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



What is Self-Supervised Learning?

An analogy

How do Infants Learn? Can Machine do the same?



How do Infants Learn? Yes! Self-Supervised Learning







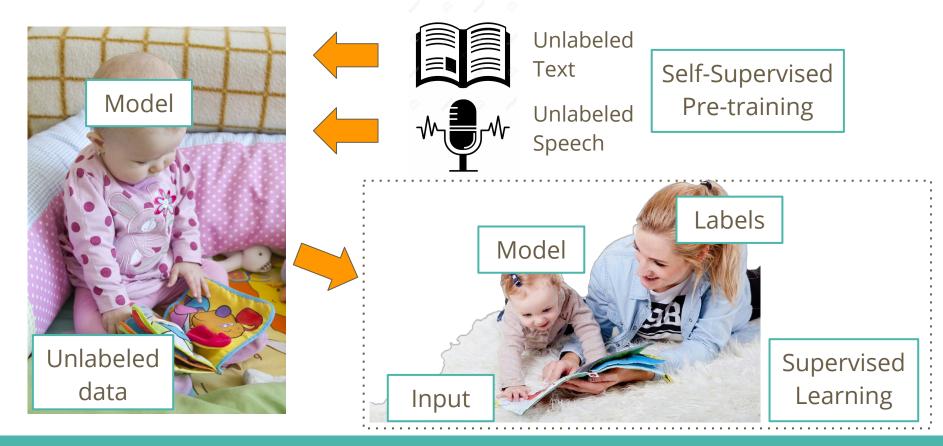
Unlabeled Text

Unlabeled

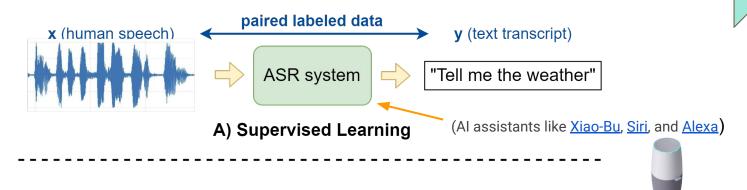
Speech

Self-Supervised Pre-training

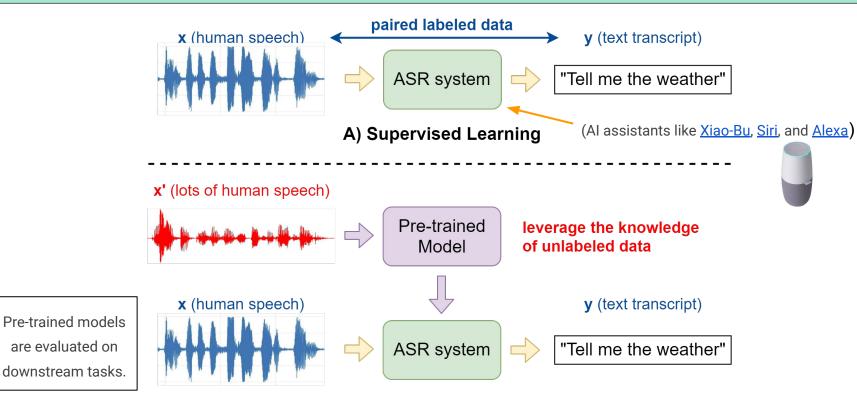
How do Infants Learn? Yes! Self-Supervised Learning



Self-Supervised Learning for Speech



Self-Supervised Learning for Speech



B) Self-Supervised Learning for Improving Supervised Systems

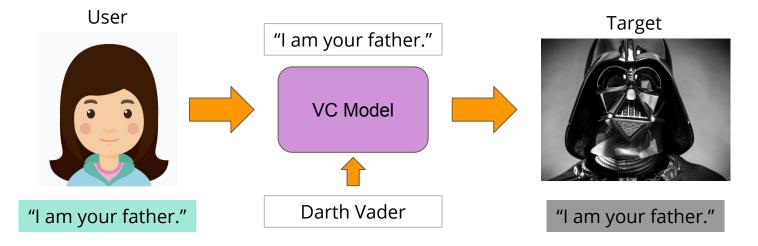
Examples Voice Conversion and BERT

An example: Voice Conversion

Given a source speech (User: I am your father), 👩 and target speaker's identity (Darth Vader),

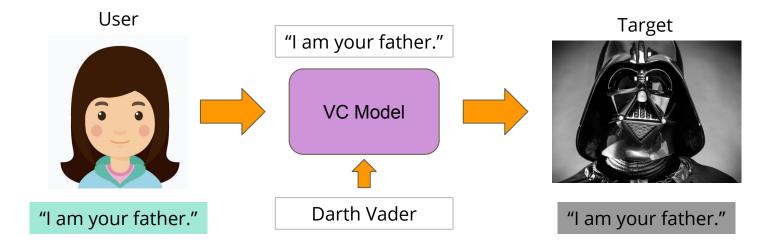


the converted speech should sound like B uttering A's content.

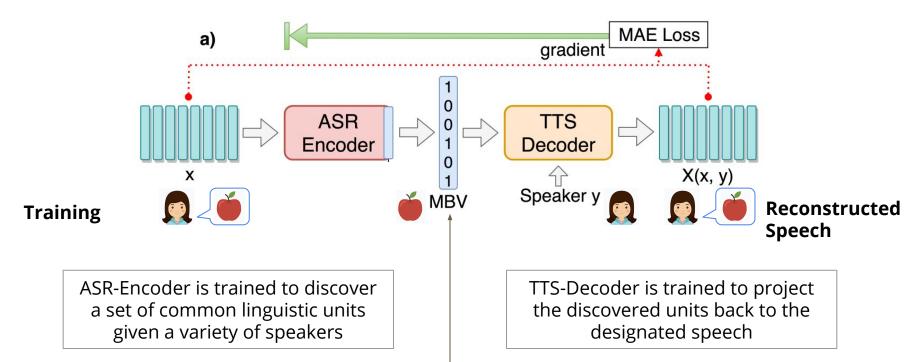


An example: Voice Conversion

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

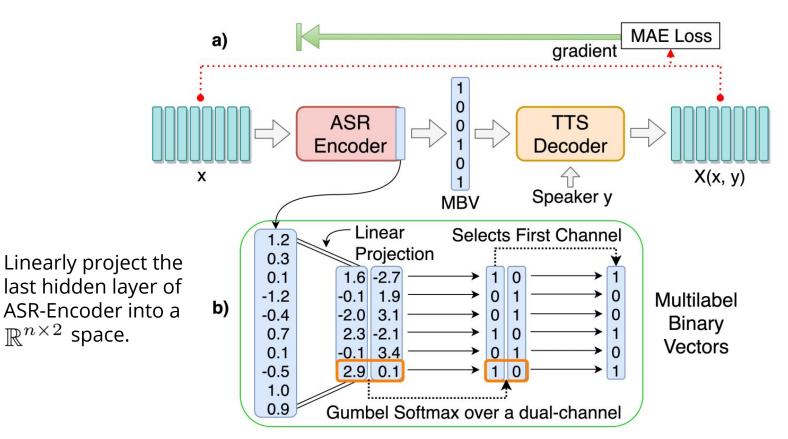


Voice Conversion (1/3) - Discrete linguistic units discovery

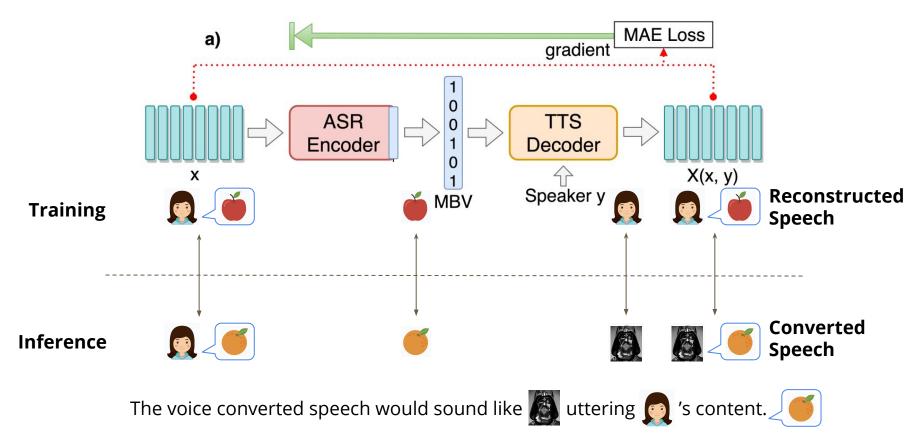


In this **self-supervised end-to-end** manner, **discrete linguistic units** are learned and **represented as multilabel binary vectors (MBVs)**.

