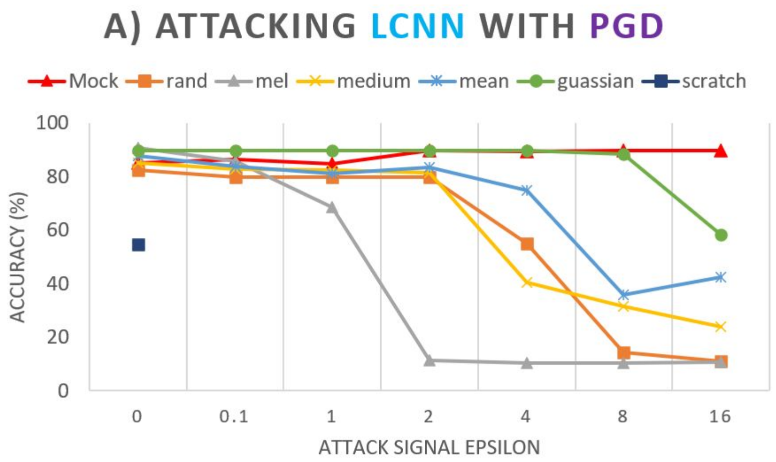


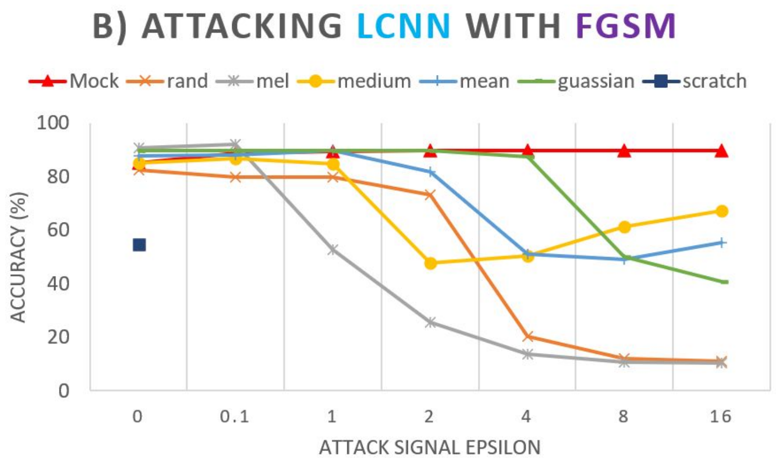
1. Adversarial Defense

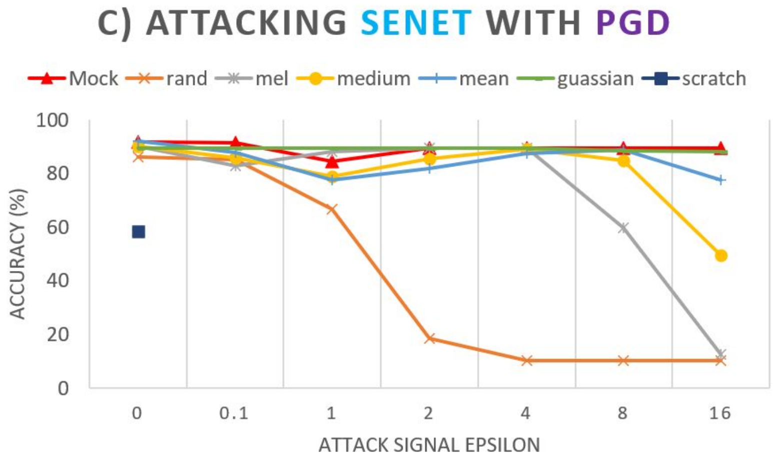
Employing Mockingjay

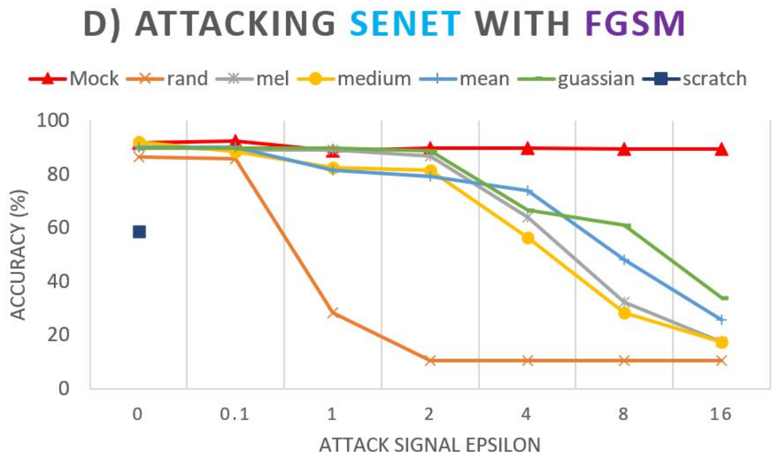


1. Adversarial Defense - Experiments

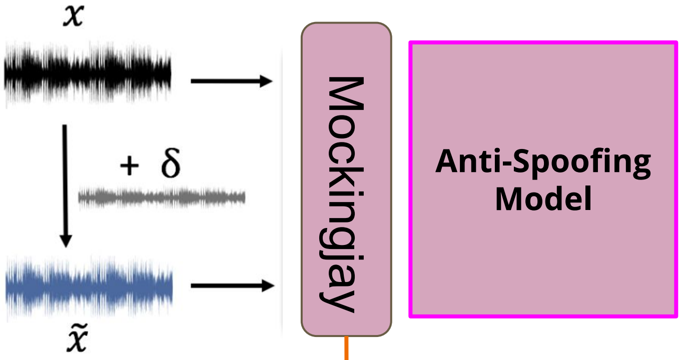


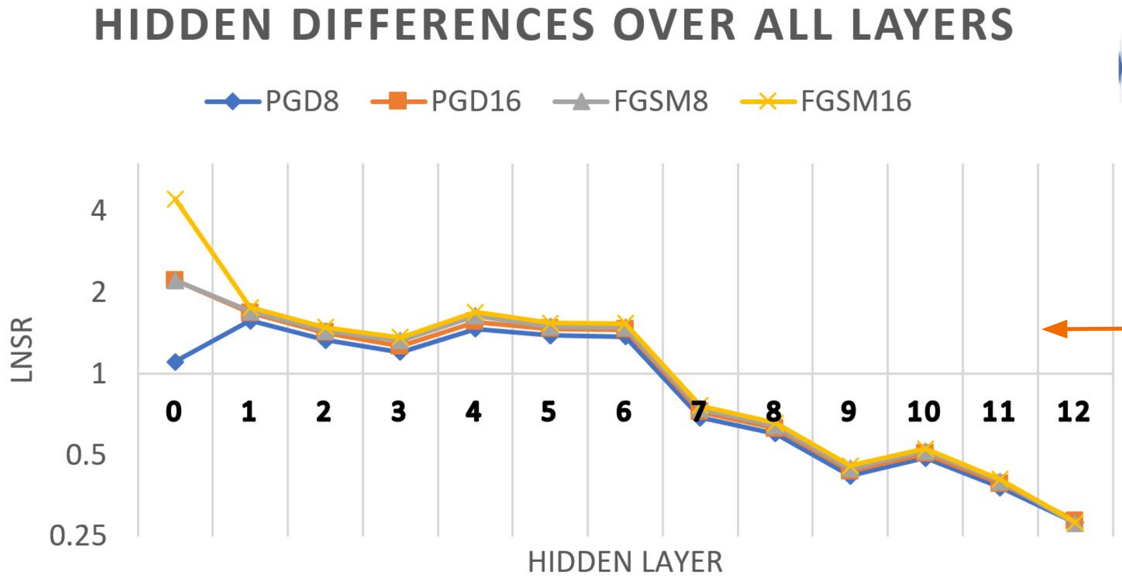






1. Adversarial Defense - Experiments





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

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

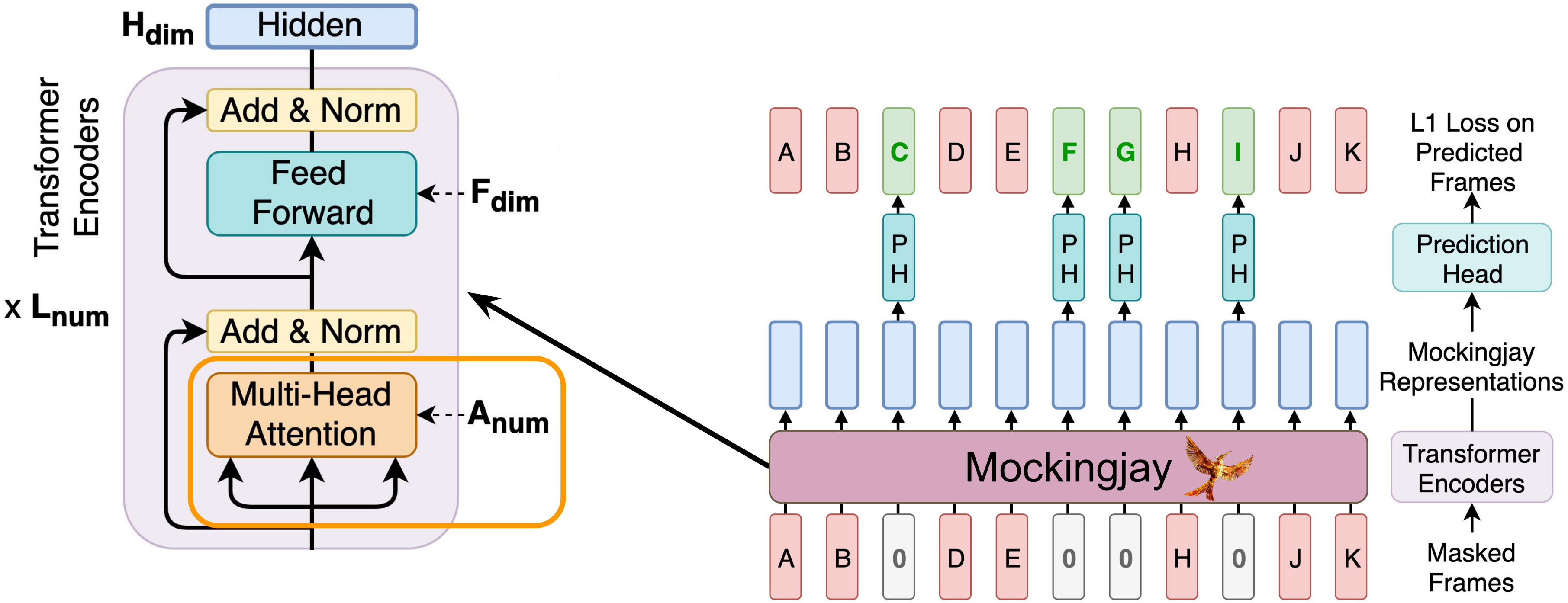
Current Works

What else can we do with Mockingjay?

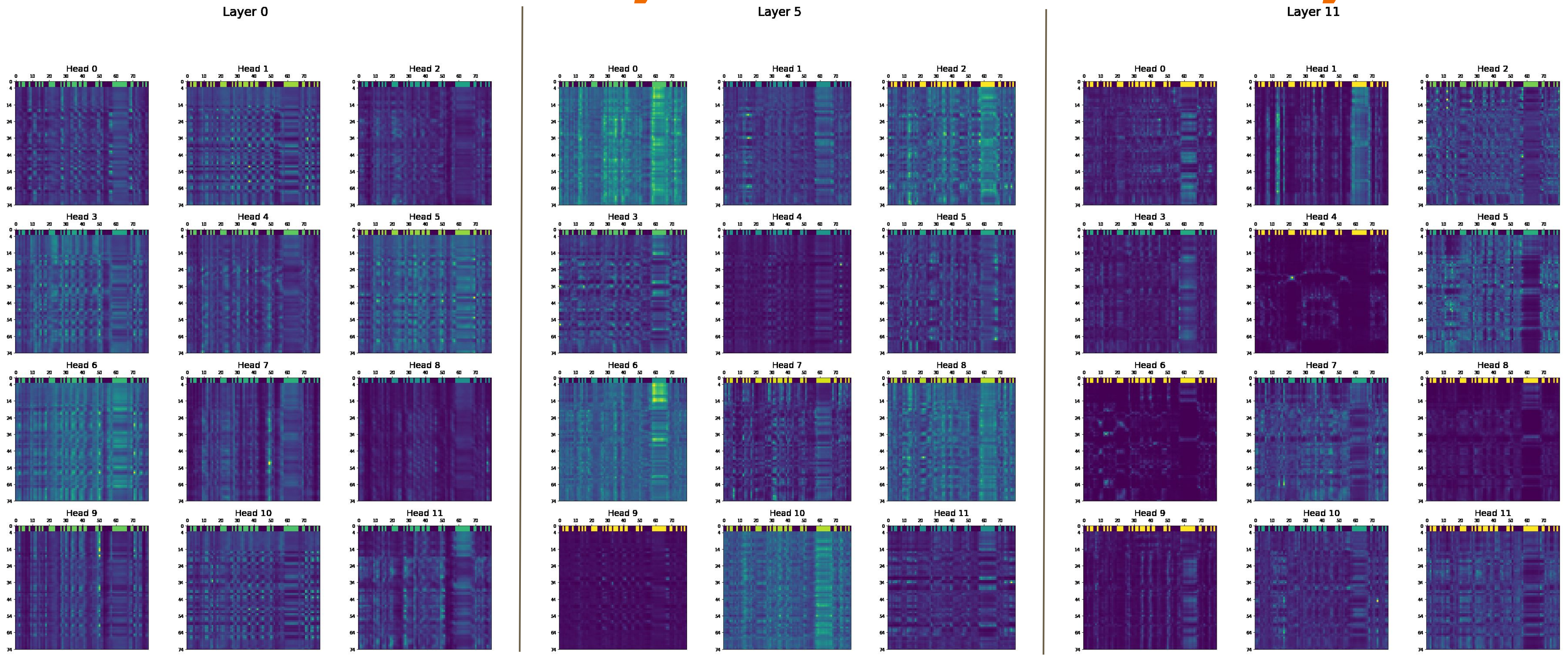
2. Understanding Self-Supervised Models

Visualize and explainable understanding of how models draw conclusion

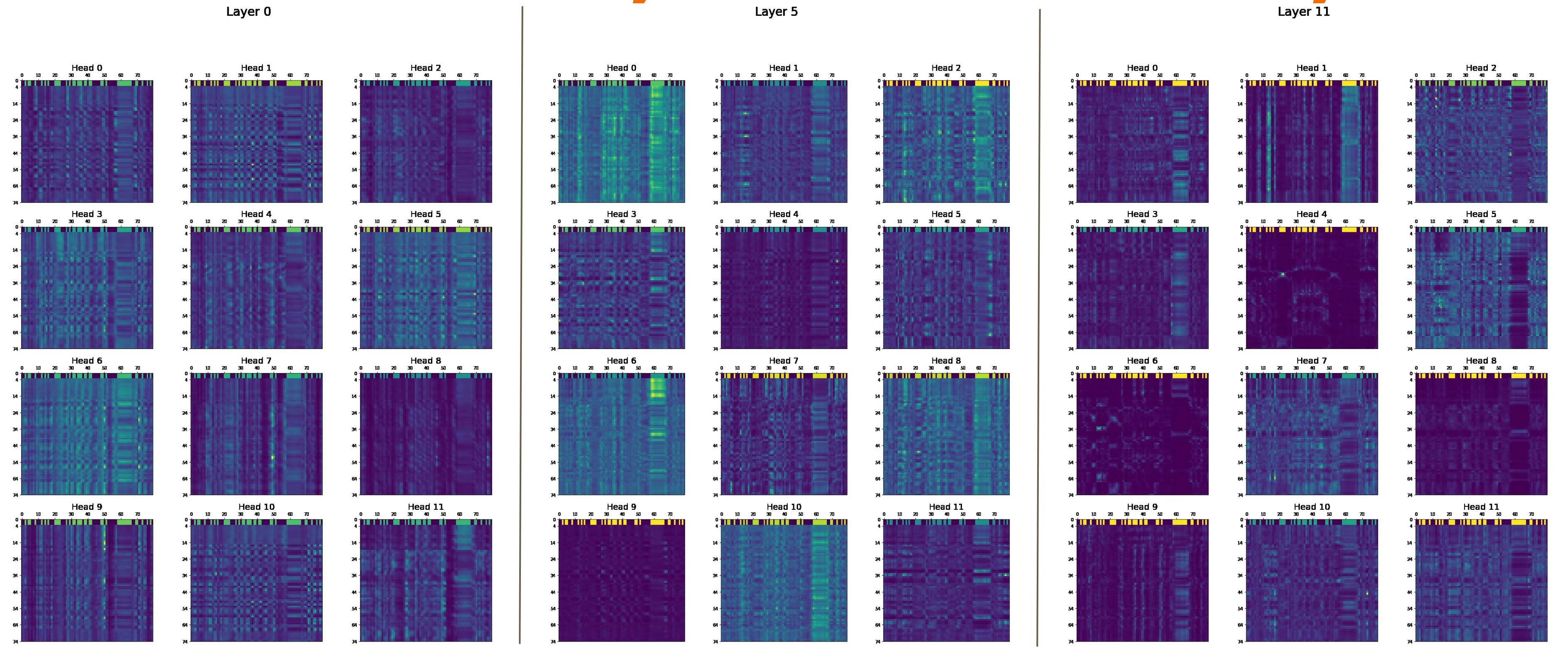
Recall: Model Architecture

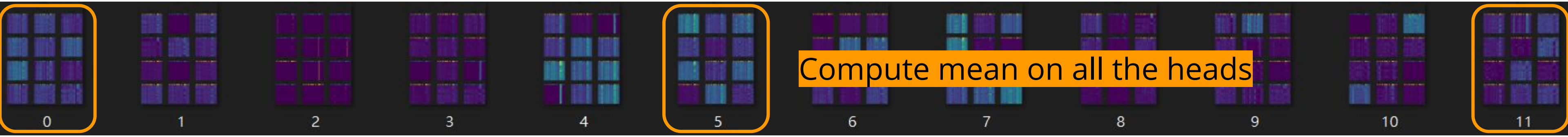


Attention of layers - 12 Heads Summary

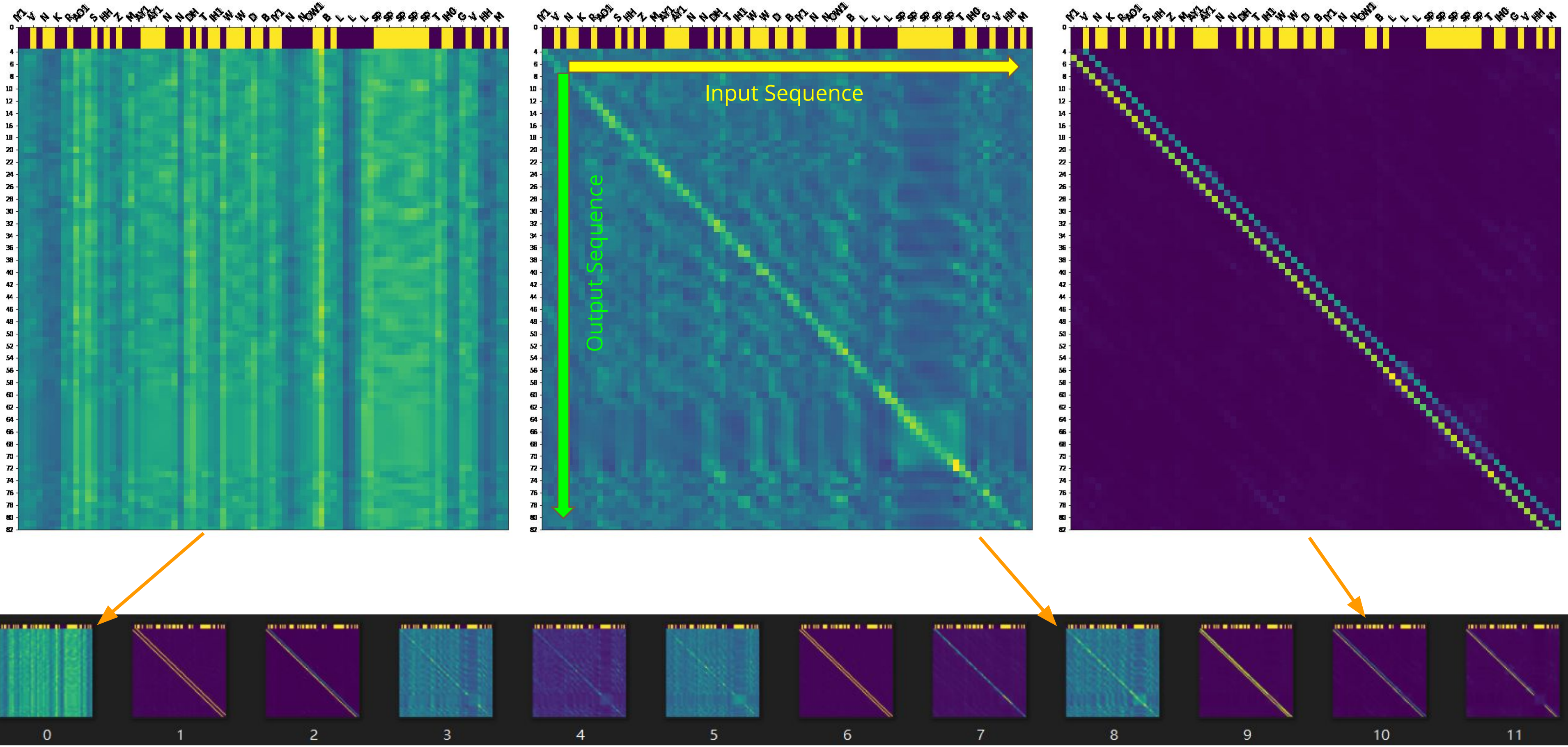


Attention of layers - 12 Heads Summary

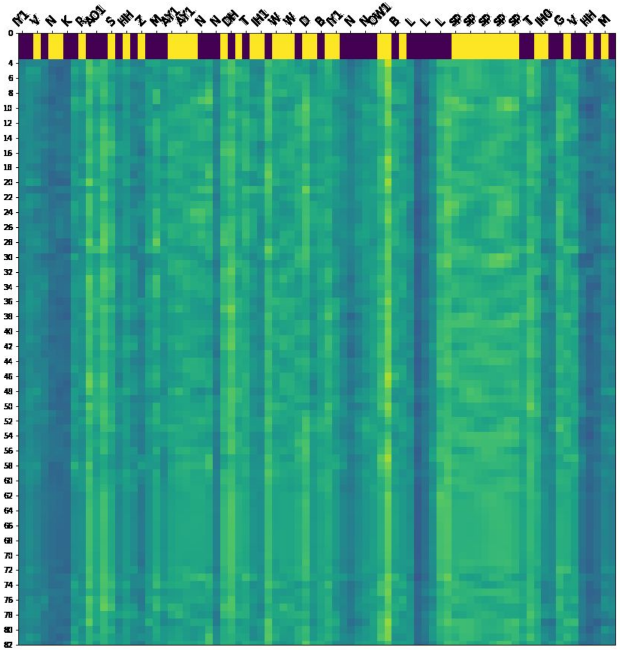




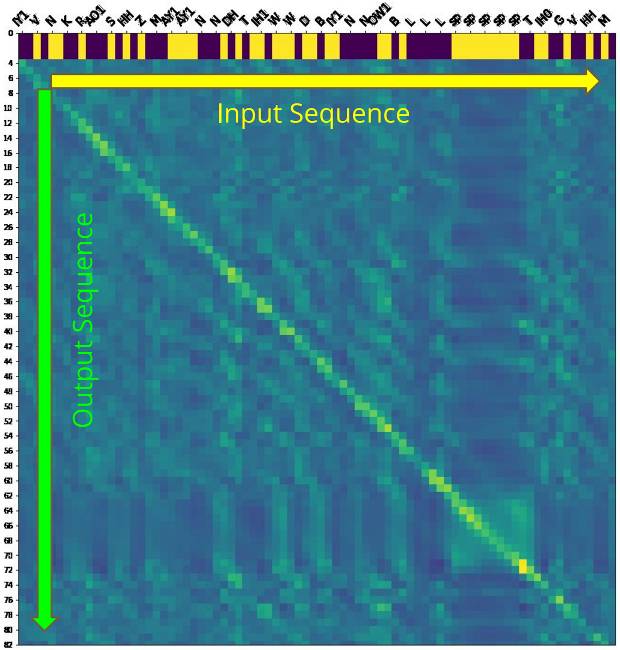
Attention of layers - What Each Layer Does?



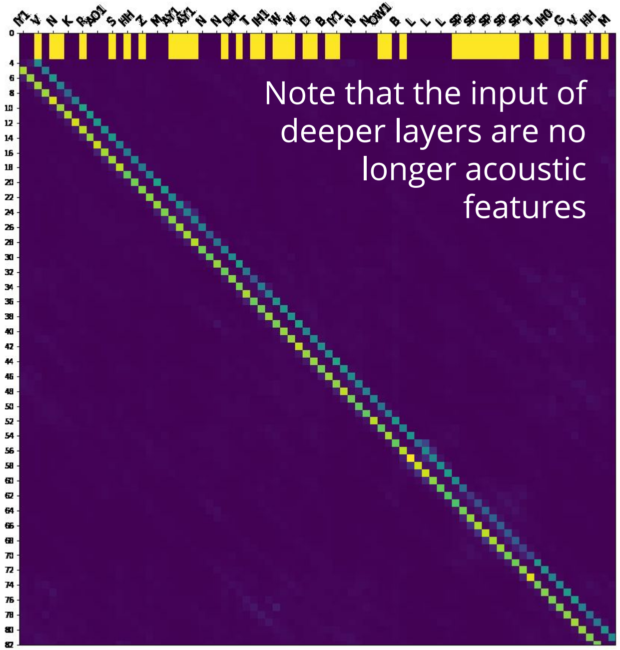
Attention of layers - What Each Layer Does?



Layer 0 - *Vertical*
Observing the model dividing phoneme boundaries



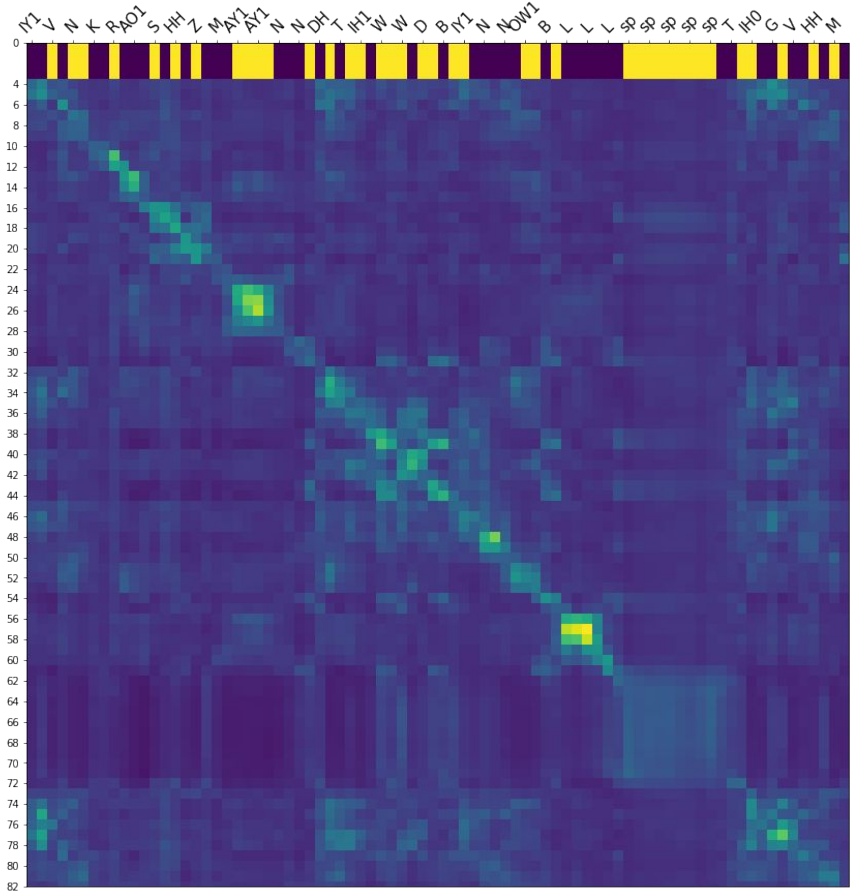
Layer 8 - *Global*
The model gathering global structures



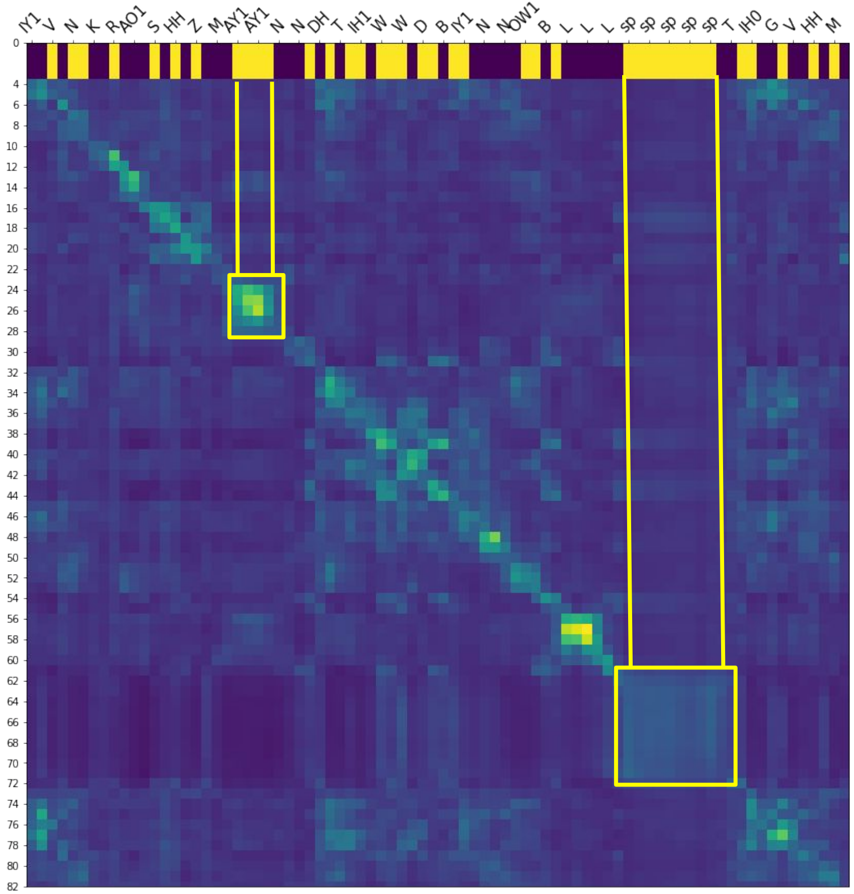
Layer 11 - *Diagonal*
The model attends on past and future context to predict output

These 3 actions are performed iteratively throughout the depth of the model

Diagonal - Observing Phoneme Boundaries

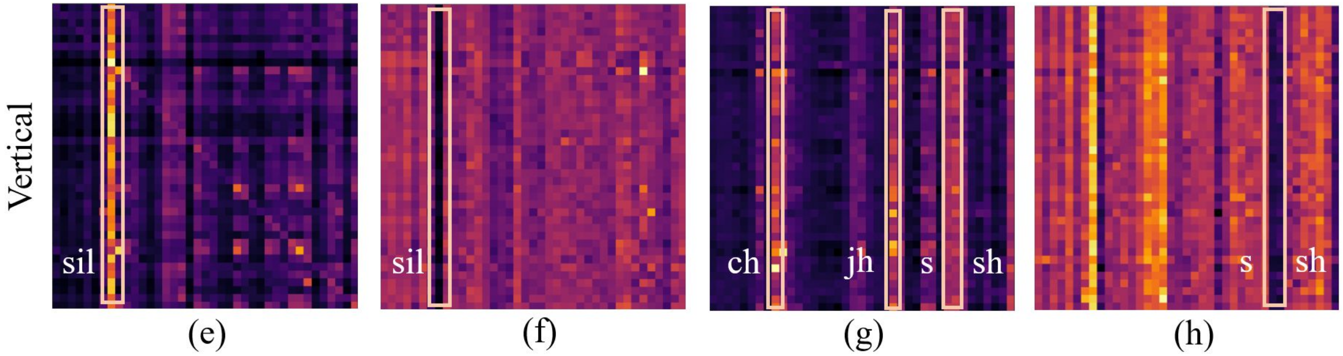


Diagonal - Observing Phoneme Boundaries



diagonal attentions are highly correlated with phoneme boundaries

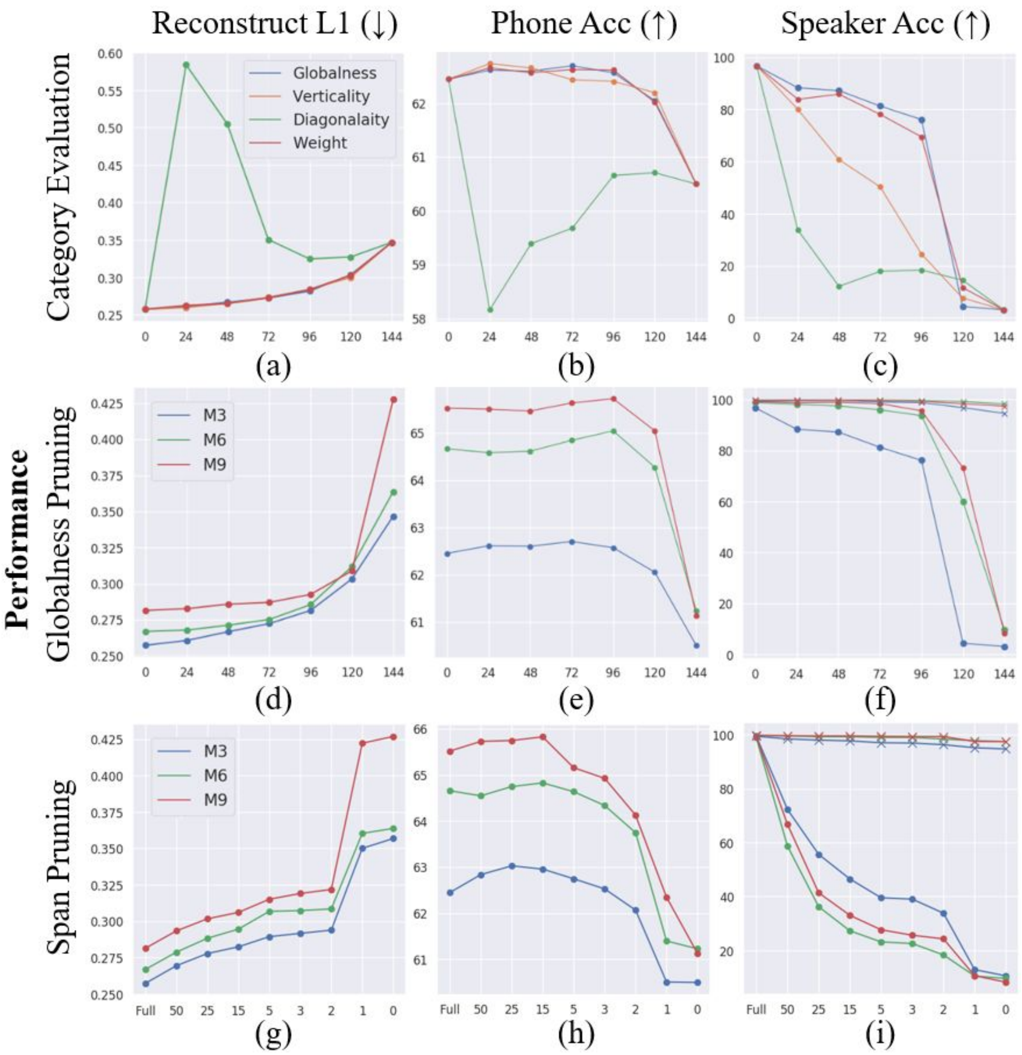
Vertical - Observing concentration



(d) not attends to identity (e) attends to sil (f) not attends to sil (g) attends to ch, jh, s, sh (h) not attends to s, sh.

vertical attentions
often concentrate on
specific phonemes

Refine Attentions

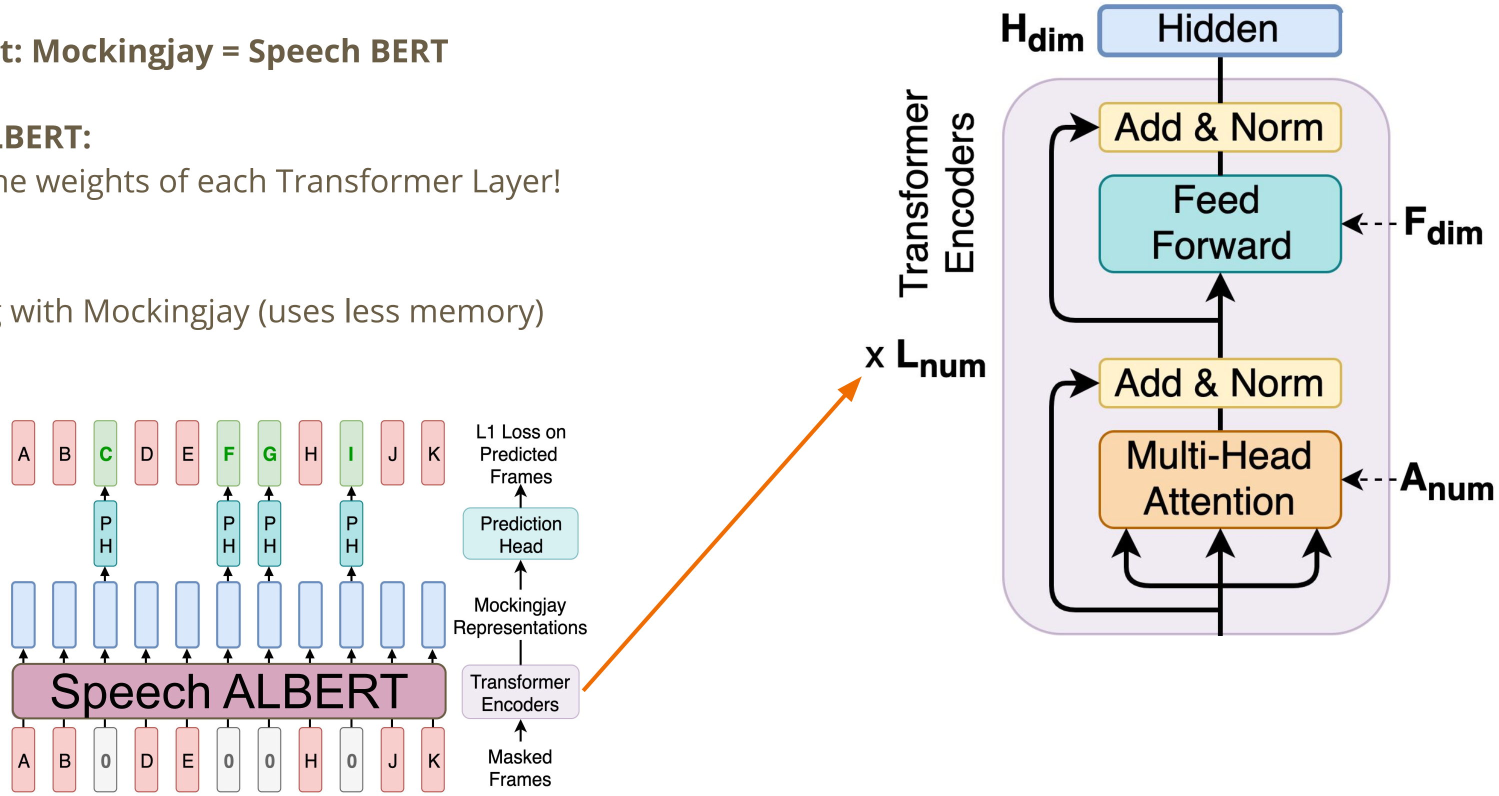


3. Speech ALBERT

Recall that: Mockingjay = Speech BERT

Speech ALBERT:
Share all the weights of each Transformer Layer!