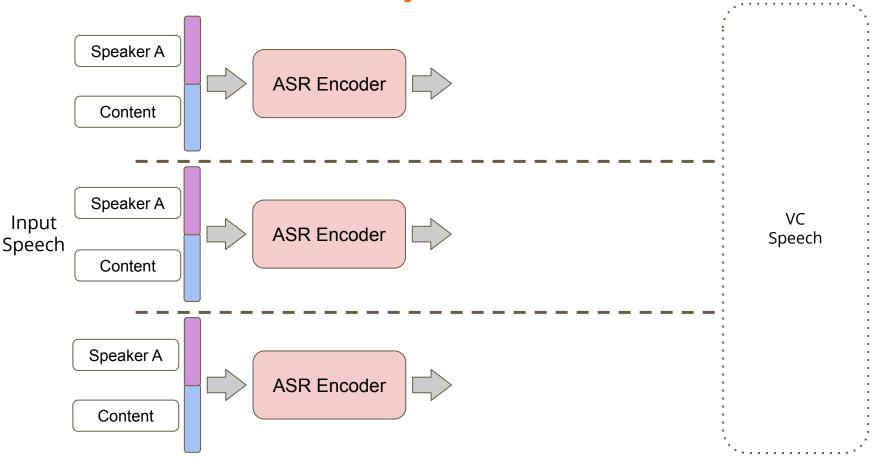
Voice Conversion (2/3) - MBV: vectors of zeros one ones



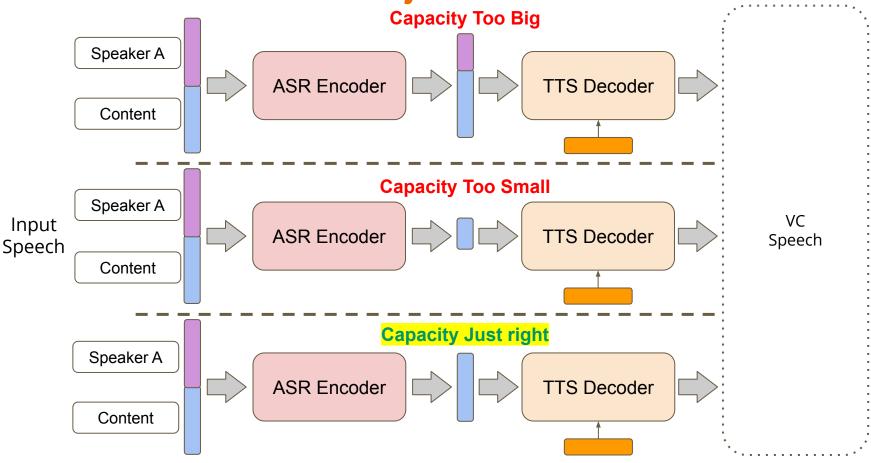
Voice Conversion (3/3) - VC using the ASR-TTS autoencoder



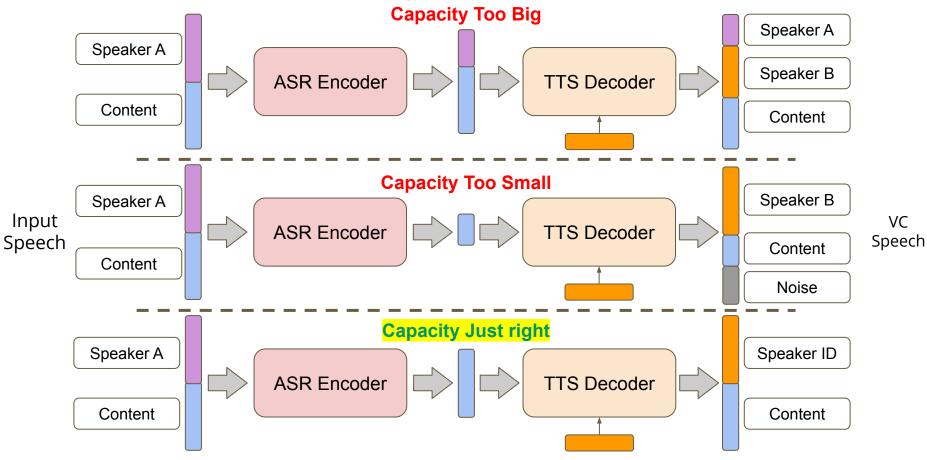
Voice Conversion - Why does it work?



Voice Conversion - Why does it work?



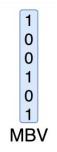
Voice Conversion - Why does it work?





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?

0

0

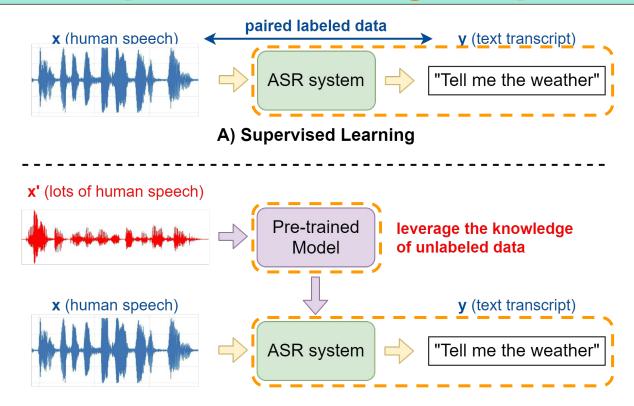
0

1 MBV

- 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).

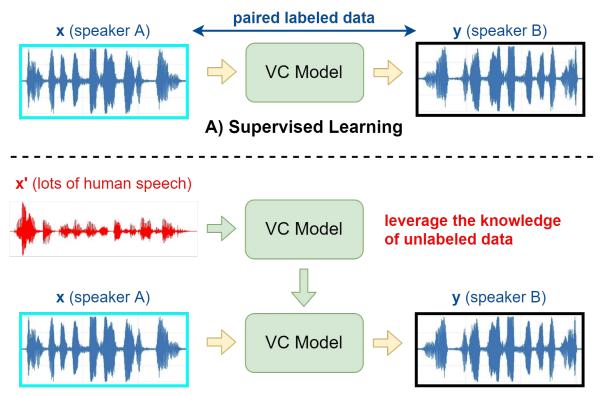
	# ≜	Authors 👙	Surprise language				
	# 👳		MOS 🍦	CER 🌢	Similarity 🗍	ABX 🌢	Bitrate 🔺
0	17	Gok et al.	1.46	0.86	3.03	27.26	29.46
٢	21	Topline	3.92	0.28	3.95	16.09	35.2
0	6	Liu <i>et al.</i>	1.69	0.81	1.97	44.25	43.95
0	18	Liu <i>et al.</i>	1.27	0.86	1.96	43.42	43.95
0	9	Kumar et al.	1.44	0.89	3.02	45.64	44.07
0	10	Kumar et al.	1.82	0.86	3.3	40.17	46.07
0	12	Kamper <i>et al.</i>	1.9 <mark>4</mark>	0. <mark>5</mark> 8	1.95	26.49	69.22
0	15	Rallabandi <i>et al.</i>	1.89	0.71	3.02	<mark>28.41</mark>	71.42
0	1	Baseline	2.07	0.62	3.41	27.46	74.55
0	3	Pandia et al.	2.02	0.48	3.21	20.77	94.15
0	2	Pandia et al.	2.53	0.43	3.58	<mark>23.5</mark> 6	115.43
0	16	Yusuf et al.	1.84	0.8	2.84	<mark>24.1</mark> 6	121.03
0	7	Kamper <i>et al.</i>	1.96	0.6	1.76	19.76	139.54
0	13	Cho et al.	1.23	0.85	1.28	12.05	143.76
0	14	Cho et al.	1.53	0.78	1.33	10.39	144.63
0	19	Tjandra <i>et al</i> .	3.25	0.35	2.67	17.8	151.77
0	5	Feng et al.	1.67	0.66	2.6	16.87	299.21
0	20	Tjandra <i>et al.</i>	3.2	0.21	2.3	13.98	362.99
0	11	Feng et al.	1.28	0.74	2.01	10.64	470.23
0	4	Horizon Robotics	2.89	0.36	1.43	24.92	842.46
0	8	Horizon Robotics	3.55	0.32	1.34	24.92	842.46

Recall: Self-Supervised Learning for Speech



B) Self-Supervised Learning for Improving Supervised Systems

Self-Supervised Learning: VC



B) Self-Supervised Learning for Improving Supervised Systems

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

An example: Machine QA



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



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 or 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 or 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

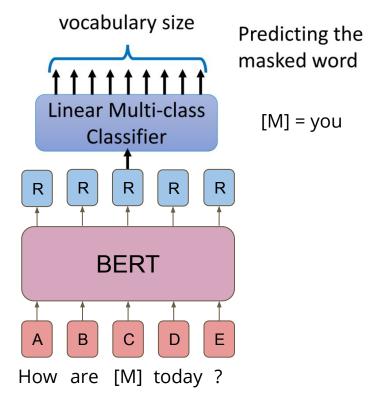
BERT (<u>B</u>idirectional <u>Encoder Representations from Transformers</u>)

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



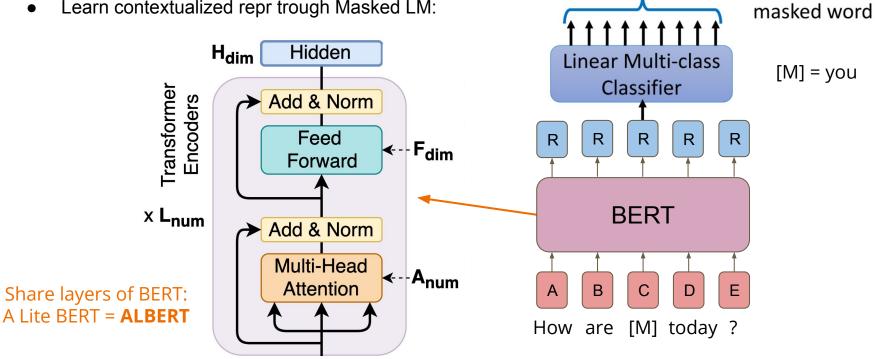
BERT (<u>B</u>idirectional <u>Encoder Representations from Transformers</u>)

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



BERT (<u>B</u>idirectional <u>Encoder Representations from Transformers</u>)

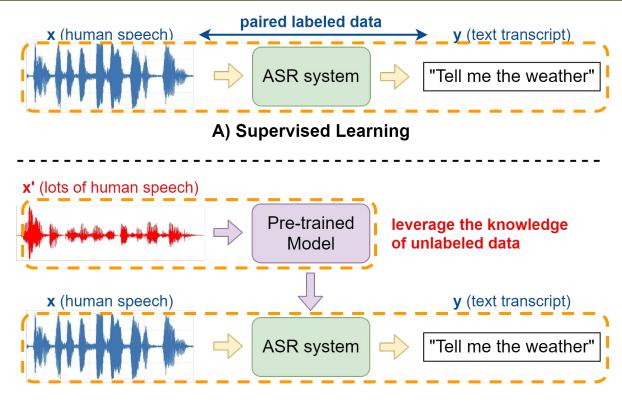
- Achieved **11 SOTA** when published.
- A technique for NLP pre-training developed by Google.
- Learn contextualized repr trough Masked LM:



vocabulary size

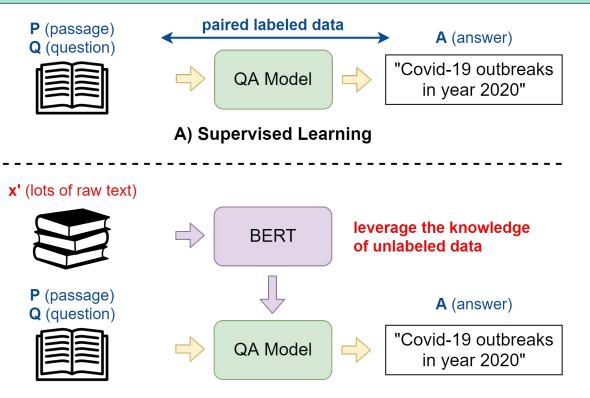
Predicting the

Recall: Self-Supervised Learning for Speech



B) Self-Supervised Learning for Improving Supervised Systems

Self-Supervised Learning: BERT

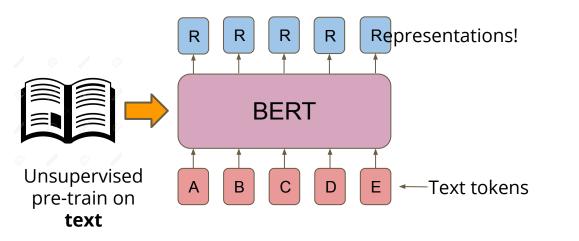


B) Self-Supervised Learning for Improving Supervised Systems

Mockingjay From BERT to Speech BERT

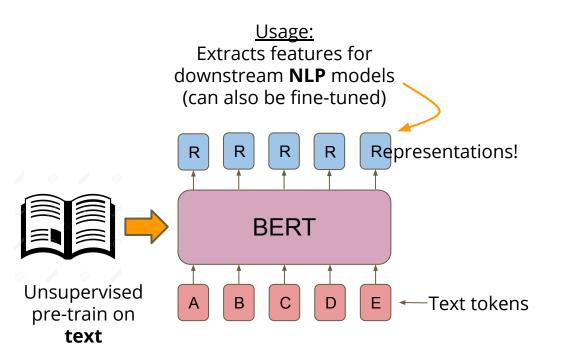
From BERT to Speech BERT

NLP BERT: Language Representation Learning



From BERT to Speech BERT

NLP BERT: Language Representation Learning



From BERT to Speech BERT

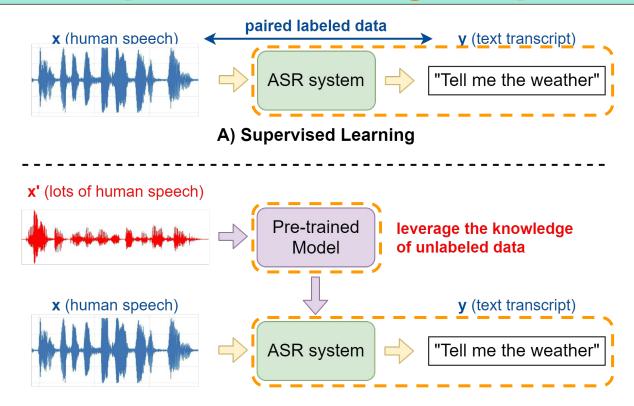
NLP BERT: Language Representation Learning

Usage: <u>Usage:</u> Extracts features for Extracts features for downstream **SLP** models downstream **NLP** models (can also be fine-tuned) (can also be fine-tuned) R R **Representations!** R R R R R R R Speech BERT BERT Unsupervised Unsupervised Е В pre-train on pre-train on text speech **Acoustic Frames**

Speech BERT:

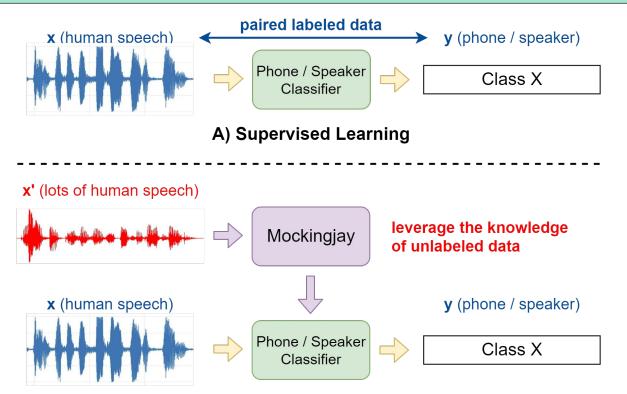
Speech Representation Learning

Recall: Self-Supervised Learning for Speech

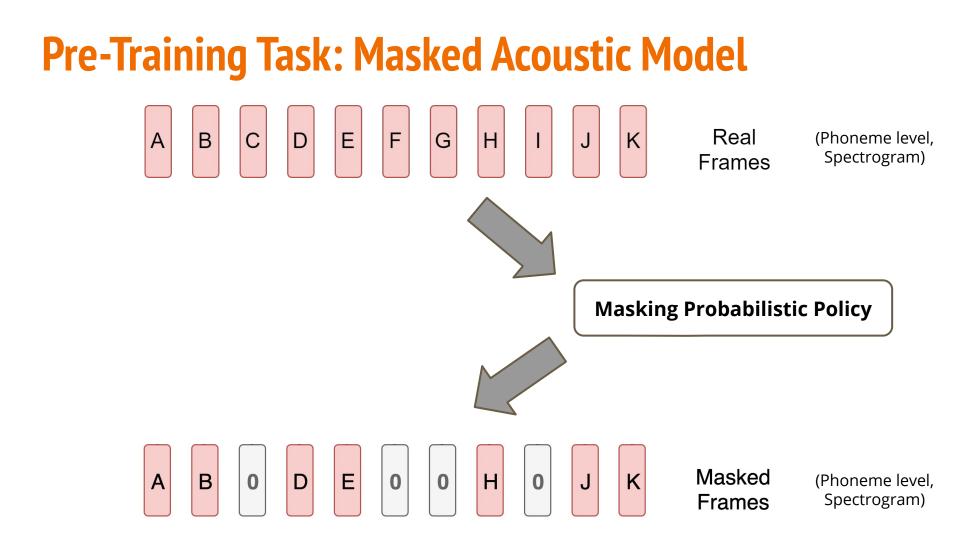


B) Self-Supervised Learning for Improving Supervised Systems

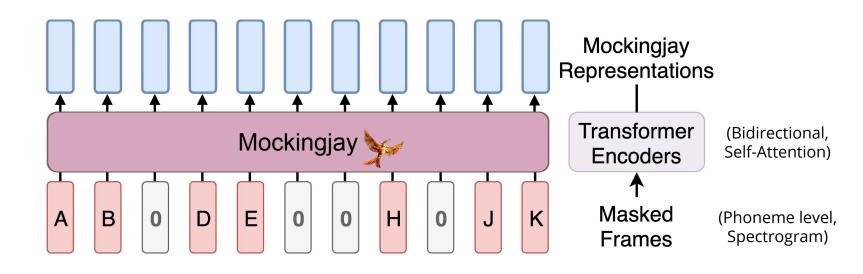
Self-Supervised Learning: Mockingjay



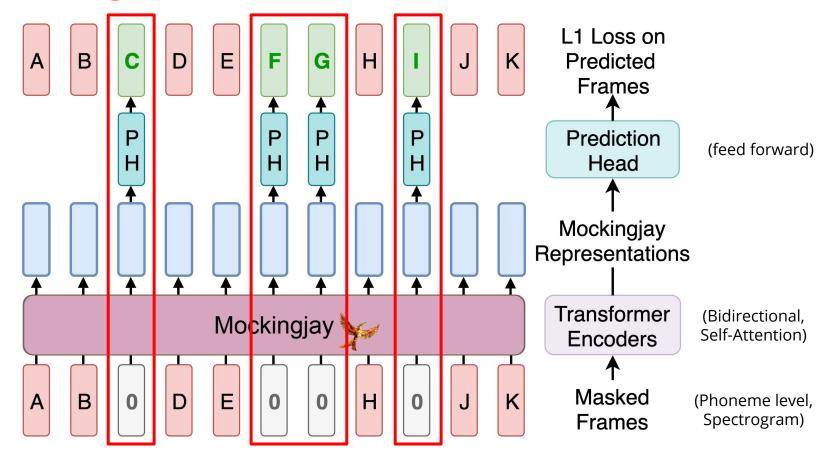
B) Self-Supervised Learning for Improving Supervised Systems



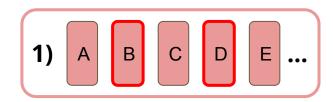
Pre-Training Task: Masked Acoustic Model A B C D E F G H I J K Real Frames



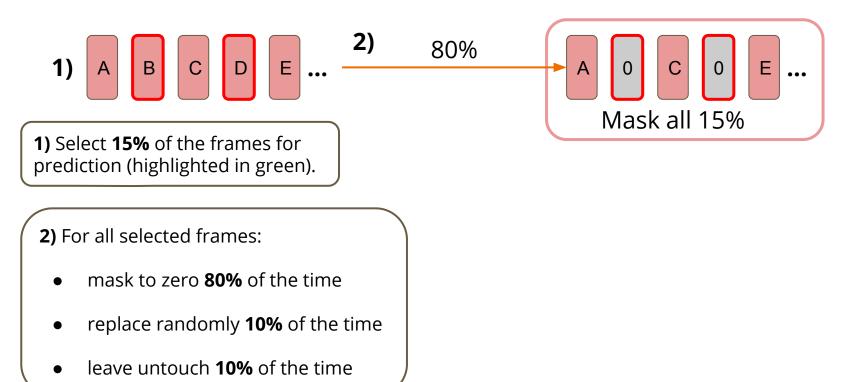
Pre-Training Task: Masked Acoustic Model

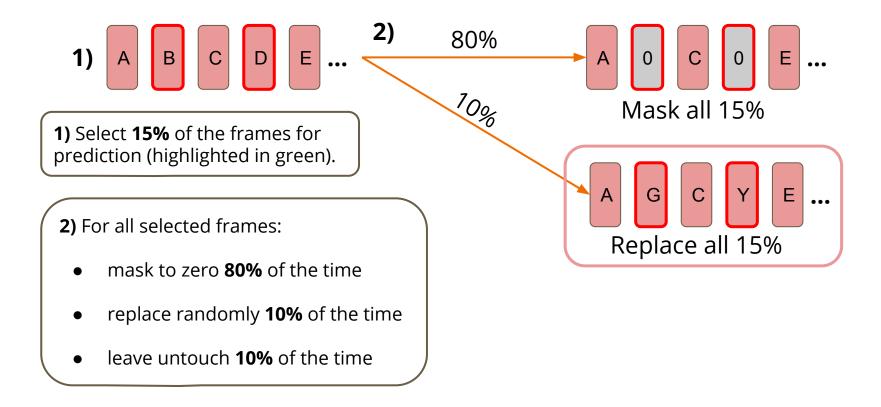


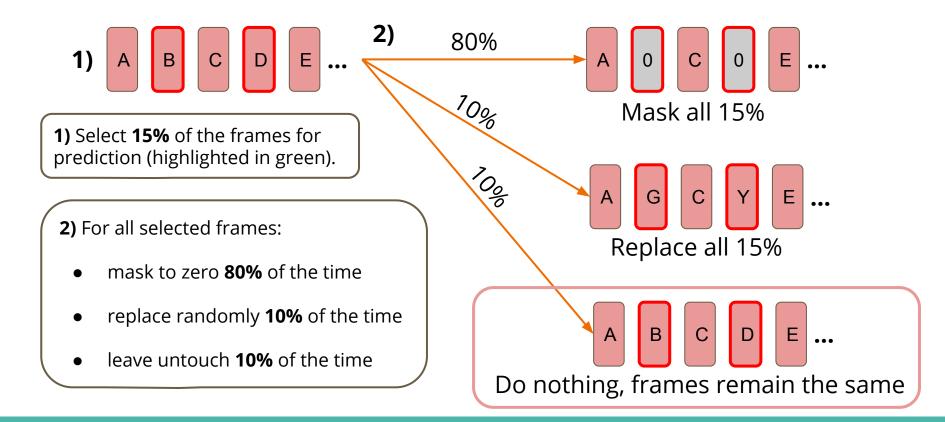
Pre-Training Task: Masked Acoustic Model L1 Loss on B С Ε F G Н Α K D J Predicted Frames Considers the Ρ Ρ Ρ Prediction Ρ (feed forward) whole utterance Н Н Н Head н Mockingjay Representations Transformer (Bidirectional, Mockingjay Self-Attention) Encoders Reconstructs Masked B Ε 0 Н K Α 0 D 0 0 J (Phoneme level, from corrupted Spectrogram) Frames input