Exps:
comparing with Mockingjay (uses less memory)

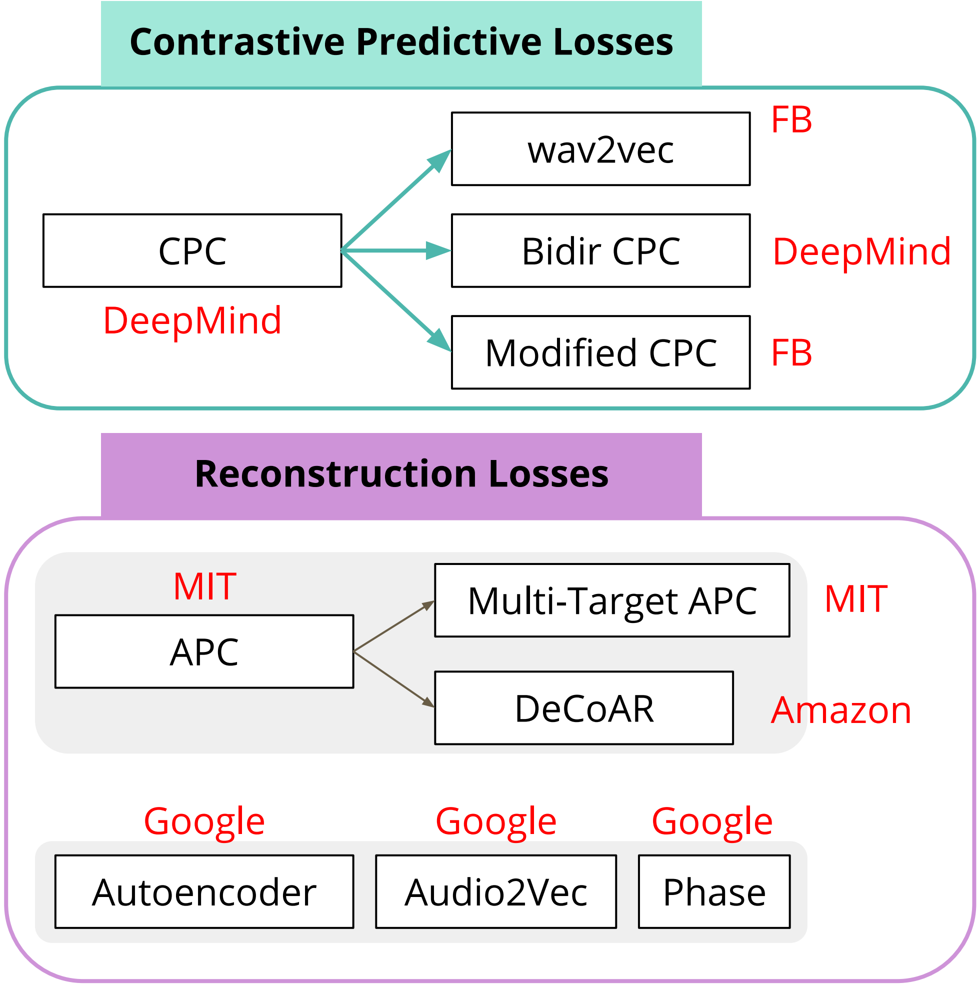


Related Works

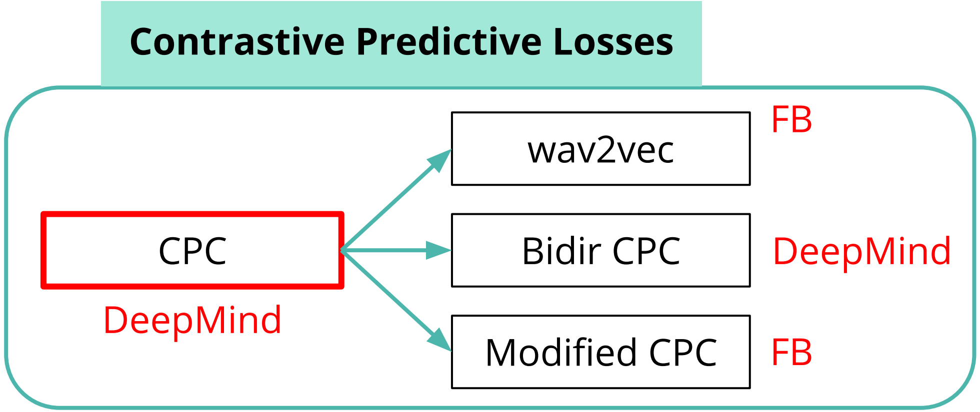
What else besides Mockingjay?

A top-down introduction to all recent related works.

Related Works

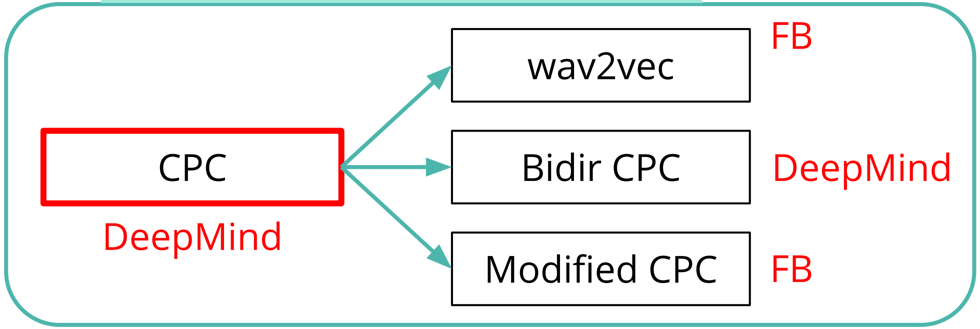


Related Works



Related Works

Contrastive Predictive Losses

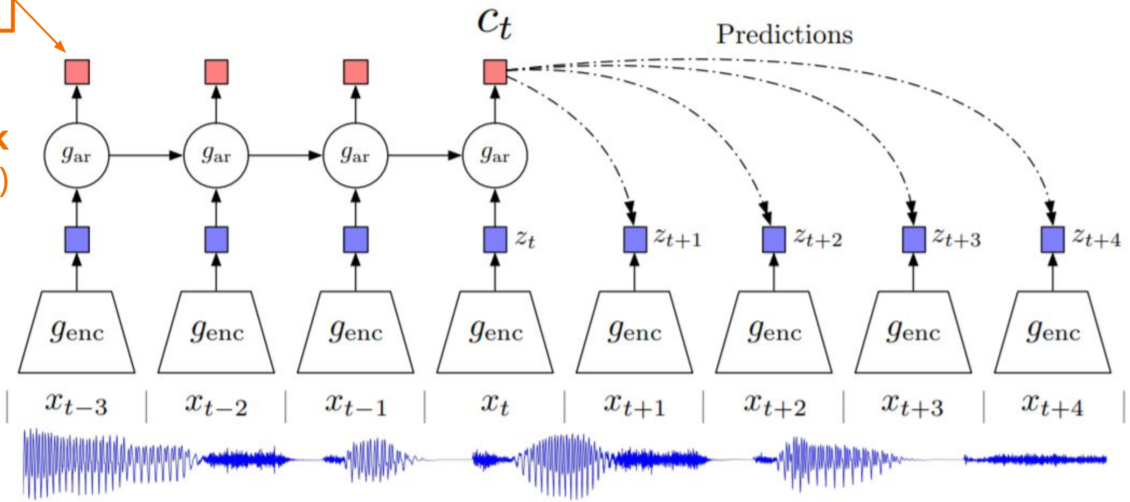


Intuition:
Pulls temporally nearby representations closer and pushes temporally distant ones further.

Learned representations

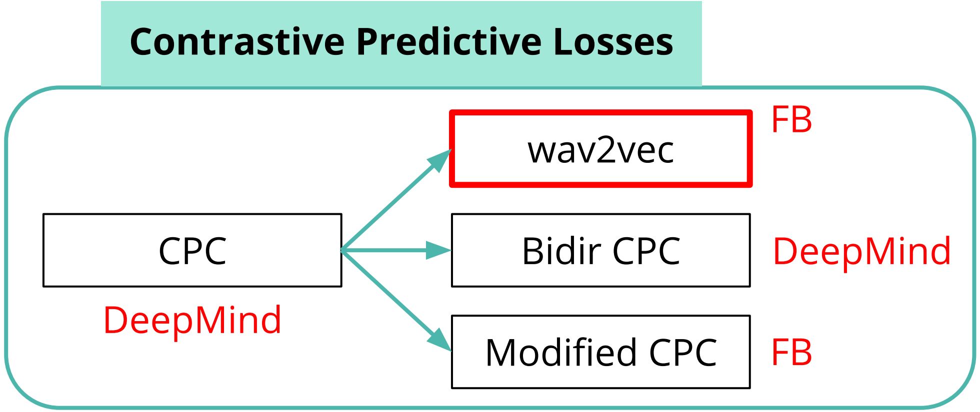
Context Network
1-layer Gated Recurrent Units (GRU)

Encoder Network
5-layer convolutional neural network (CNN)



Important exps: Phone / Speaker Classification
*We will compare with this in our work.

Related Works



Related Works

Contrastive Predictive Losses

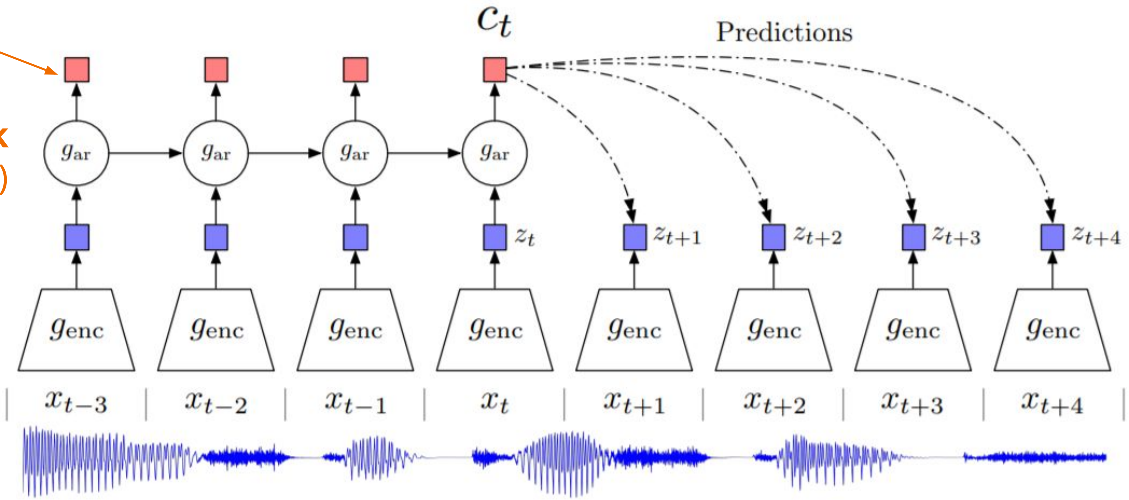
- CPC — DeepMind
  - wav2vec — FB
  - Bidir CPC — DeepMind
  - Modified CPC — FB

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

Used as input for ASR models, replace acoustic features

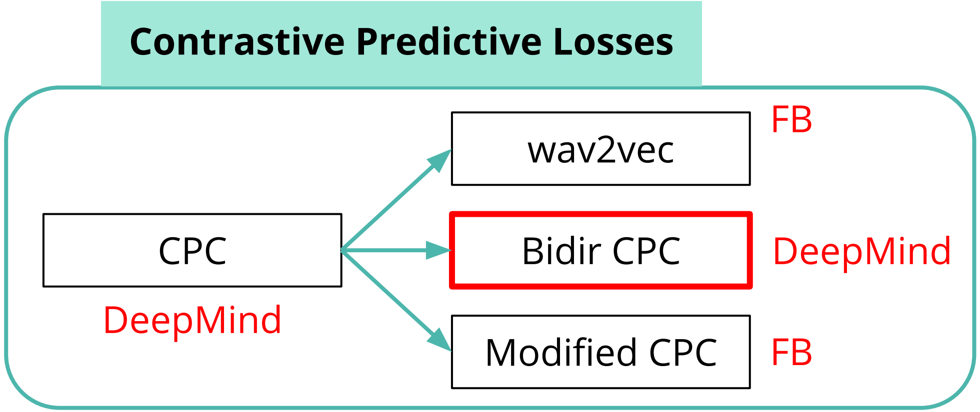
Context Network
9-layer convolutional neural network (CNN)

Encoder Network
5-layer convolutional neural network (CNN)

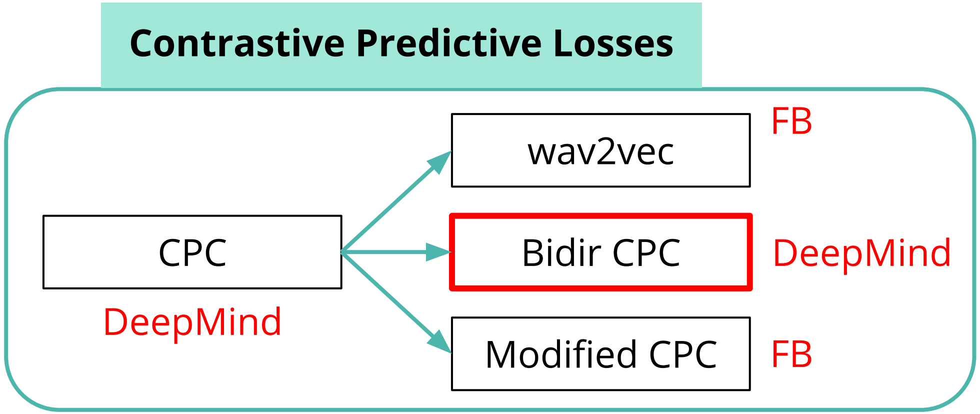


Important exps: ASR on WSJ / TIMIT
*We will compare with this in our work.

Related Works

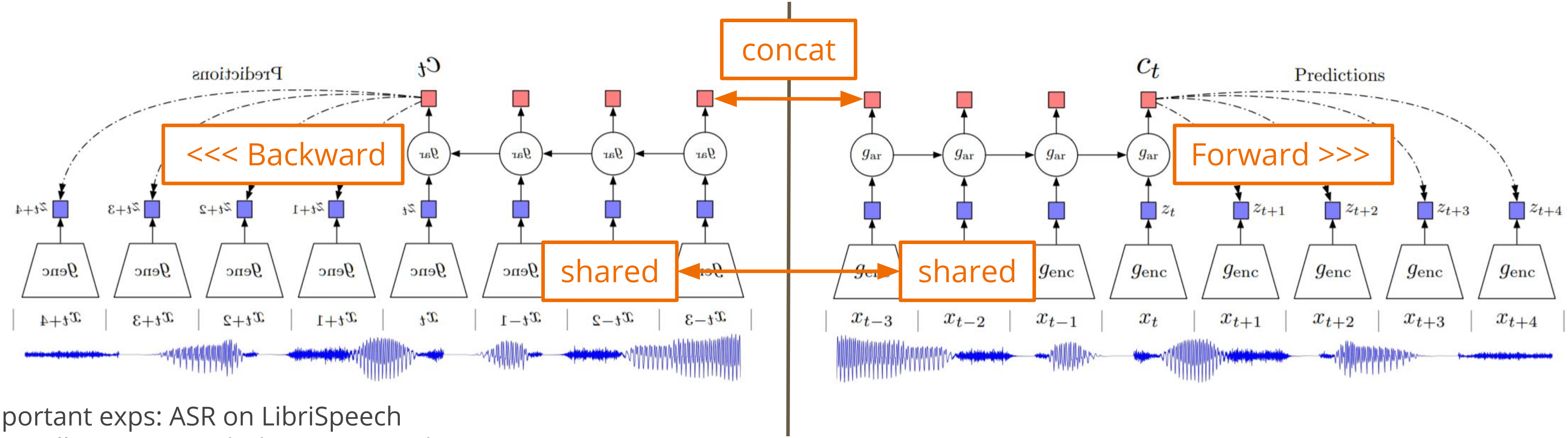


Related Works



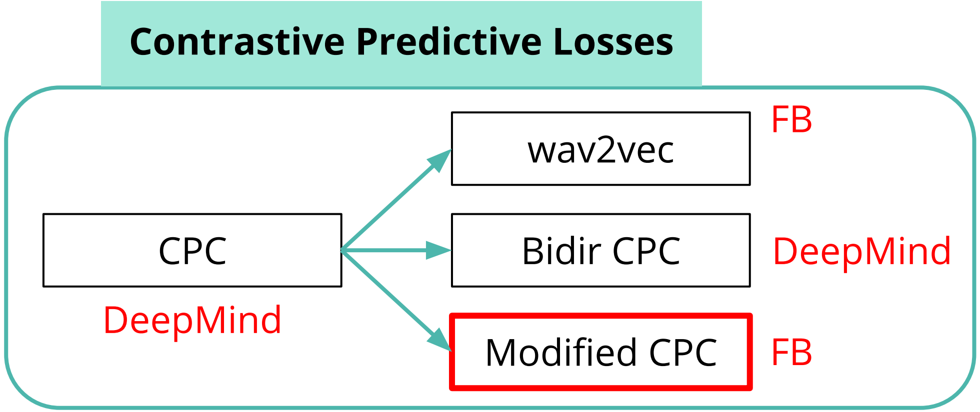
Contribution:
bidirectional context + ASR

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

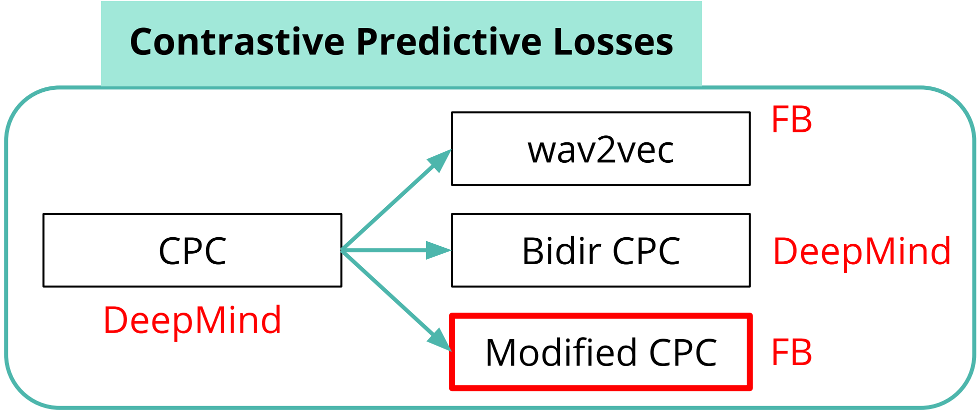


Important exps: ASR on LibriSpeech
*We will compare with this in our work.