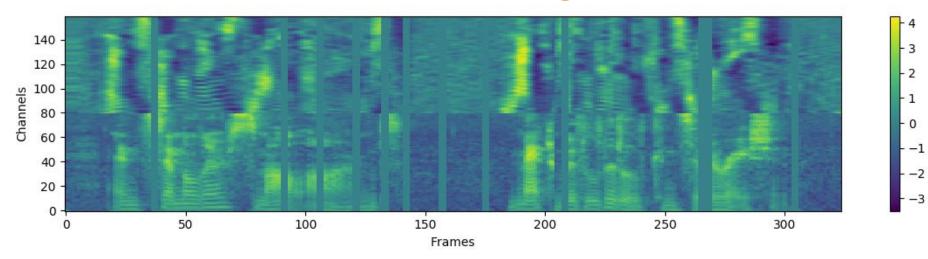
1) Select **15%** of the frames for prediction (highlighted in green).



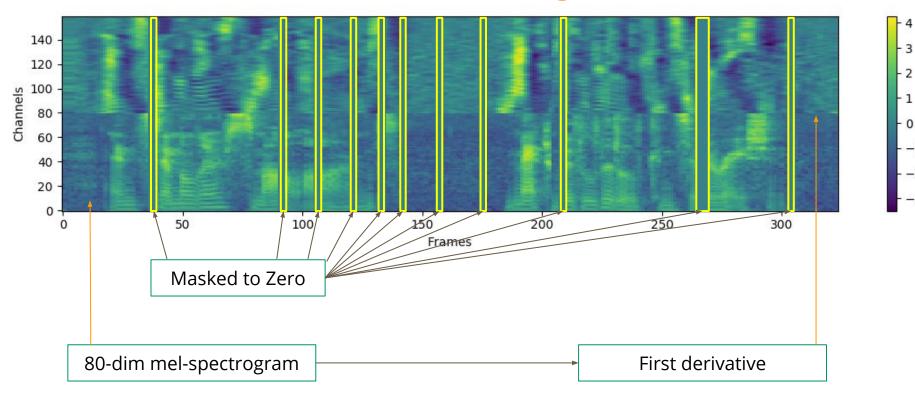


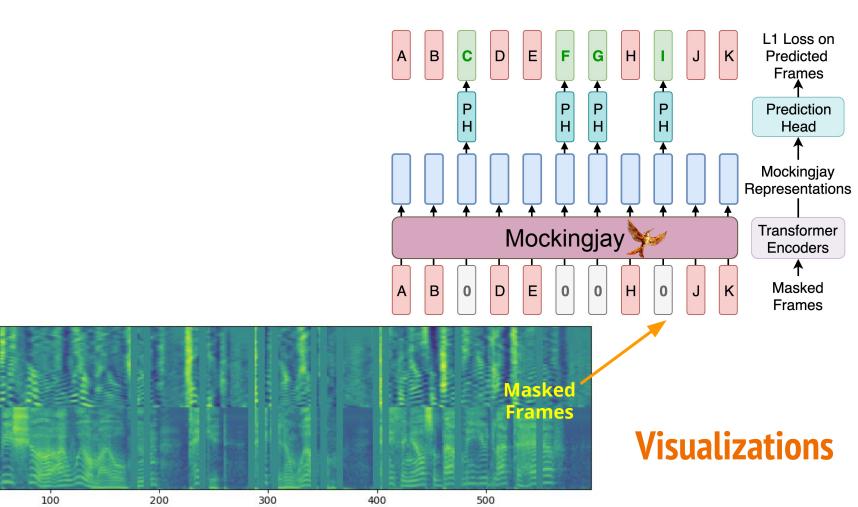


Input Feature: Masked Spectrogram



Input Feature: Masked Spectrogram



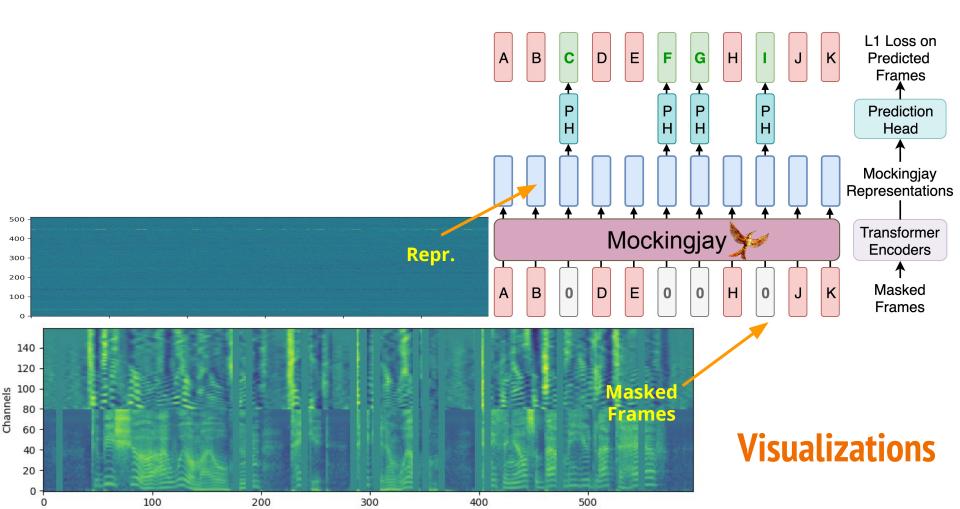


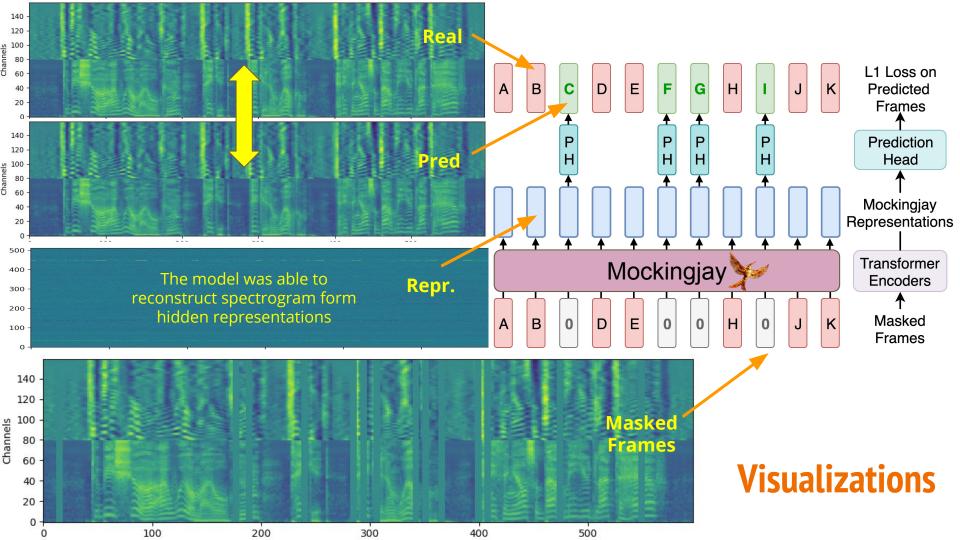
140 120

001 Channels 09 Channels

> 40 -20 -0 -

> > 0





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>

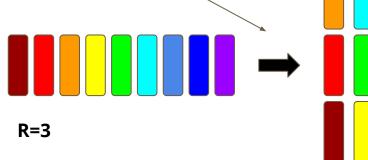


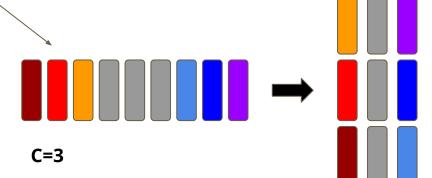
Migrating from text to speech

Acoustic Features: long and locally smooth in nature,

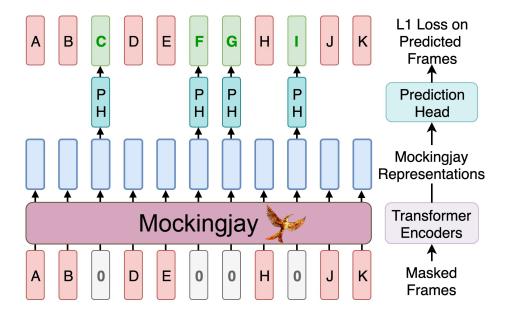
need to 1) shorten the sequence and 2) mask over a longer span

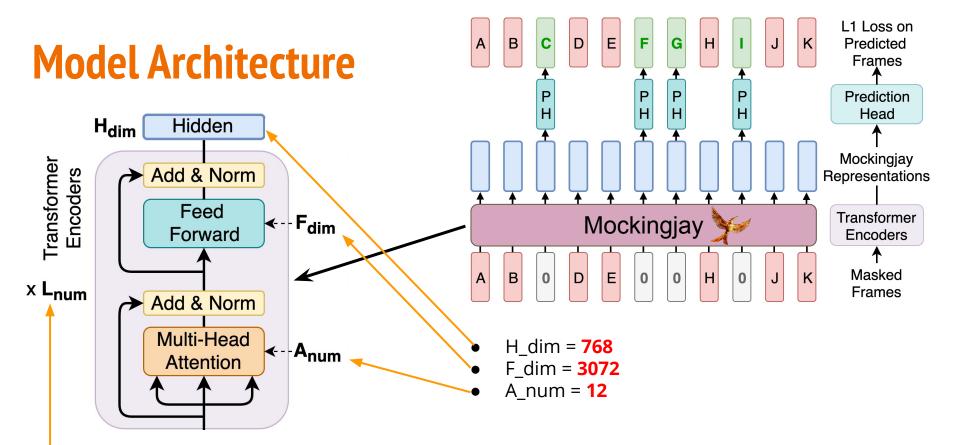






Model Architecture

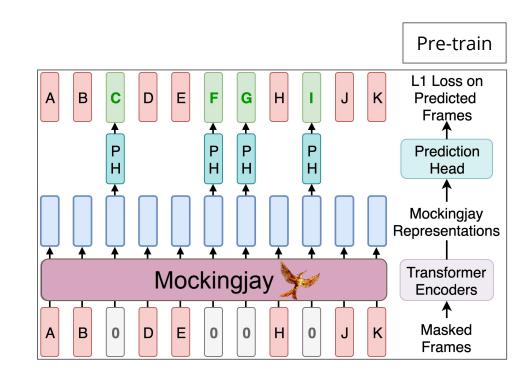




- Train on LibriSpeech **360 hrs**
- Pre-train steps = **500k**
- Fine-tune steps = **50k** (2-epochs)

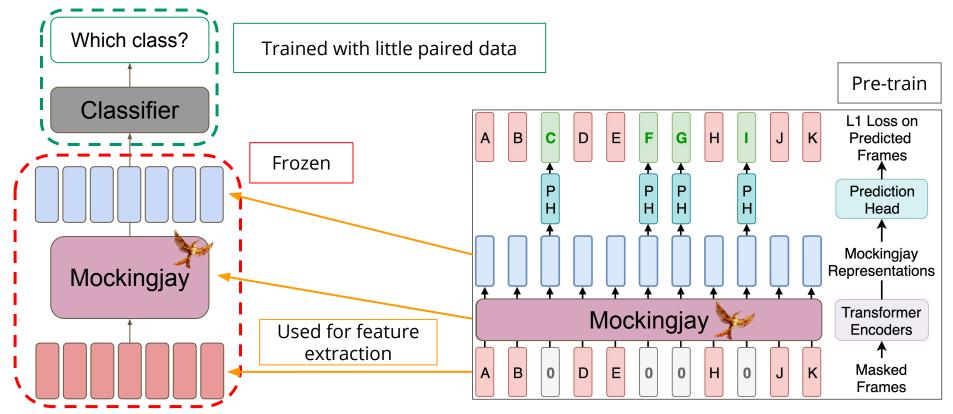
BASE (L=3)
 LARGE (L=12)

Incorporating with Downstream Tasks 1) Feature Extraction

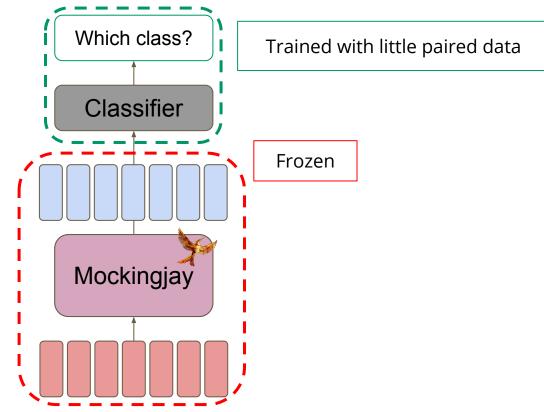


Incorporating with Downstream Tasks

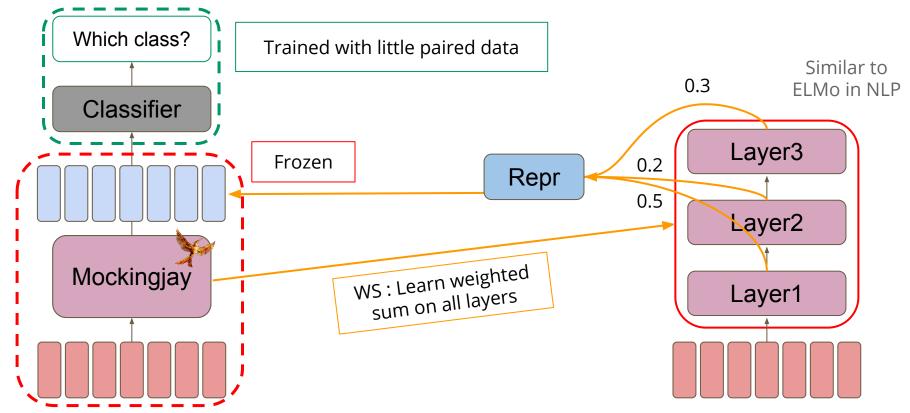
1) Feature Extraction



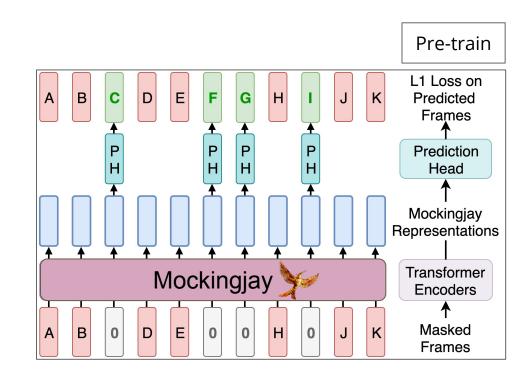
Incorporating with Downstream Tasks 2) Weighted Sum from All Layers (WS)



Incorporating with Downstream Tasks 2) Weighted Sum from All Layers (WS)

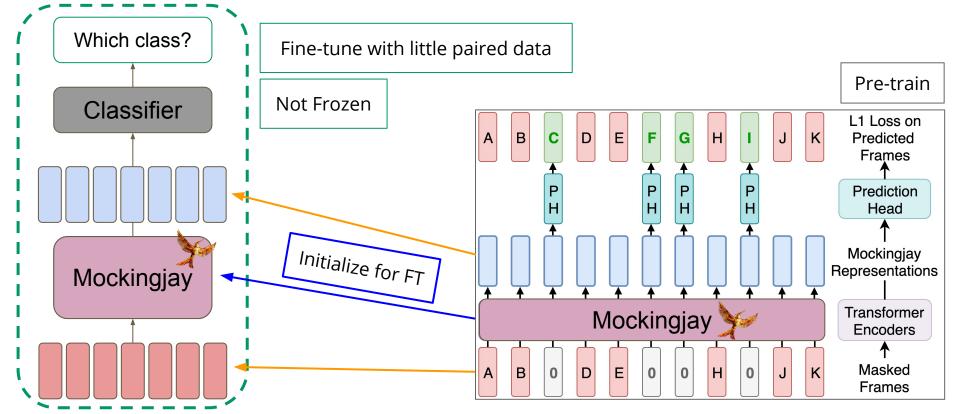


Incorporating with Downstream Tasks 3) Fine-tune (FT2)