Related Works



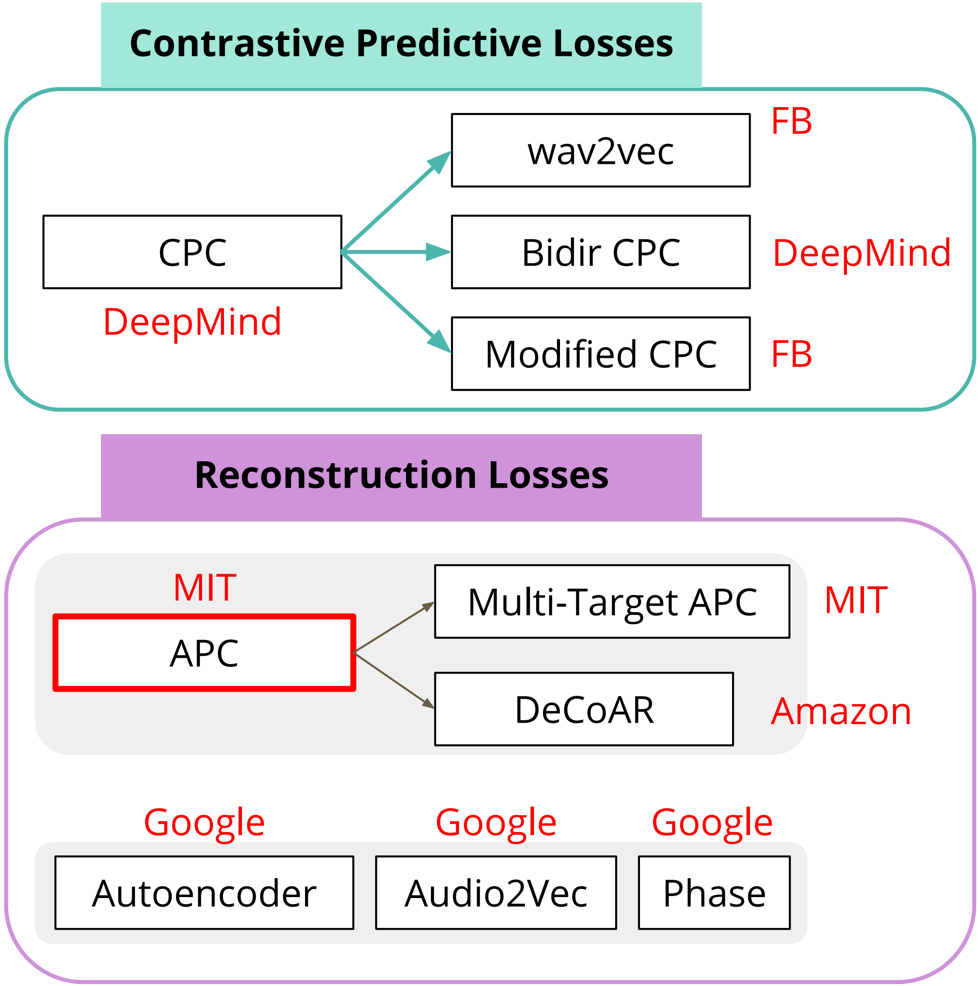
Related Works



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
*We will compare with this in our work.

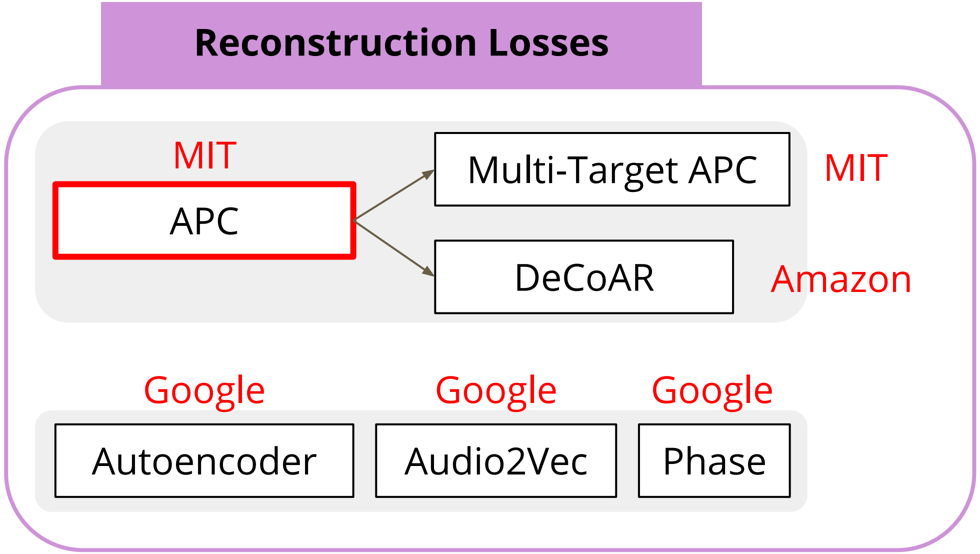
Related Works



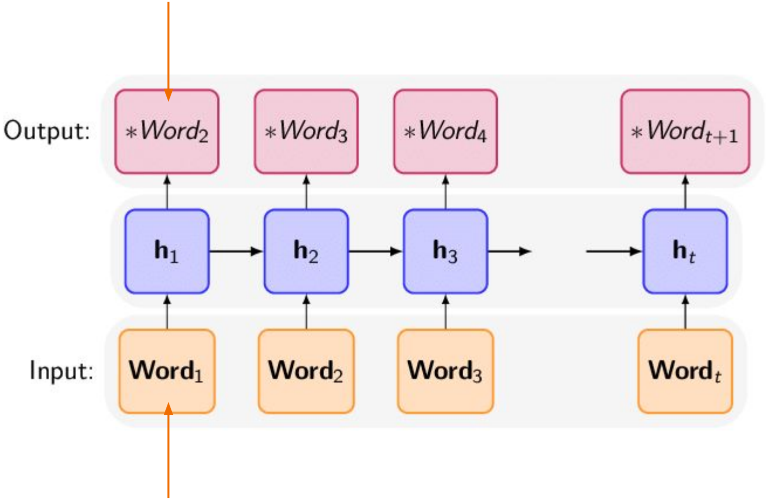
Important exps: Phone / Speaker Classification
*We compared with this in our previous work.

Related Works

Intuition:
Speech Version of a RNN Language Model

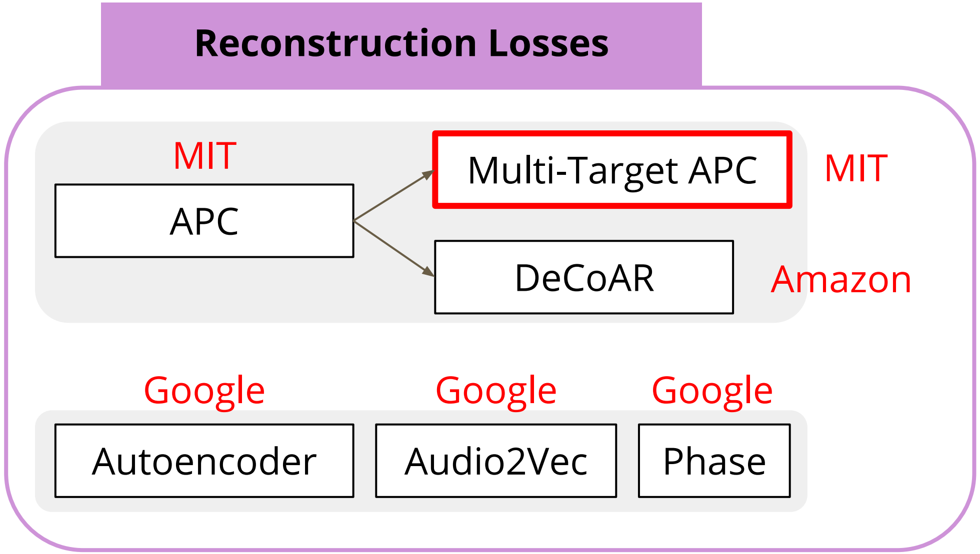


Change the softmax layer to regression layer for reconstruction



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

Related Works



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

Related Works

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.

Reconstruction Losses

MIT
APC

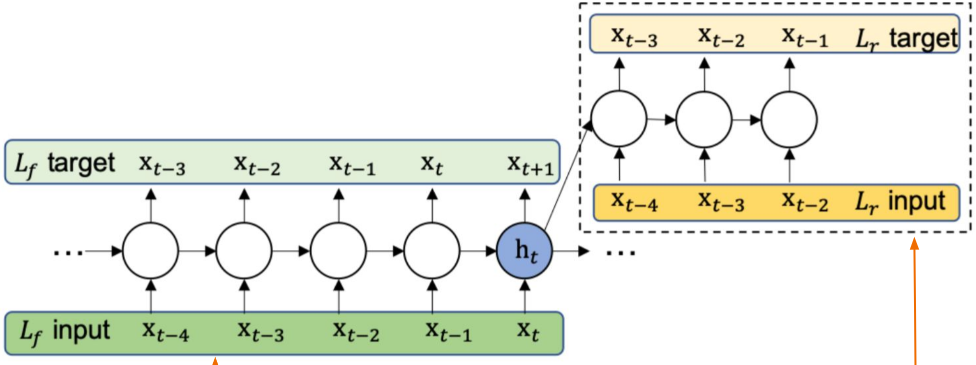
Multi-Target APC — MIT

DeCoAR — Amazon

Autoencoder — Google

Audio2Vec — Google

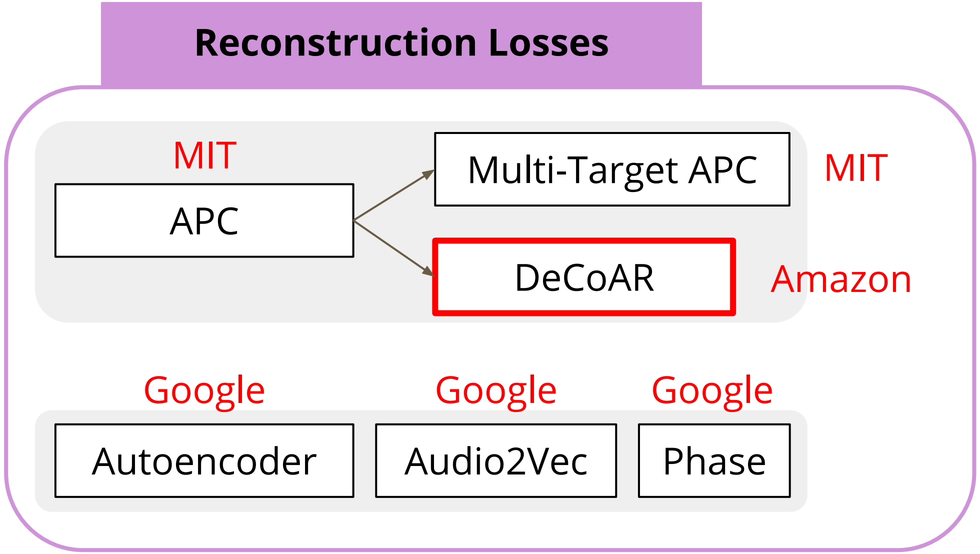
Phase — Google



predicts the future frame conditioning on previous context,

but also predicts the past memory through reconstruction.

Related Works



Important exps: ASR on WSJ / LibriSpeech
*We will compare with this in our work.

Related Works

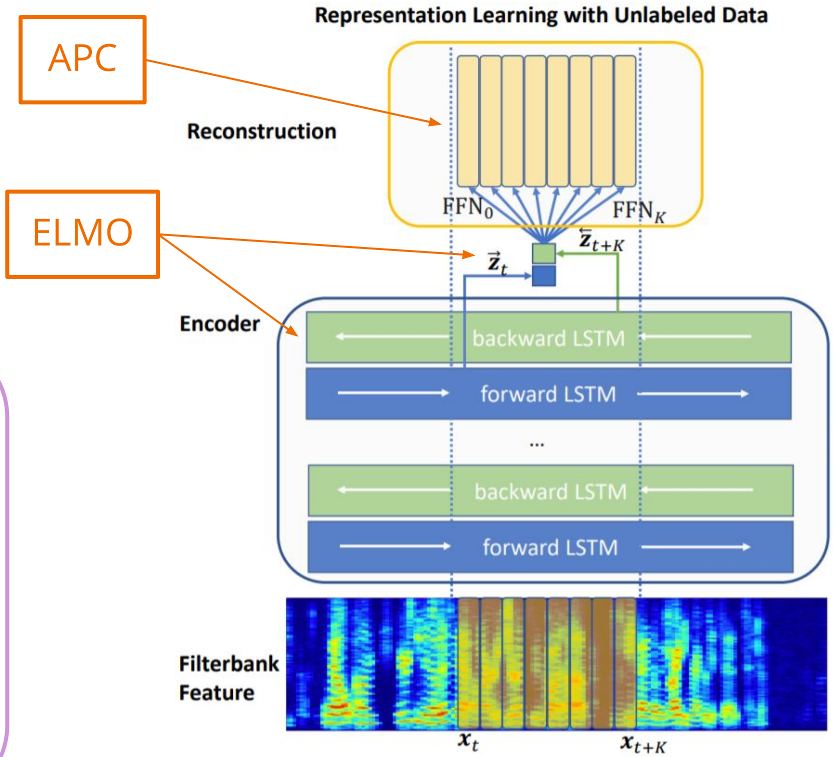
Intuition:
Deep Contextualized Acoustic Representations

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

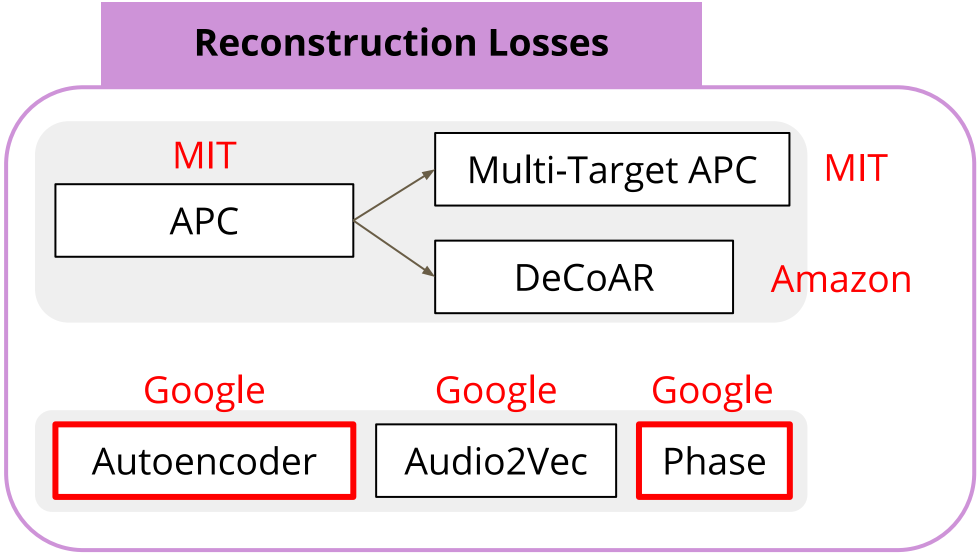
Reconstruction Losses

MIT
APC
Multi-Target APC — MIT
DeCoAR — Amazon

Google: Autoencoder
Google: Audio2Vec
Google: Phase

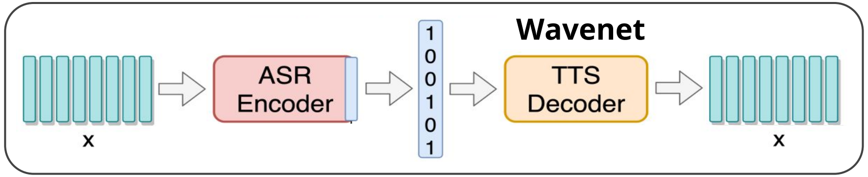


Related Works

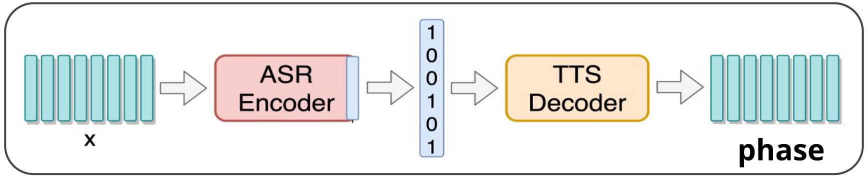


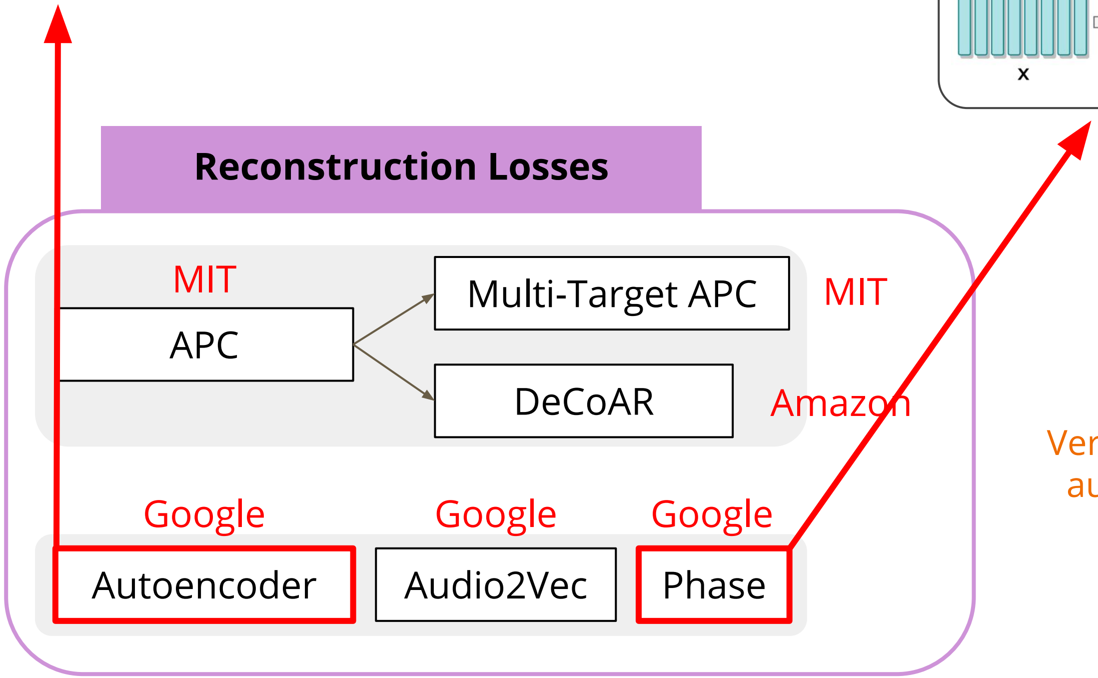
Related Works

Important exps: Phone Classification



Important exps: Linear Classifications

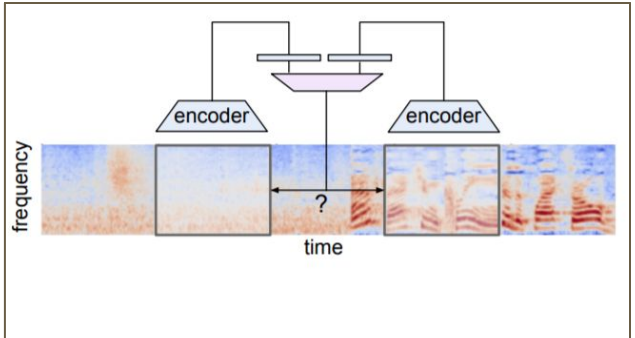




Intuition:

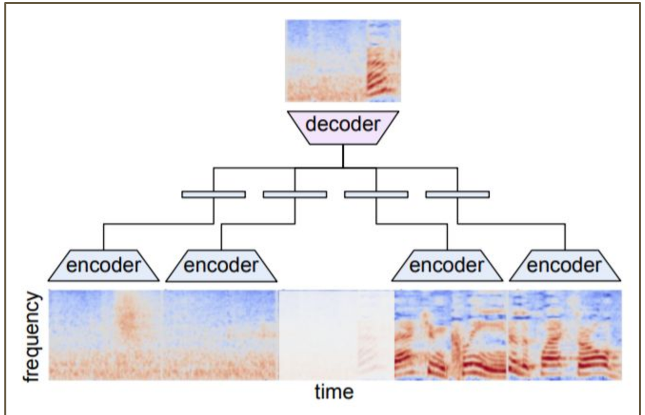
Very similar to the VC structure, learn through autoencoder bottleneck and reconstruction

Important exps: Linear Classifications



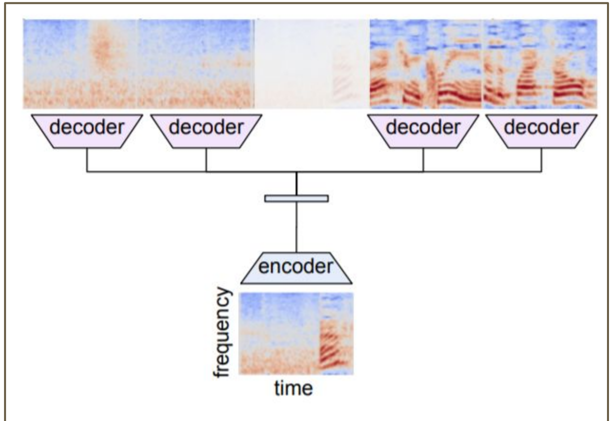
c) Temporal gap

Related Works



a) Audio2Vec - CBoW

Intuition:
Audio version of Word2Vec



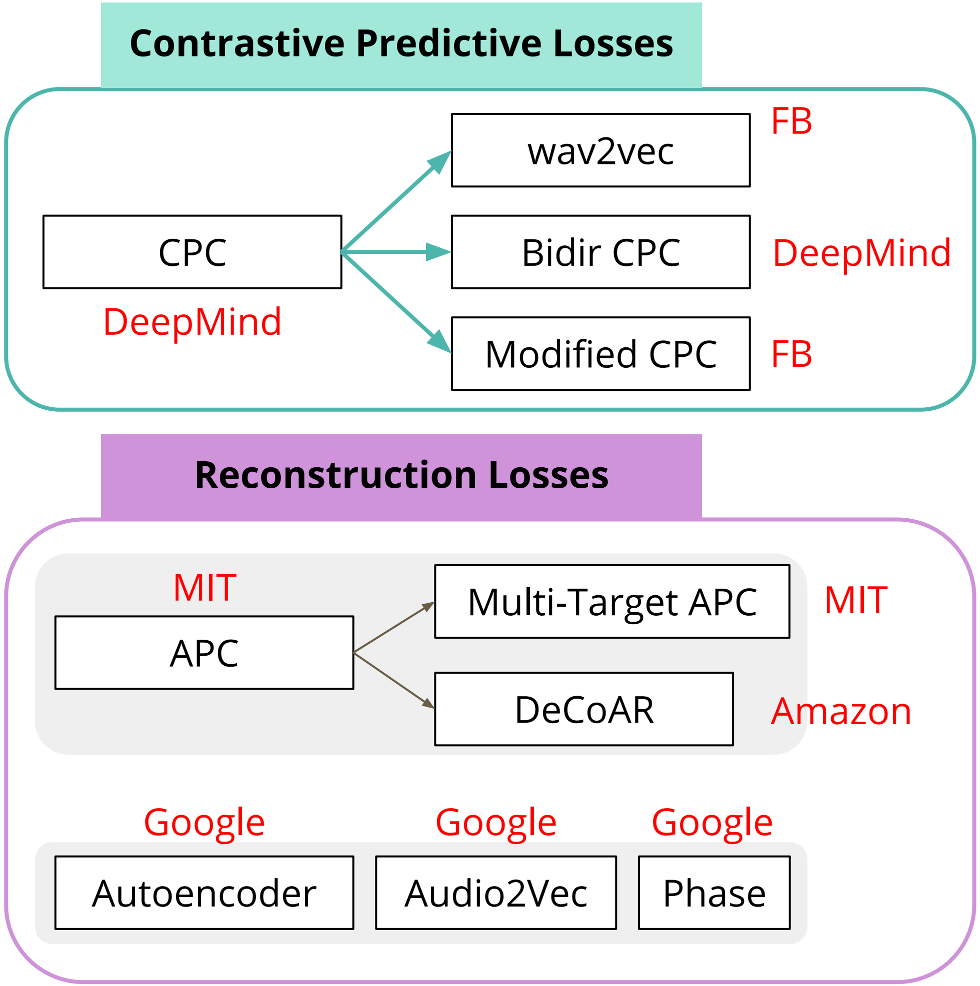
b) Audio2Vec - skip-gram

Reconstruction Losses

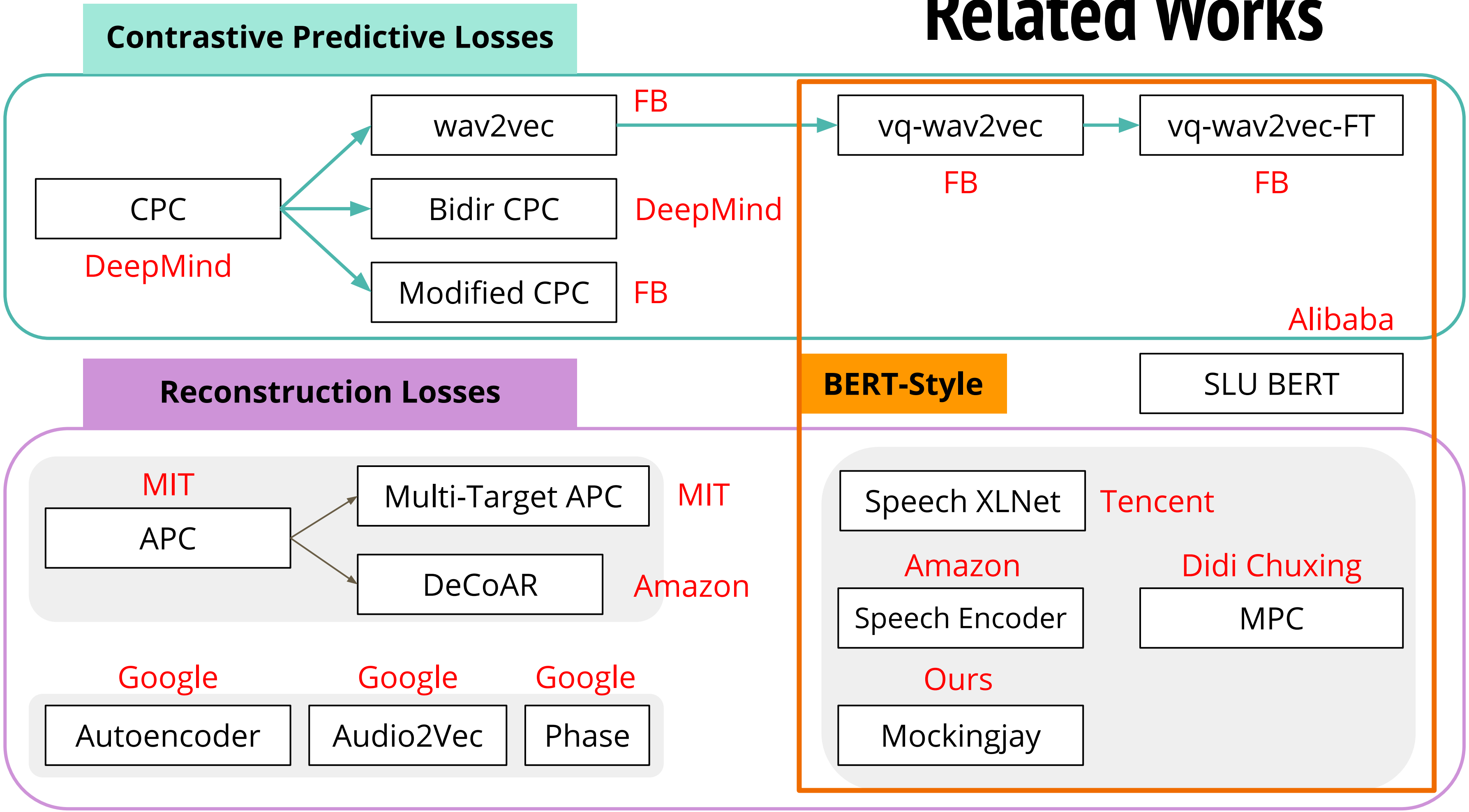
MIT
APC → Multi-Target APC (MIT)
APC → DeCoAR (Amazon)

Google: Autoencoder
Google: Audio2Vec
Google: Phase

Related Works



Related Works



Related Works

vq-wav2vec → vq-wav2vec-FT

FB FB

Alibaba

BERT-Style

SLU BERT

Speech XLNet Tencent

Amazon

Speech Encoder

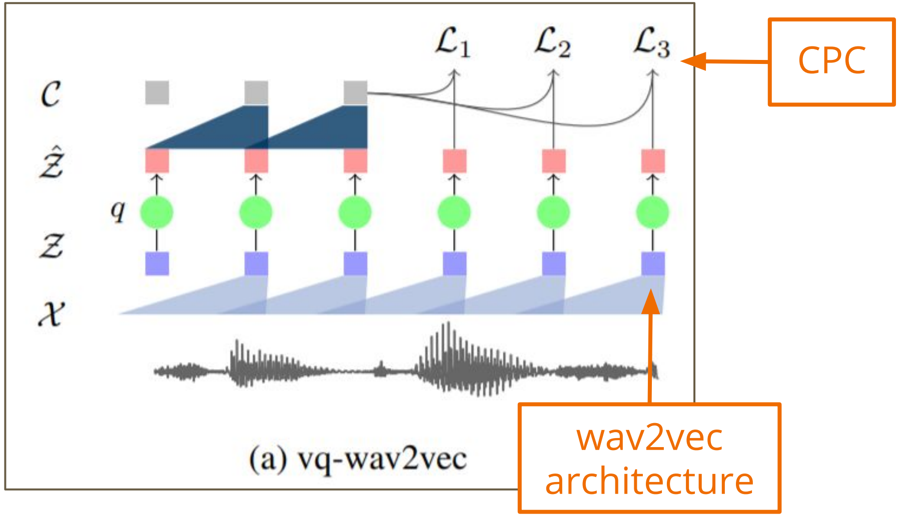
Didi Chuxing

MPC

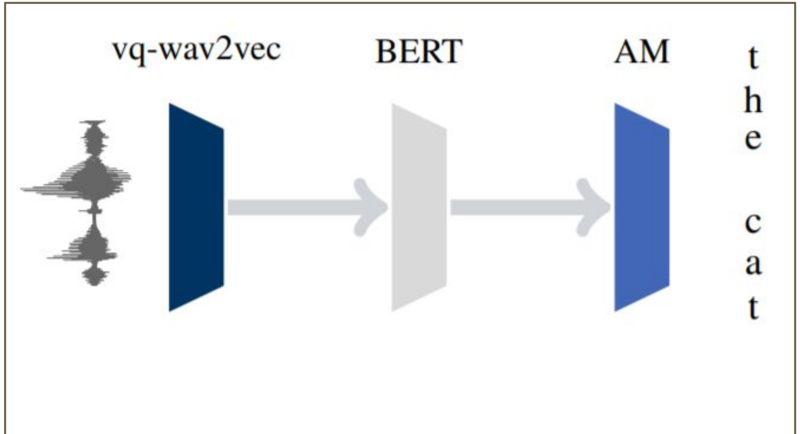
Ours

Mockingjay

Related Works



(a) vq-wav2vec



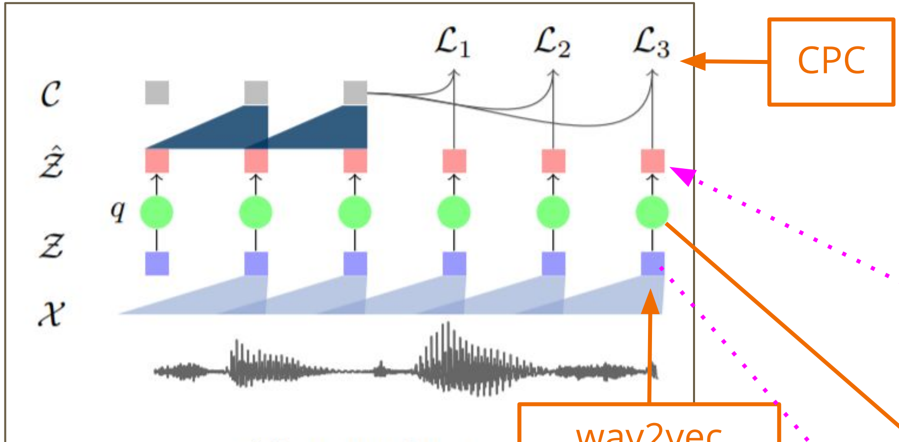
(b) Discretized speech training pipeline

vq-wav2vec → vq-wav2vec-FT

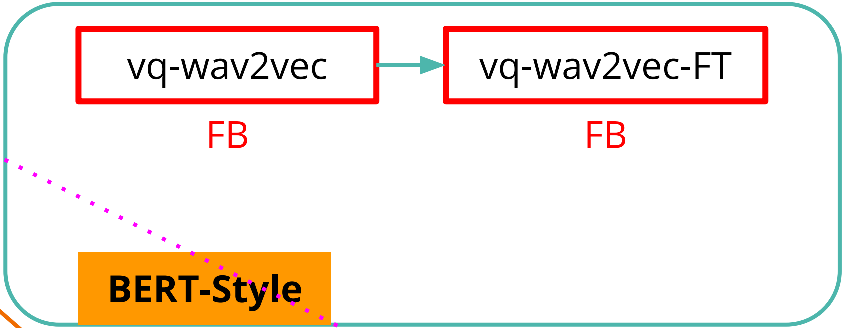
FB FB

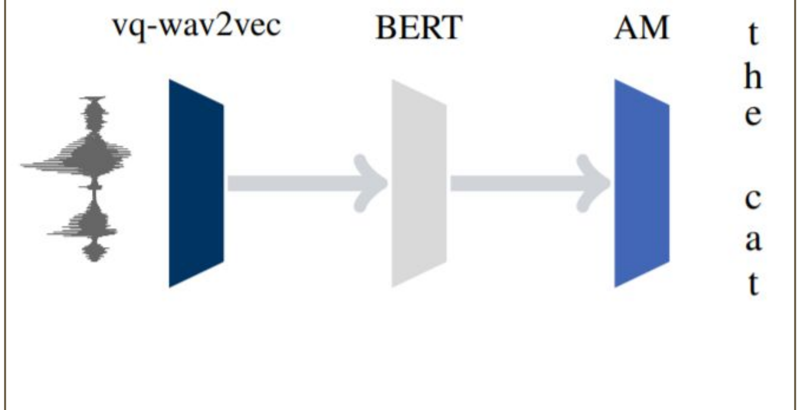
BERT-Style

Related Works

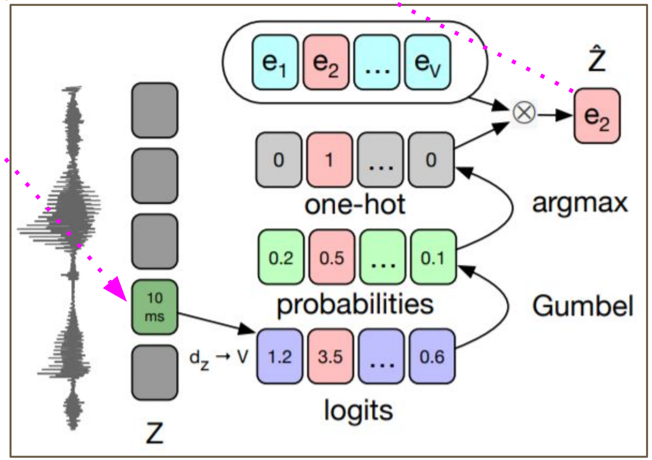


(a) vq-wav2vec

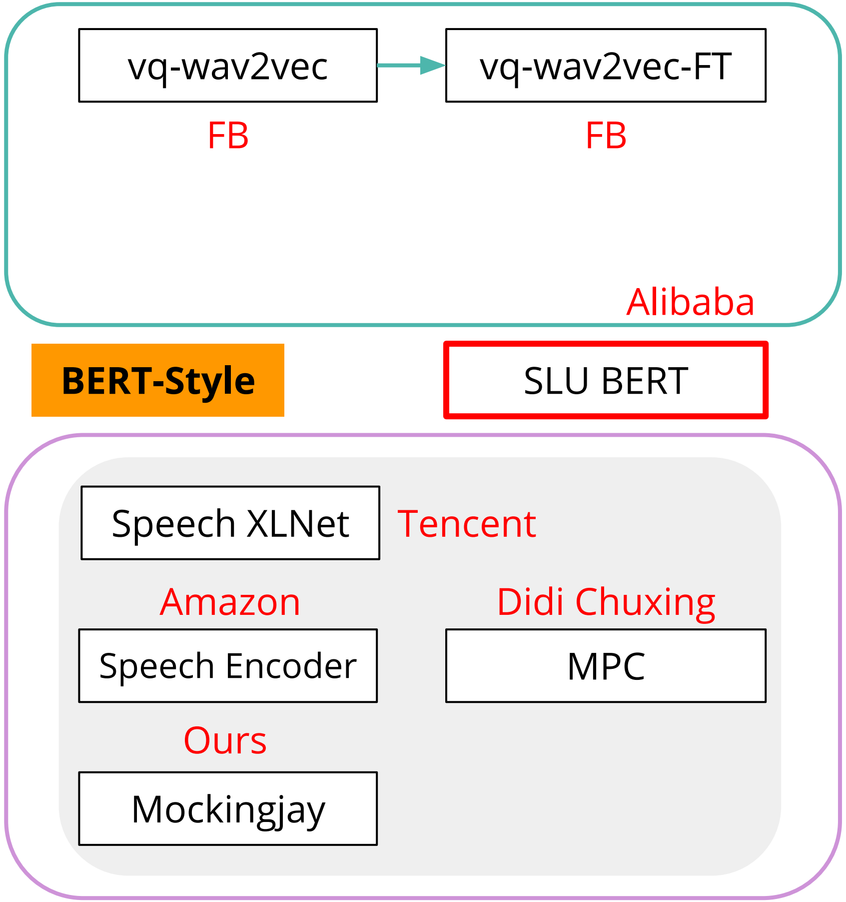




(b) Discretized speech training pipeline



Related Works



Related Works

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

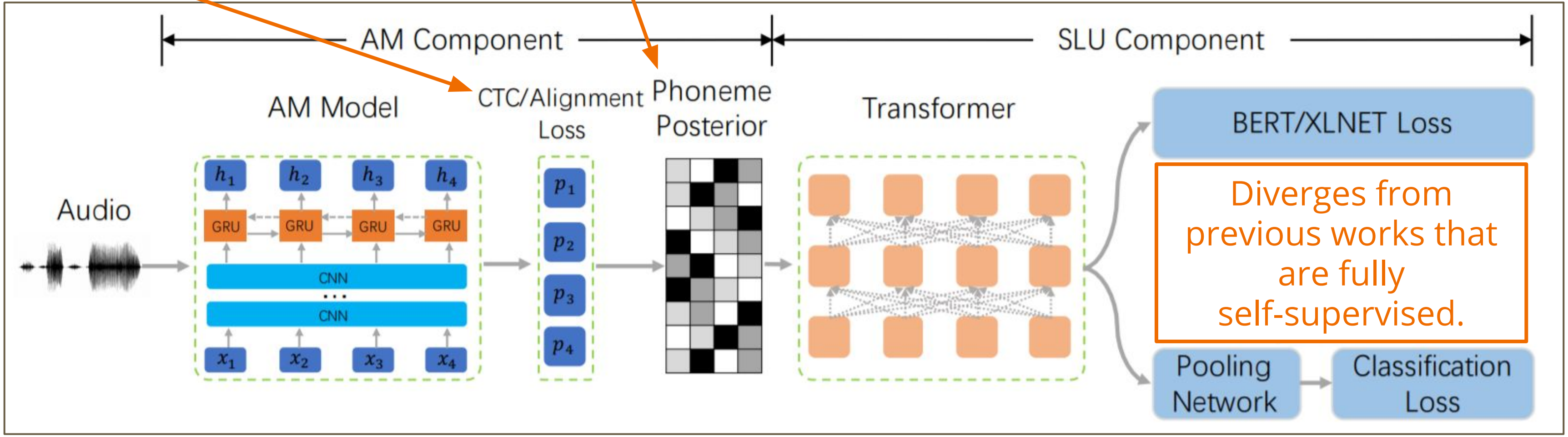
vq-wav2vec → vq-wav2vec-FT

FB FB

Alibaba

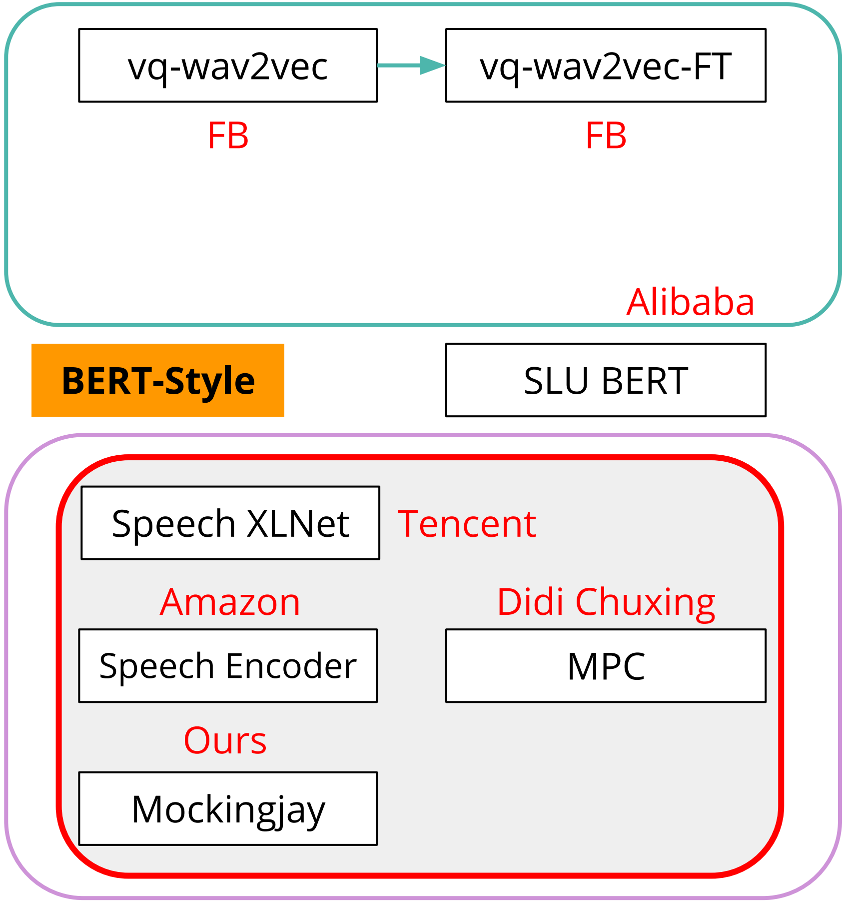
BERT-Style

SLU BERT



Diverges from previous works that are fully self-supervised.