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

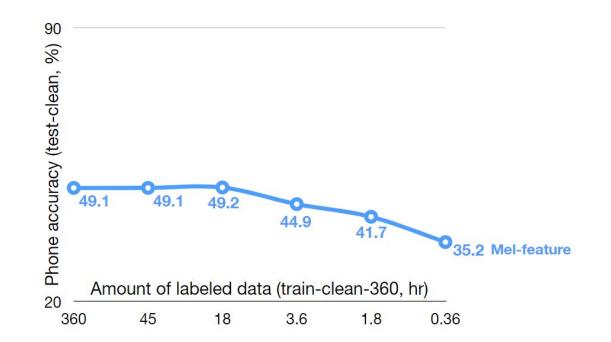
Consistent results over all three tasks: Mel < BASE < LARGE

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

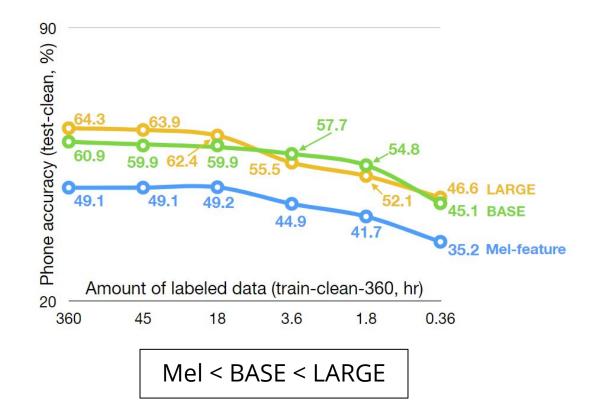


We demonstrate how pre-training on speech can improve supervised training in low resource scenarios, we train with reduced amount of labels.

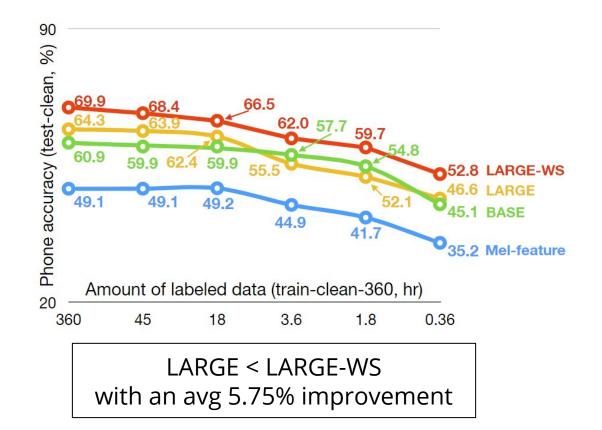
Low-Resource Experiments - 2/6



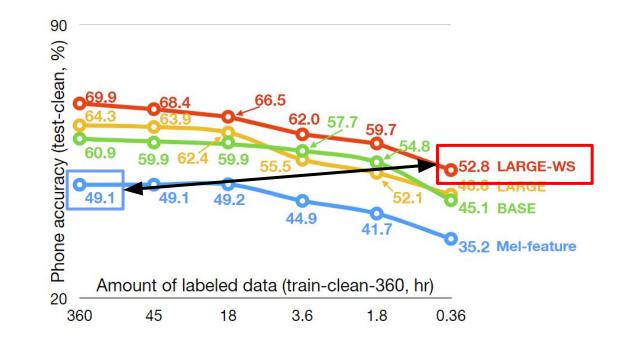
Low-Resource Experiments - 3/6



Low-Resource Experiments - 4/6



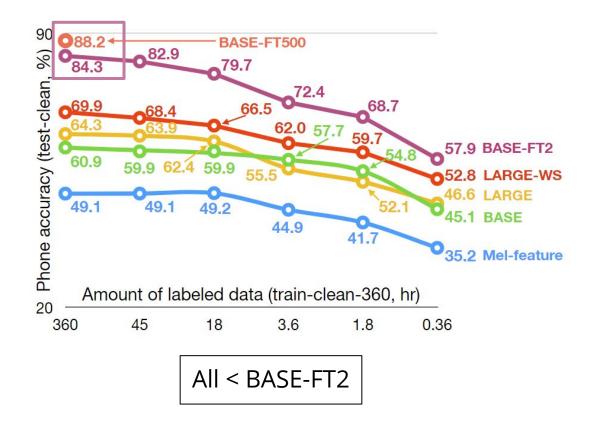
Low-Resource Experiments - 4/6



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



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.

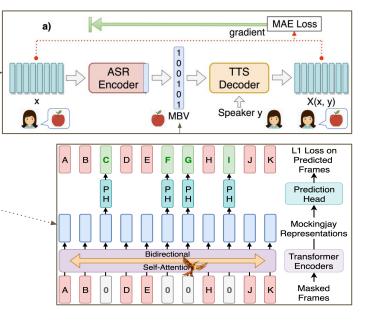
From here to beyond

- 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

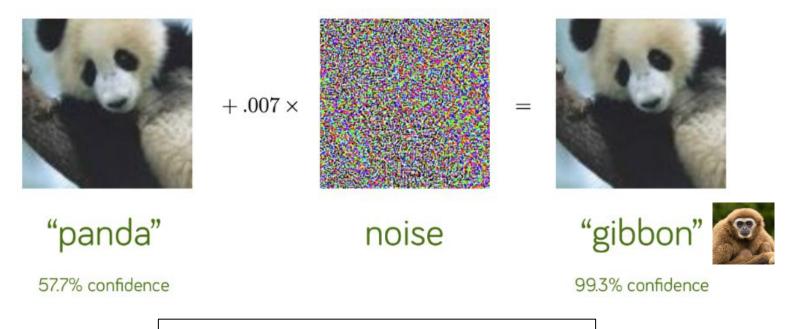
What else can we do with Mockingjay?

1. Adversarial Defense

Employ Mockingjay to protect models against adversarial attacks

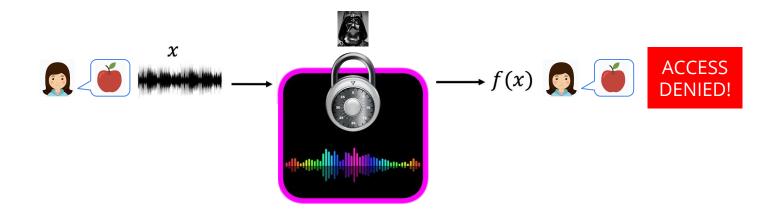
1. Adversarial Defense

What is Adversarial Attack?



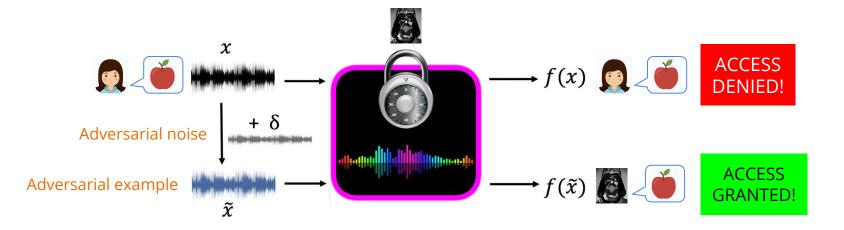
Hacking AI security systems: Face ID / Voice ID