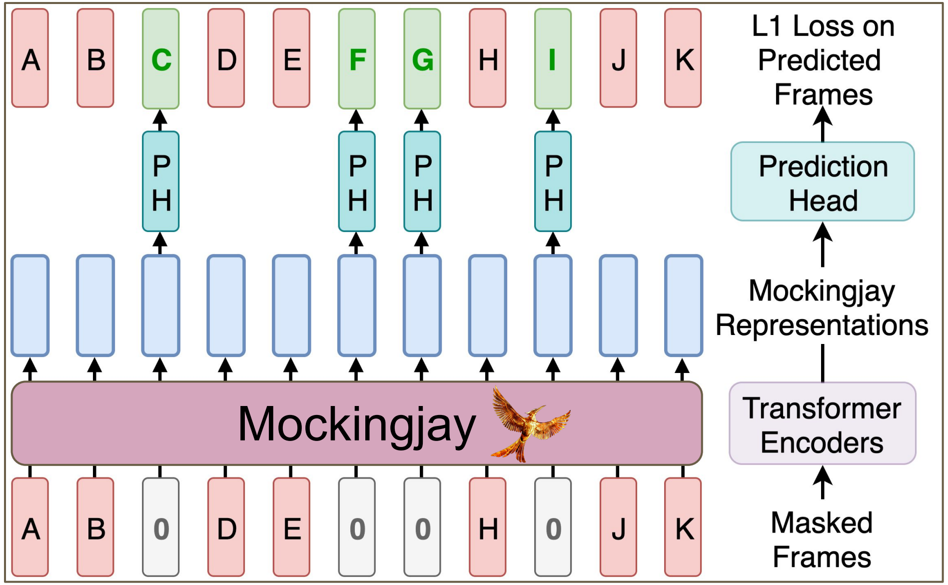
Related Works

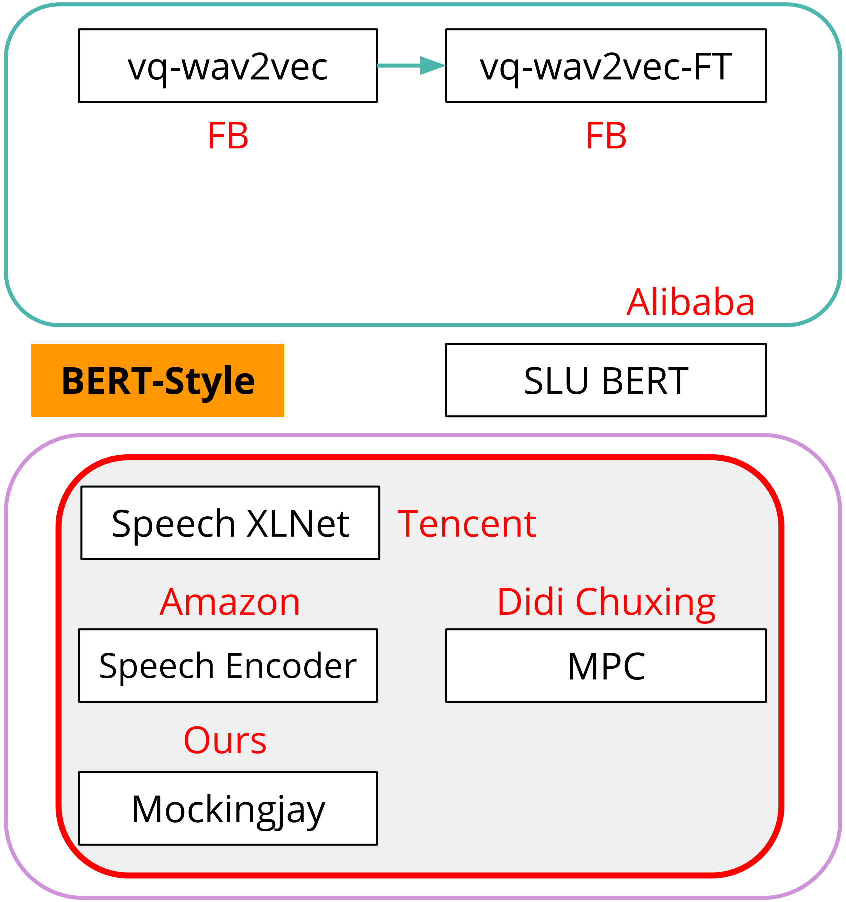


The Trend:
All of these works emerges around October, 2019.
All submitted to ICASSP 2020

(Speech XLNet and MPC did not make it)

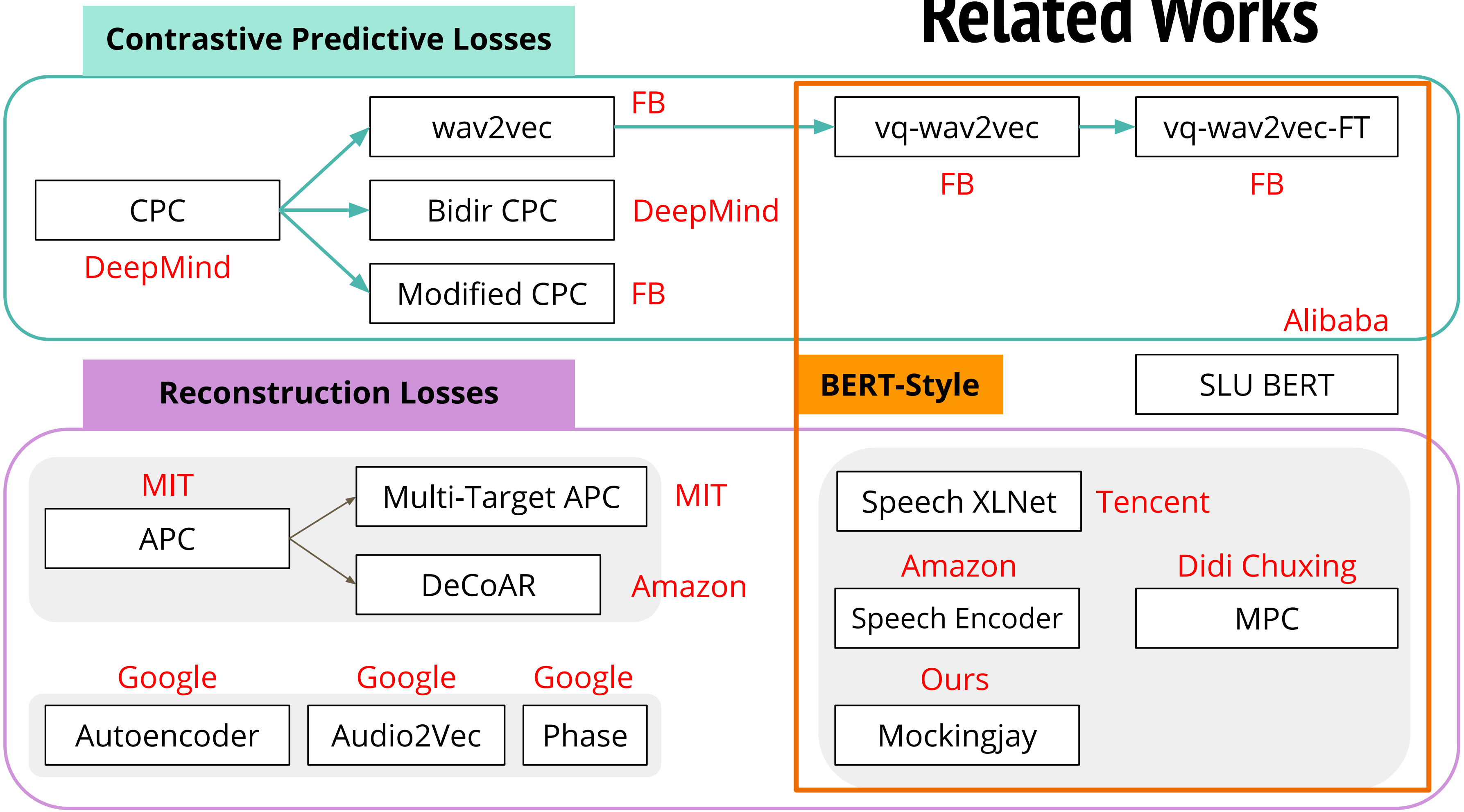
Related Works



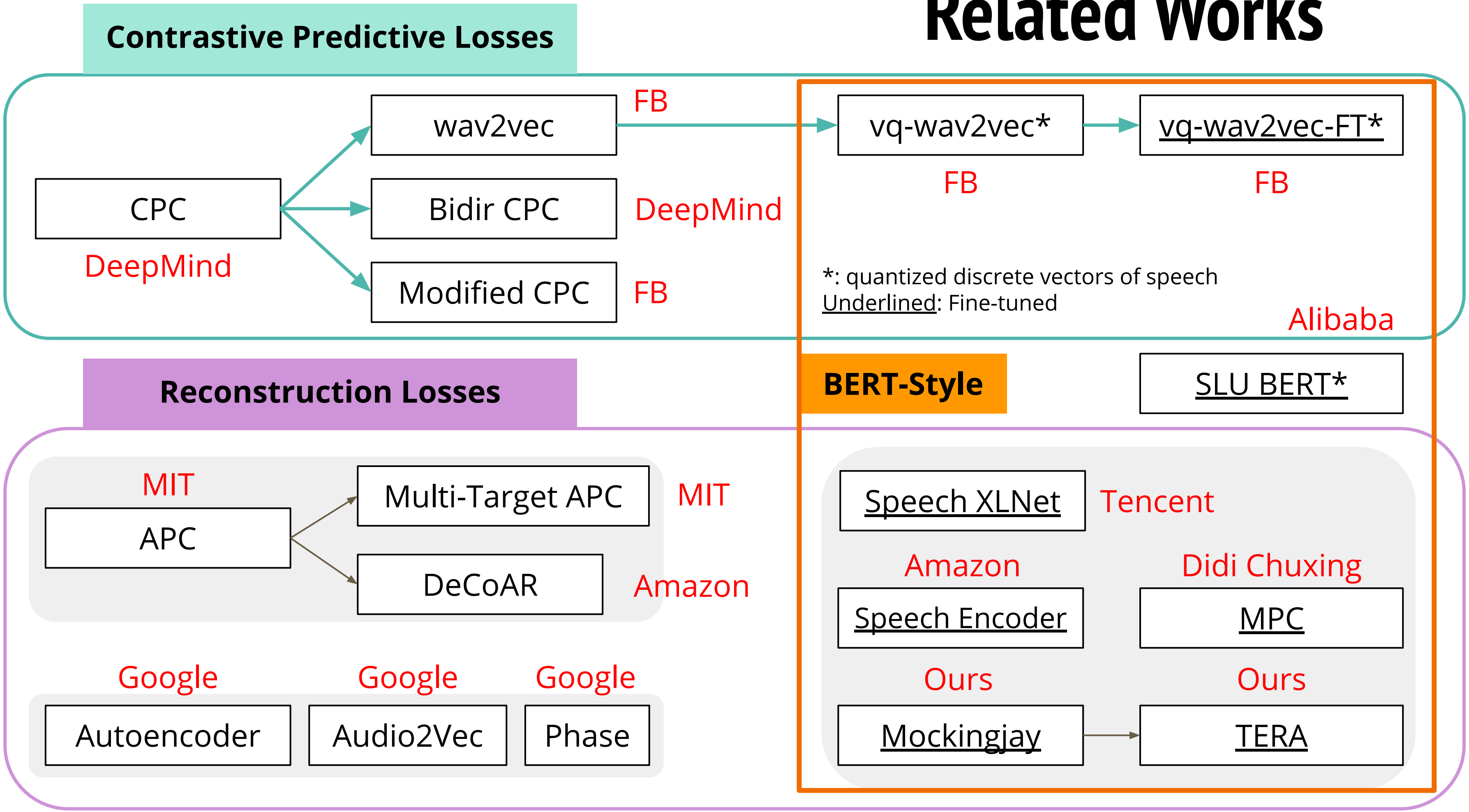


Intuition:
A model that can predict the partial loss
of small segments of speech,
should provide a contextualized
understanding of previous and later content.

Related Works

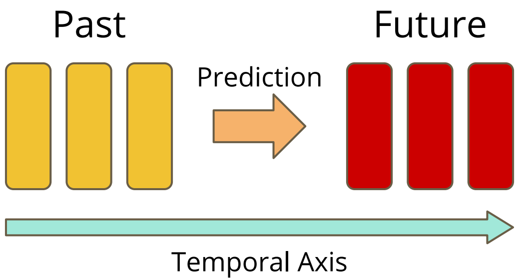


Related Works

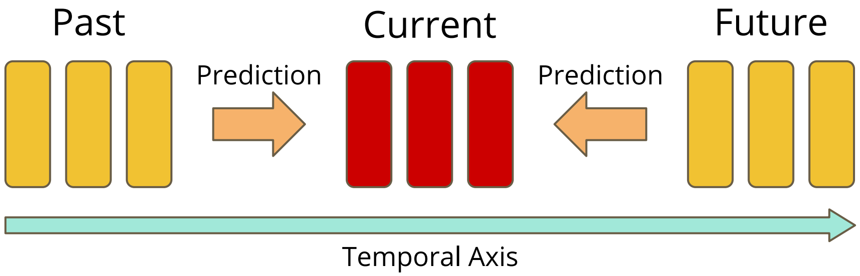


Related Work Summary

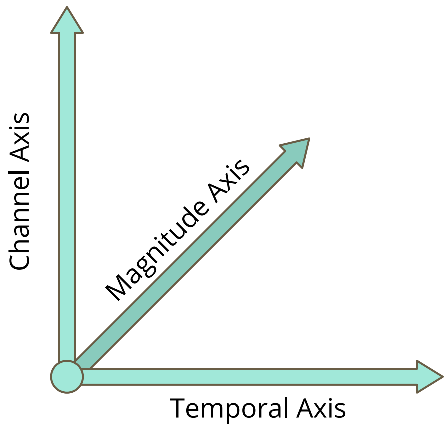
The design of auxiliary task fundamentally defines what the model learns!



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

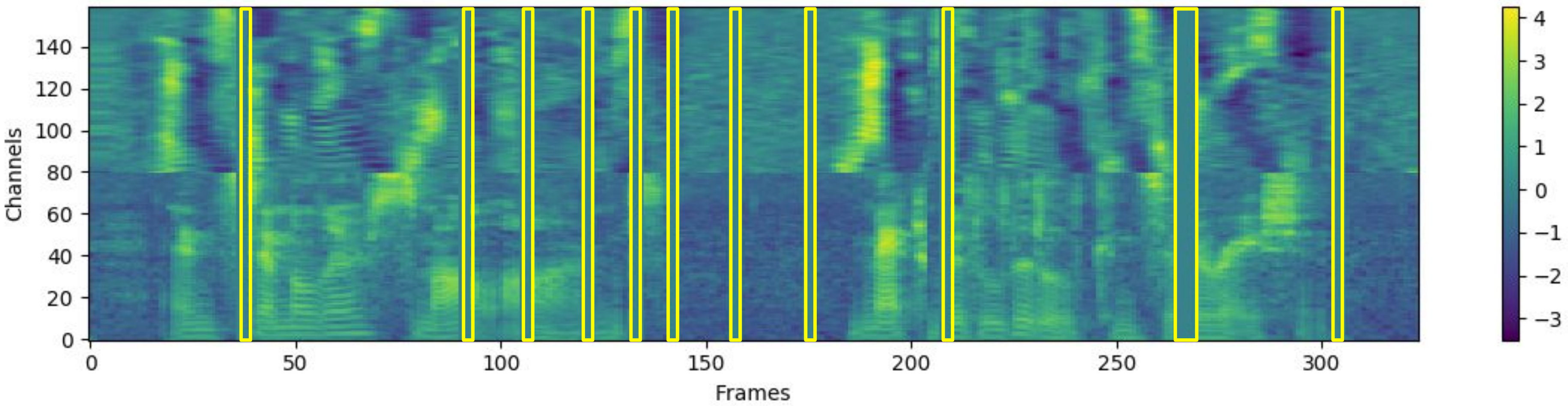


TERA

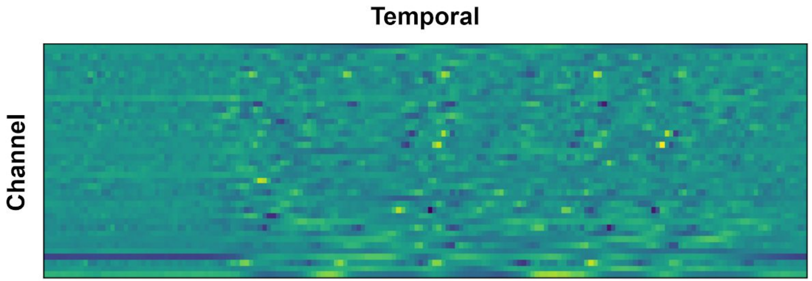
Transformer Encoder Representations from Alteration

Extending Mockingjay to multi-target learning on three dimensions

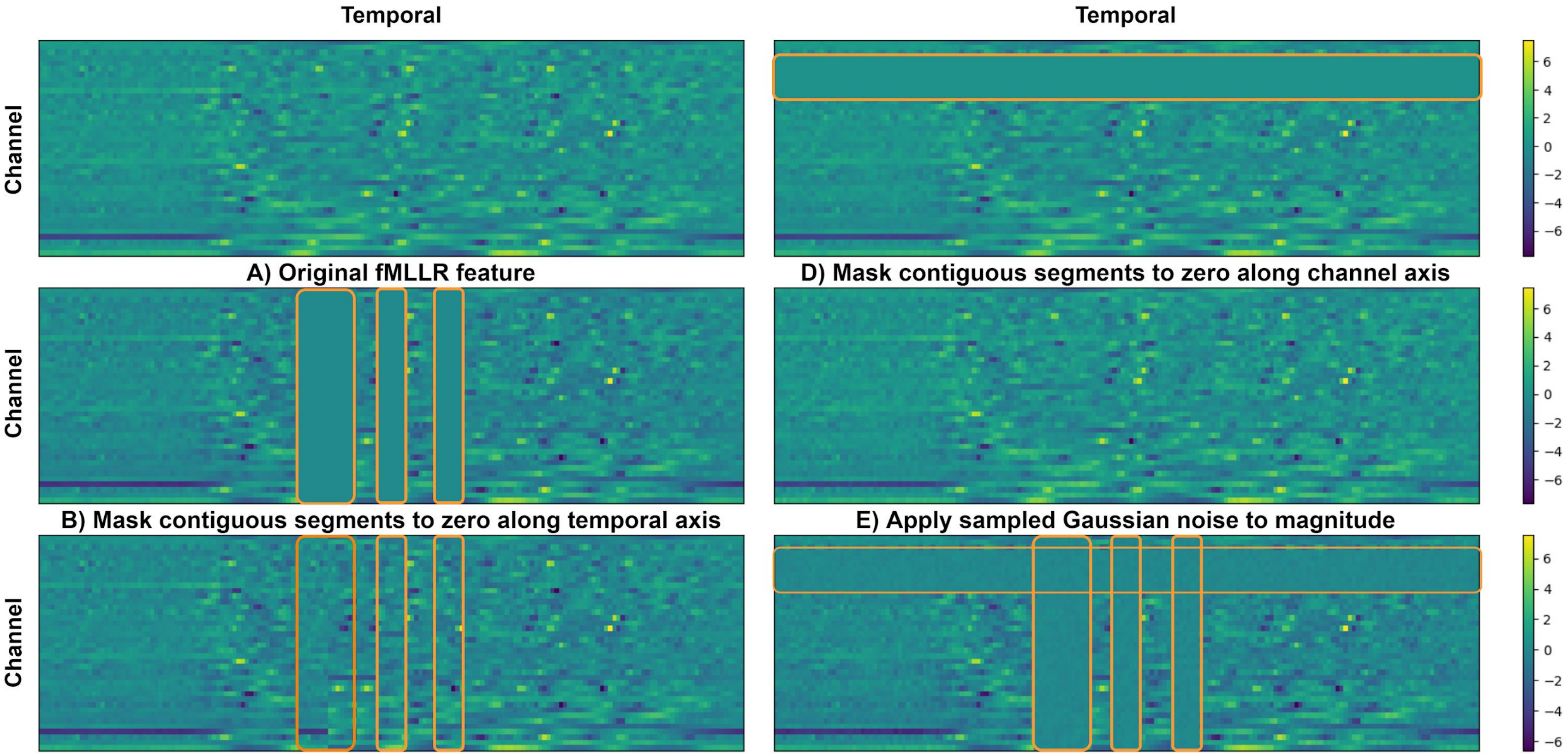
Recall: we mask mel spectrogram on time axis



Consider fMLLR on 3 Axis:



Multi-target Pre-training



A) Original fMLLR feature

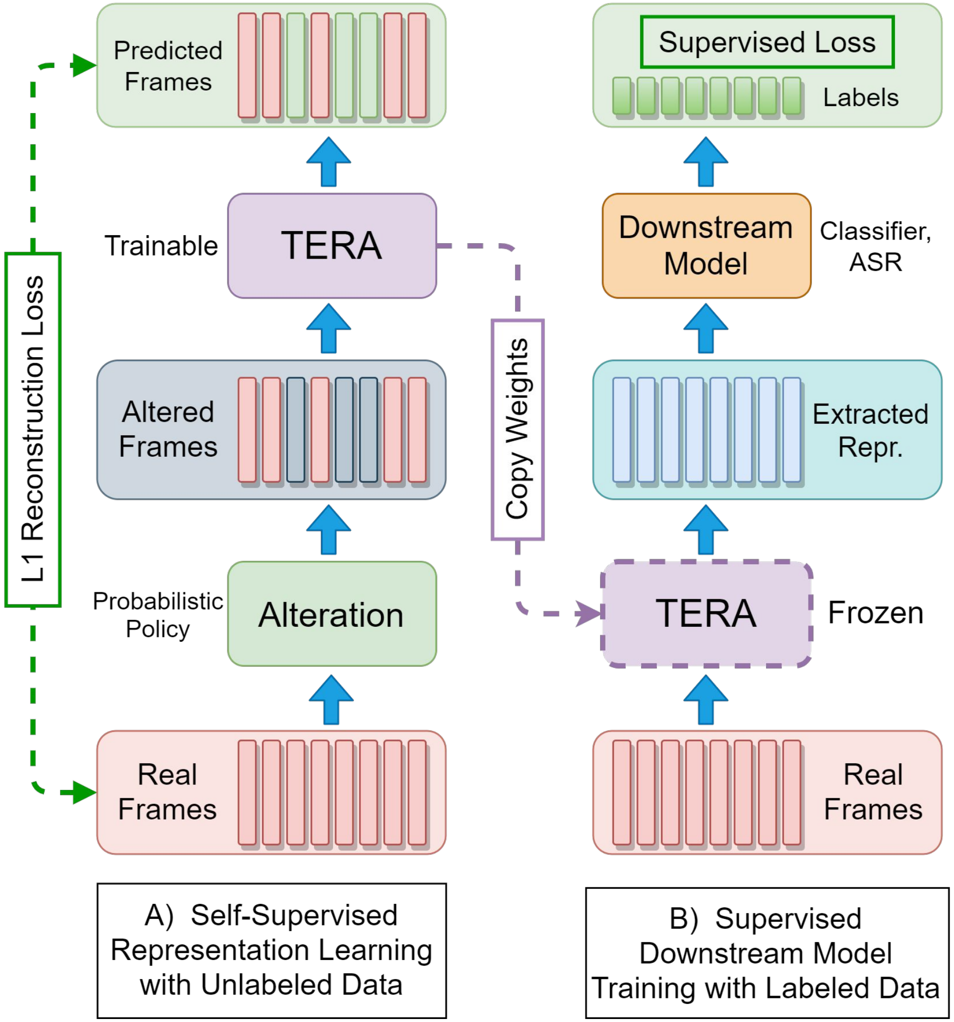
D) Mask contiguous segments to zero along channel axis

B) Mask contiguous segments to zero along temporal axis

E) Apply sampled Gaussian noise to magnitude

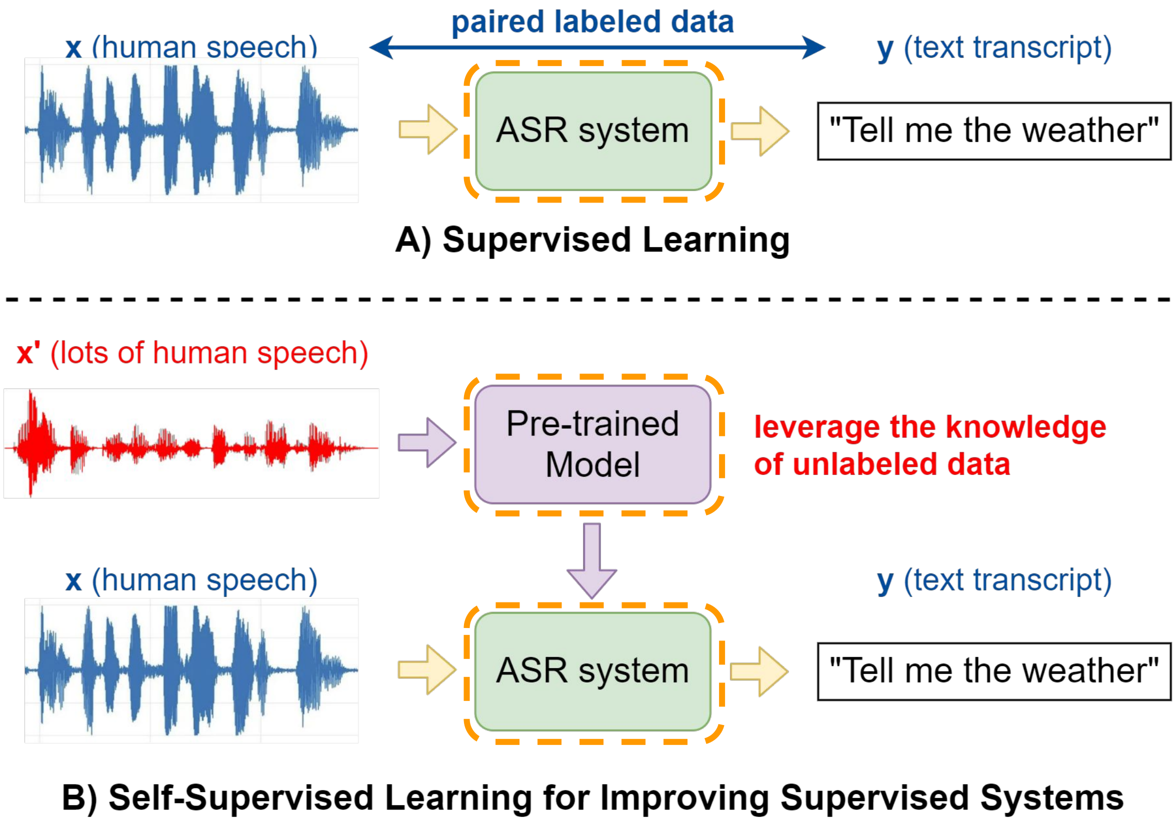
C) Replace contiguous segments with random segments

F) Combining the alterations in B), D), and E)

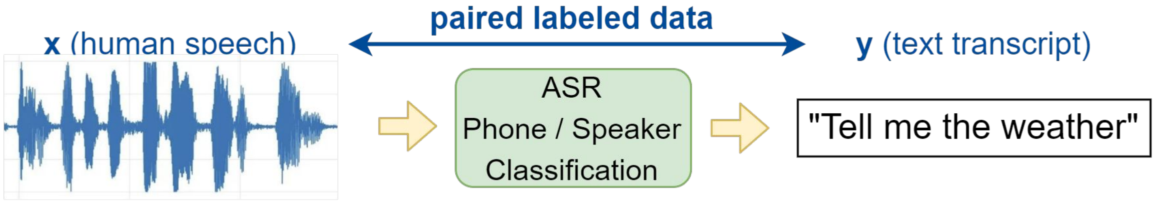


Predicted Frames
Trainable
TERA
Altered Frames
Probabilistic Policy
Alteration
Real Frames
L1 Reconstruction Loss
Copy Weights
Supervised Loss
Labels
Downstream Model
Classifier, ASR
Extracted Repr.
TERA
Frozen
Real Frames
A) Self-Supervised Representation Learning with Unlabeled Data
B) Supervised Downstream Model Training with Labeled Data

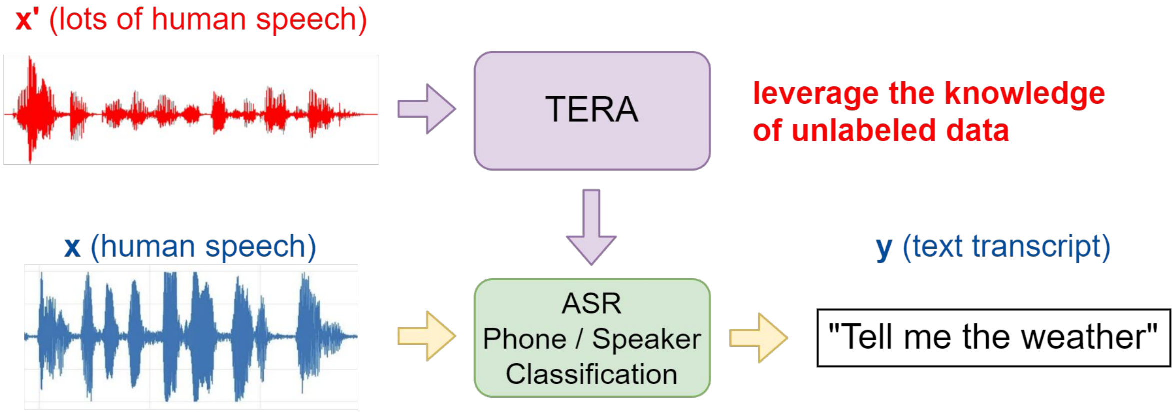
Recall: Self-Supervised Learning for Speech



Self-Supervised Learning: TERA



A) Supervised Learning

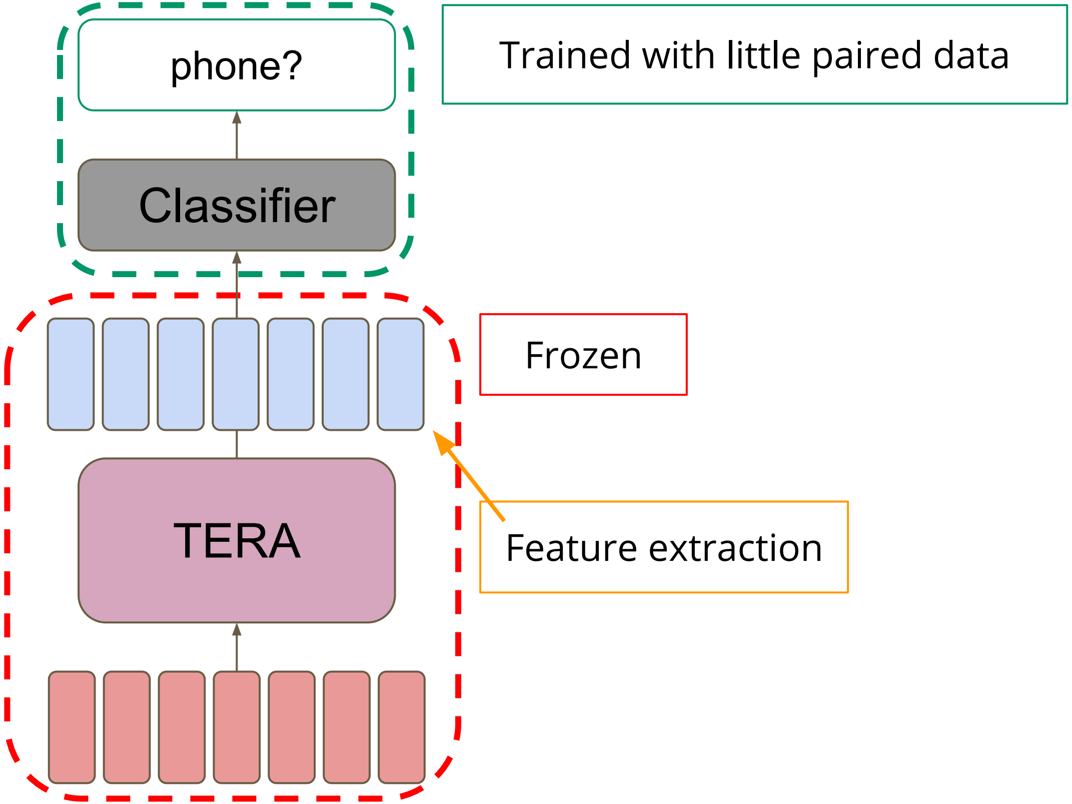


B) Self-Supervised Learning for Improving Supervised Systems

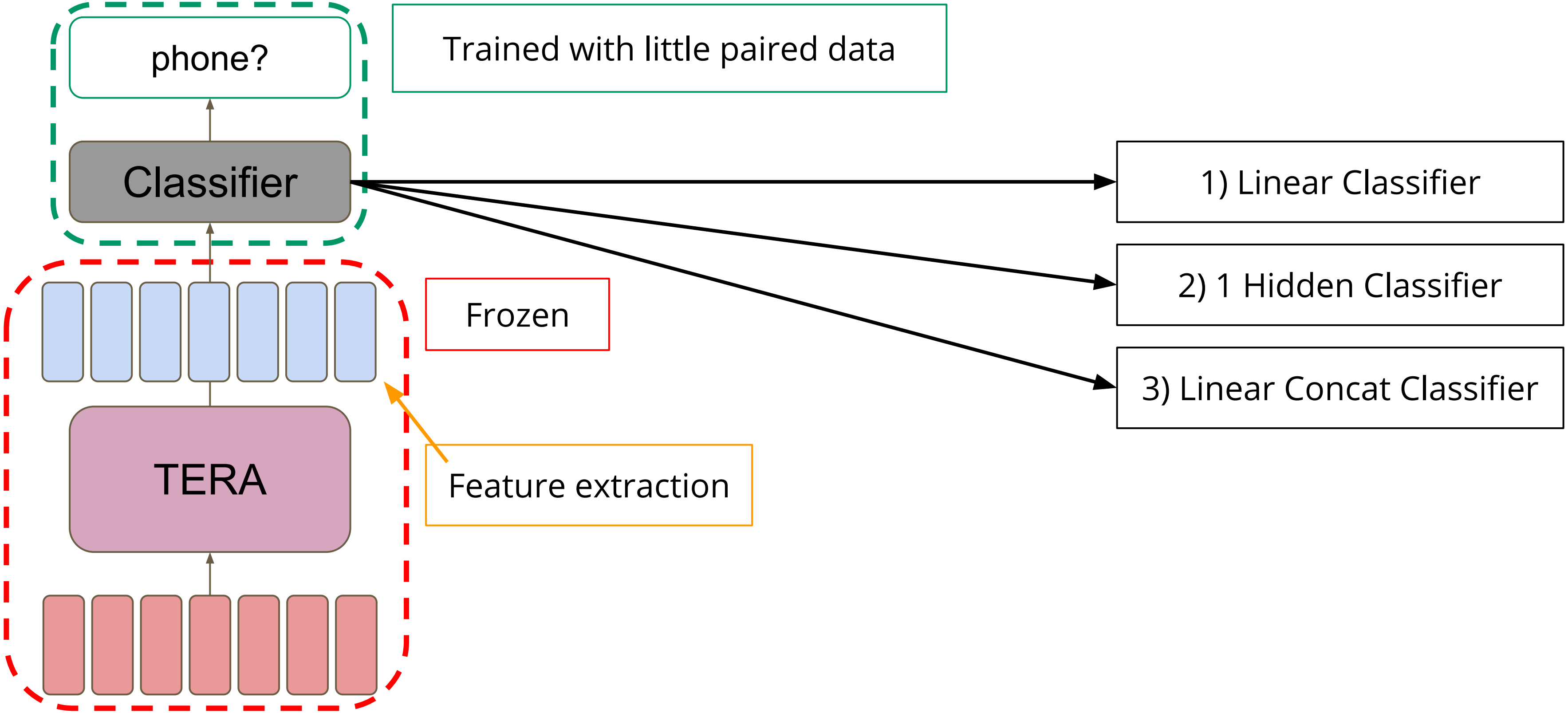
Progress: 10 pages, 90% Done

Submitting to the IEEE/ACM Journal of TASL around summer

Frame-wise phone classification on LibriSpeech



Frame-wise phone classification on LibriSpeech



Frame-wise phone classification on LibriSpeech

Representation \ Pre-train	100 hr	
	Linear	1 Hidden
CPC [8]	64.6	72.5
TERA-base: time	64.3	76.8
TERA-base: time + mag	64.1	77.1
TERA-base: time + channel	**65.2**	**77.4**
TERA-base: time + channel + mag	65.1	77.3
MFCC	39.7	59.9
fMLLR	52.6	68.4

Frame-wise phone classification on LibriSpeech

Representation \ Pre-train	100 hr Linear	100 hr 1 Hidden
CPC [8]	64.6	72.5
TERA-base: time	64.3	76.8
TERA-base: time + mag	64.1	77.1
TERA-base: time + channel	**65.2**	**77.4**
TERA-base: time + channel + mag	65.1	77.3
MFCC	39.7	59.9
fMLLR	52.6	68.4

Outperformed CPC

Using more objectives also improves performance!

Baseline feature was outperformed by TERA features.

Frame-wise phone classification on LibriSpeech

Representation \ Pre-train	100 hr		460 hr		960 hr	
	Linear	1 Hidden	Linear	1 Hidden	Linear	1 Hidden
CPC [8]	64.6	72.5	-			
TERA-base: time	64.3	76.8	64.4	77.0	**67.0**	**79.1**
TERA-base: time + mag	64.1	77.1	64.5	77.3	64.7	77.8
TERA-base: time + channel	**65.2**	**77.4**	66.0	78.1	65.9	78.5
TERA-base: time + channel + mag	65.1	77.3	**66.3**	**78.3**	66.4	78.9
MFCC	39.7	59.9	← traditional features with linear classifier			
fMLLR	52.6	68.4				

More pre-training data increases performance.

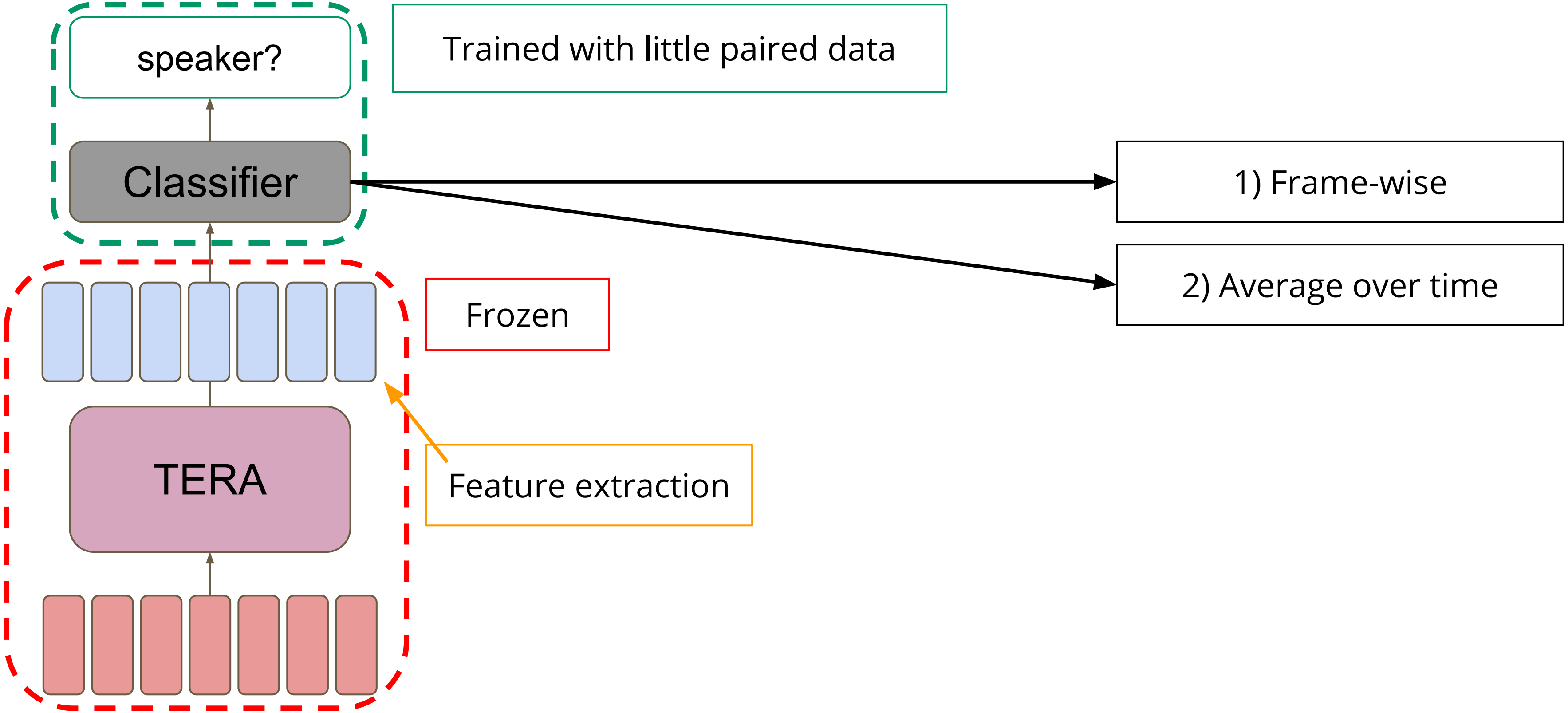
Having more alteration is like having more data.

Frame-wise phone classification on LibriSpeech

Representation \ Pre-train	100 hr		460 hr		960 hr	
	Linear	1 Hidden	Linear	1 Hidden	Linear	1 Hidden
CPC [8]	64.6	72.5		-		
TERA-base: time	64.3	76.8	64.4	77.0	**67.0**	**79.1**
TERA-base: time + mag	64.1	77.1	64.5	77.3	64.7	77.8
TERA-base: time + channel	**65.2**	**77.4**	66.0	78.1	65.9	78.5
TERA-base: time + channel + mag	65.1	77.3	**66.3**	**78.3**	66.4	78.9
MFCC	39.7	59.9	← traditional features with linear classifier			
fMLLR	52.6	68.4				

When real data is limited (<= 460 hr): more alterations are helpful.
When real data is vast (960 hr): augmentation is not required, but comparable.

Speaker linear classification on LibriSpeech



Speaker linear classification on LibriSpeech

Representation \ Pre-train	100 hr	
	Frame	Average
CPC [8]	97.4	-
TERA-base: time	68.4	96.1
TERA-base: time + mag	70.8	96.1
TERA-base: time + channel	93.6	98.5
TERA-base: time + channel + mag	**98.9**	**99.2**
MFCC	17.6	10.8
fMLLR	0.4	2.6

Baseline feature fails to encode speaker information.

Speaker linear classification on LibriSpeech

Representation \ Pre-train	100 hr	
	Frame	Average
CPC [8]	97.4	-
TERA-base: time	68.4	96.1
TERA-base: time + mag	70.8	96.1
TERA-base: time + channel	93.6	98.5
TERA-base: time + channel + mag	**98.9**	**99.2**
MFCC	17.6	10.8
fMLLR	0.4	2.6

Outperformed CPC

Using more objectives also improves performance!

Although we train on fMLLR, we recover the speaker information through the proposed objectives.

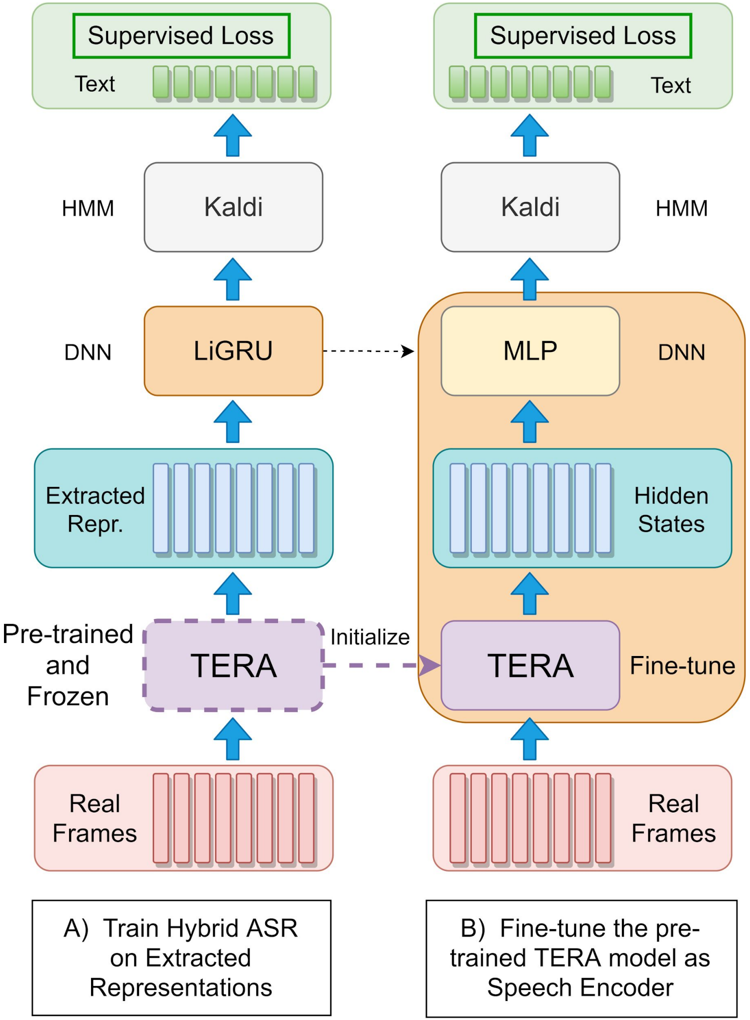
Speaker linear classification on LibriSpeech

Representation \ Pre-train	100 hr		460 hr		960 hr	
	Frame	Average	Frame	Average	Frame	Average
CPC [8]	97.4	-	-			
TERA-base: time	68.4	96.1	86.9	97.3	99.3	99.7
TERA-base: time + mag	70.8	96.1	88.0	98.0	89.2	98.8
TERA-base: time + channel	93.6	98.5	**99.4**	**99.5**	**99.5**	**99.8**
TERA-base: time + channel + mag	**98.9**	**99.2**	99.0	99.5	99.4	99.8
MFCC	17.6	10.8	← traditional features with linear classifier			
fMLLR	0.4	2.6				

Pre-training on more data also gives benefit!

ASR Framework

Hybrid DNN/HMM ASR



Complete

In Progress

Mixtures of auxiliary objectives - ASR

Models \ Pre-train	100 hr	
	WER	Rescore
liGRU + TERA-base: time	8.46	6.12
liGRU + TERA-base: time + mag	8.43	6.11
liGRU + TERA-base: time + channel	8.57	6.16
liGRU + TERA-base: time + channel + mag	**8.32**	**6.01**

More augmentations are helpful

Effect of Amount of Pre-training Data - ASR 1/4

Pre-train / Models	100 hr		460 hr		960 hr	
	WER	Rescore	WER	Rescore	WER	Rescore
liGRU + TERA-base: time	8.46	6.12	8.38	6.10	**8.31**	**5.99**
liGRU + TERA-base: time + mag	8.43	6.11	8.38	6.04	8.40	6.03
liGRU + TERA-base: time + channel	8.57	6.16	8.49	6.08	8.35	6.07
liGRU + TERA-base: time + channel + mag	**8.32**	**6.01**	**8.29**	**6.00**	**8.31**	6.01

More pre-training data increases performance,
Consistent WER drop for the red rows

Effect of Amount of Pre-training Data - ASR 2/4

Models \ Pre-train	100 hr		460 hr		960 hr	
	WER	Rescore	WER	Rescore	WER	Rescore
liGRU + TERA-base: time	8.46	6.12	8.38	6.10	**8.31**	**5.99**
liGRU + TERA-base: time + mag	8.43	6.11	8.38	6.04	8.40	6.03
liGRU + TERA-base: time + channel	8.57	6.16	8.49	6.08	8.35	6.07
liGRU + TERA-base: time + channel + mag	**8.32**	**6.01**	**8.29**	**6.00**	**8.31**	6.01

More pre-training data increases performance,
Consistent WER drop for the red rows

Using all three auxiliary objectives is potentially
increasing the amount of pre-training data.
100 hr == 960 hr

Effect of Amount of Pre-training Data - ASR 3/4

Pre-train / Models	100 hr		460 hr		960 hr	
	WER	Rescore	WER	Rescore	WER	Rescore
liGRU + TERA-base: time	8.46	6.12	8.38	6.10	**8.31**	**5.99**
liGRU + TERA-base: time + mag	8.43	6.11	8.38	6.04	8.40	6.03
liGRU + TERA-base: time + channel	8.57	6.16	8.49	6.08	8.35	6.07
liGRU + TERA-base: time + channel + mag	**8.32**	**6.01**	**8.29**	**6.00**	**8.31**	6.01

More pre-training data increases performance,
Consistent WER drop for the red rows

Using all three auxiliary objectives is potentially
increasing the amount of pre-training data.
100 hr == 960 hr

100 hr block to 460 hr block gives performance
boost, saturates for 460 hr to 960 hr

Effect of Amount of Pre-training Data - ASR 4/4

Models \ Pre-train	100 hr		460 hr		960 hr	
	WER	Rescore	WER	Rescore	WER	Rescore
liGRU + TERA-base: time	8.46	6.12	8.38	6.10	**8.31**	**5.99**
liGRU + TERA-base: time + mag	8.43	6.11	8.38	6.04	8.40	6.03
liGRU + TERA-base: time + channel	8.57	6.16	8.49	6.08	8.35	6.07
liGRU + TERA-base: time + channel + mag	**8.32**	**6.01**	**8.29**	**6.00**	**8.31**	6.01

Using all three auxiliary objectives is potentially increasing the amount of pre-training data

Ablation Study

Representation	Pre-train Context	Phone Classification		Speaker Recognition		Speech Recognition	
		Linear	1 Hidden	Frame	Average	WER	Rescore
MFCC	none	39.7	59.9	17.6	10.8	8.66	6.42
fMLLR	none	52.6	68.4	0.4	2.6	8.63	6.25
TERA-base: random	none	15.3	4.8	0.4	0.7	16.96	13.68
TERA-base: none	unidirectional	57.0	66.2	1.3	12.7	9.67	7.17
TERA-base: mag	unidirectional	59.7	69.5	2.3	28.8	9.32	6.93
TERA-base: channel	unidirectional	65.0	76.6	96.7	99.0	9.41	6.91
TERA-base: channel + mag	unidirectional	64.2	75.7	97.3	99.2	9.33	6.87
TERA-base: time	bidirectional	64.3	76.8	68.4	96.1	8.46	6.12
TERA-base: time + mag	bidirectional	64.1	77.1	70.8	96.1	8.43	6.11
TERA-base: time + channel	bidirectional	**65.2**	**77.4**	93.6	98.5	8.57	6.16
TERA-base: time + channel + mag	bidirectional	65.1	77.3	**98.9**	**99.2**	**8.32**	**6.01**
TERA-base: time + channel + mag (MFCC)	bidirectional	61.5	74.2	95.5	98.8	10.84	8.06

Ablation Study - 1) Importance of Bidirectionality

Representation	Pre-train Context	Phone Classification		Speaker Recognition		Speech Recognition	
		Linear	1 Hidden	Frame	Average	WER	Rescore
MFCC	none	39.7	59.9	17.6	10.8	8.66	6.42
fMLLR	none	52.6	68.4	0.4	2.6	8.63	6.25
TERA-base: random	none	15.3	4.8	0.4	0.7	16.96	13.68
TERA-base: none	unidirectional	57.0	66.2	1.3	12.7	9.67	7.17
TERA-base: mag	unidirectional	59.7	69.5	2.3	28.8	9.32	6.93
TERA-base: channel	unidirectional	65.0	76.6	96.7	99.0	9.41	6.91
TERA-base: channel + mag	unidirectional	64.2	75.7	97.3	99.2	9.33	6.87
TERA-base: time	bidirectional	64.3	76.8	68.4	96.1	8.46	6.12
TERA-base: time + mag	bidirectional	64.1	77.1	70.8	96.1	8.43	6.11
TERA-base: time + channel	bidirectional	**65.2**	**77.4**	93.6	98.5	8.57	6.16
TERA-base: time + channel + mag	bidirectional	65.1	77.3	**98.9**	**99.2**	**8.32**	**6.01**
TERA-base: time + channel + mag (MFCC)	bidirectional	61.5	74.2	95.5	98.8	10.84	8.06

The time objective leads the model to learn bidirectional context!

Ablation Study - 2) Learning Speaker Identity

Representation	Pre-train Context	Phone Classification		Speaker Recognition		Speech Recognition	
		Linear	1 Hidden	Frame	Average	WER	Rescore
MFCC	none	39.7	59.9	17.6	10.8	8.66	6.42
fMLLR	none	52.6	68.4	0.4	2.6	8.63	6.25
TERA-base: random	none	15.3	4.8	0.4	0.7	16.96	13.68
TERA-base: none	unidirectional	57.0	66.2	1.3	12.7	9.67	7.17
TERA-base: mag	unidirectional	59.7	69.5	2.3	28.8	9.32	6.93
TERA-base: channel	unidirectional	65.0	76.6	96.7	99.0	9.41	6.91
TERA-base: channel + mag	unidirectional	64.2	75.7	97.3	99.2	9.33	6.87
TERA-base: time	bidirectional	64.3	76.8	68.4	96.1	8.46	6.12
TERA-base: time + mag	bidirectional	64.1	77.1	70.8	96.1	8.43	6.11
TERA-base: time + channel	bidirectional	**65.2**	**77.4**	93.6	98.5	8.57	6.16
TERA-base: time + channel + mag	bidirectional	65.1	77.3	**98.9**	**99.2**	**8.32**	**6.01**
TERA-base: time + channel + mag (MFCC)	bidirectional	61.5	74.2	95.5	98.8	10.84	8.06

The channel objective leads the model to learn speaker identity,
While it does not compromise ASR performance!

Ablation Study - 3) Using Different features

Representation	Pre-train Context	Phone Classification		Speaker Recognition		Speech Recognition	
		Linear	1 Hidden	Frame	Average	WER	Rescore
MFCC	none	39.7	59.9	17.6	10.8	8.66	6.42
fMLLR	none	52.6	68.4	0.4	2.6	8.63	6.25
TERA-base: random	none	15.3	4.8	0.4	0.7	16.96	13.68
TERA-base: none	unidirectional	57.0	66.2	1.3	12.7	9.67	7.17
TERA-base: mag	unidirectional	59.7	69.5	2.3	28.8	9.32	6.93
TERA-base: channel	unidirectional	65.0	76.6	96.7	99.0	9.41	6.91
TERA-base: channel + mag	unidirectional	64.2	75.7	97.3	99.2	9.33	6.87
TERA-base: time	bidirectional	64.3	76.8	68.4	96.1	8.46	6.12
TERA-base: time + mag	bidirectional	64.1	77.1	70.8	96.1	8.43	6.11
TERA-base: time + channel	bidirectional	**65.2**	**77.4**	93.6	98.5	8.57	6.16
TERA-base: time + channel + mag	bidirectional	65.1	77.3	**98.9**	**99.2**	**8.32**	**6.01**
TERA-base: time + channel + mag (MFCC)	bidirectional	61.5	74.2	95.5	98.8	10.84	8.06

Using fMLLR outperformes MFCC on all measures!

Ablation Study - 4) Comparing with baselines

Representation	Pre-train Context	Phone Classification		Speaker Recognition		Speech Recognition	
		Linear	1 Hidden	Frame	Average	WER	Rescore
MFCC	none	39.7	59.9	17.6	10.8	8.66	6.42
fMLLR	none	52.6	68.4	0.4	2.6	8.63	6.25
TERA-base: random	none	15.3	4.8	0.4	0.7	16.96	13.68
TERA-base: none	unidirectional	57.0	66.2	1.3	12.7	9.67	7.17
TERA-base: mag	unidirectional	59.7	69.5	2.3	28.8	9.32	6.93
TERA-base: channel	unidirectional	65.0	76.6	96.7	99.0	9.41	6.91
TERA-base: channel + mag	unidirectional	64.2	75.7	97.3	99.2	9.33	6.87
TERA-base: time	bidirectional	64.3	76.8	68.4	96.1	8.46	6.12
TERA-base: time + mag	bidirectional	64.1	77.1	70.8	96.1	8.43	6.11
TERA-base: time + channel	bidirectional	**65.2**	**77.4**	93.6	98.5	8.57	6.16
TERA-base: time + channel + mag	bidirectional	65.1	77.3	**98.9**	**99.2**	**8.32**	**6.01**
TERA-base: time + channel + mag (MFCC)	bidirectional	61.5	74.2	95.5	98.8	10.84	8.06

Pre-training leads to better performance!

Comparing different depth and size

Representation	#L	Pre-train / #Param	100 hr Linear	100 hr 1 Hidden	360 hr Linear	360 hr Concat	960 hr Linear	960 hr 1 Hidden
CPC [8] [2]	6	-	64.6	72.5	-	65.5	-	
Modified CPC [2]	6	-	-		-	68.9	-	
TERA-base	3	21.3M	65.1	77.3	66.4	68.3	66.4	**78.9**
TERA-medium	6	42.6M	65.9	77.5	66.6	68.9	**67.3**	78.8
TERA-large	12	85.1M	66.8	**77.7**	**67.5**	**71.7**	67.2	78.5
TERA-xlarge	24	170.1M	**66.9**	77.6	67.1	71.2	67.3	78.3

A deeper model helps when data is limited!

Comparison of recent approaches on ASR

Models	Pre-train	Labels	WER	Rescore
Bidir-CPC [1]	960 hr	96 hr	14.96	9.41
Bidir-CPC [1]	8000 hr	96 hr	13.69	8.70
vq-wav2vec [10]	960 hr	960 hr	6.2	-
wav2vec-large [12]	960 hr	100 hr	-	6.92
DeCoAR [12]	960 hr	100 hr	-	6.10
liGRU + MFCC	None	100 hr	8.66	6.42
liGRU + fMLLR	None	100 hr	8.63	6.25
liGRU + TERA-base	960 hr	100 hr	**8.31**	**6.01**
liGRU + TERA-medium	960 hr	100 hr	8.37	6.05
liGRU + TERA-large	960 hr	100 hr	8.35	6.01
liGRU + TERA-xlarge	960 hr	100 hr	8.47	6.03

The proposed approach outperformed all previous methods!

Conclusion

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

Thank You

Q&A