1. Adversarial Defense What is Adversarial Attack?



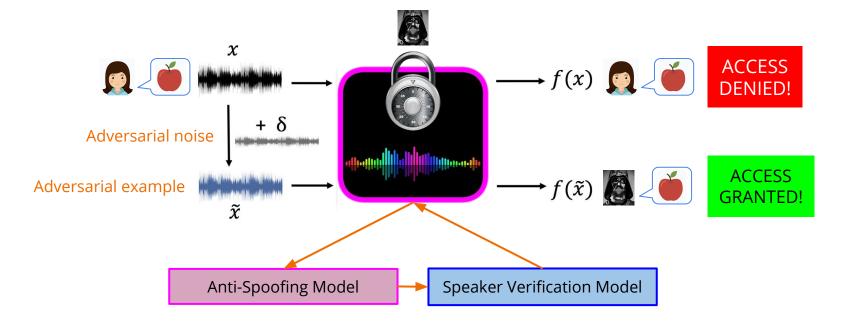
Hacking AI security systems: Face ID / Voice ID

1. Adversarial Defense What is Adversarial Attack?

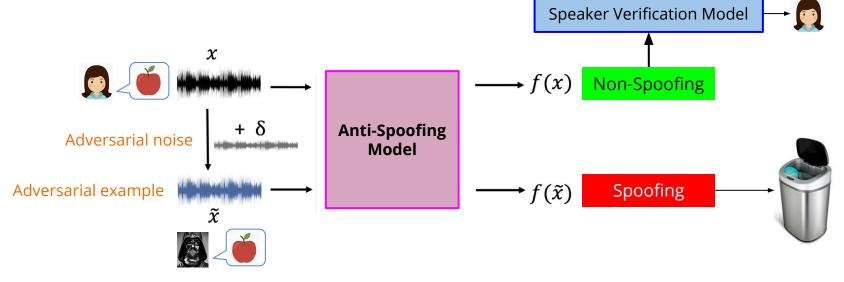


Hacking AI security systems: Face ID / Voice ID

1. Adversarial Defense What is Adversarial Attack?



1. Adversarial Defense How to Attack?



1. Adversarial Defense How to Attack? Speaker Verification Model х $\rightarrow f(x)$ Non-Spoofing **()** +δ **Anti-Spoofing** Adversarial noise Model Non-Spoofing Adversarial example $\bullet f(\tilde{x})$

 $Diff(f(x), f(\tilde{x}))$

Gradient Descent

ñ

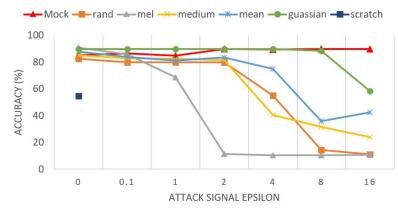
12

•

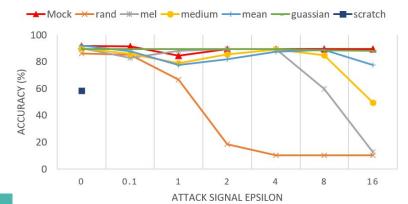
1. Adversarial Defense Employing Mockingjay Speaker Verification Model х f(x)Non-Spoofing Mockingjay +δ **Anti-Spoofing** Adversarial noise Model Adversarial example Spoofing $f(\tilde{x})$ ñ

1. Adversarial Defense - Experiments

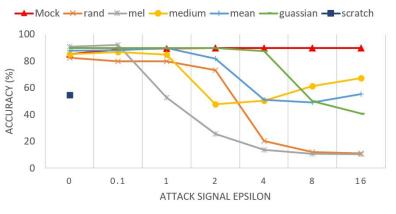
A) ATTACKING LCNN WITH PGD



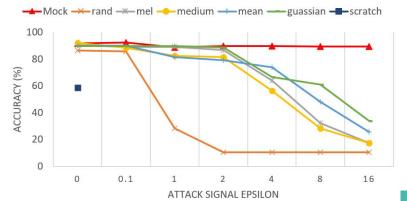
C) ATTACKING SENET WITH PGD

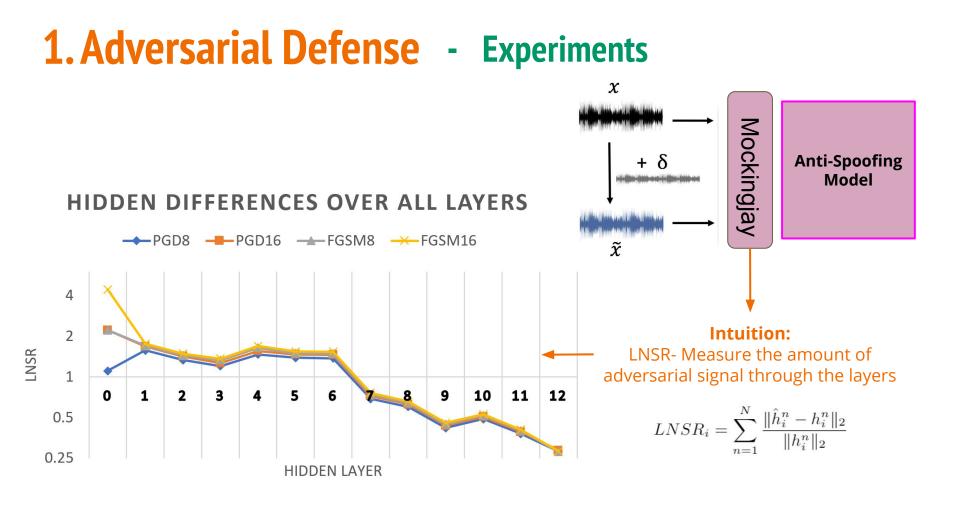


B) ATTACKING LCNN WITH FGSM



D) ATTACKING SENET WITH FGSM





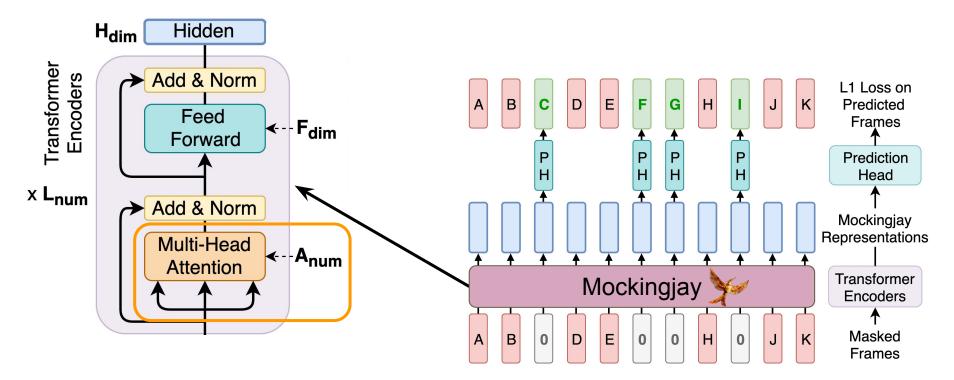
Current Works

What else can we do with Mockingjay?

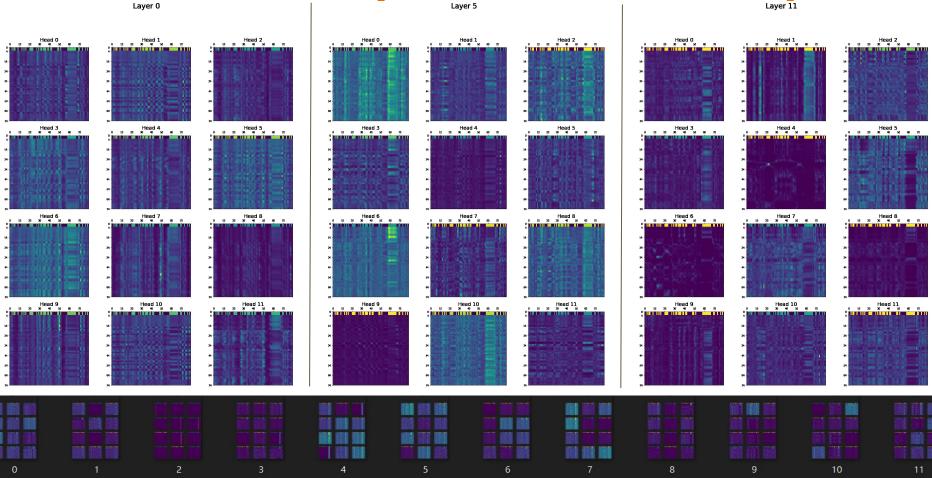
2. Understanding Self-Supervised Models

Visualize and explainable understanding of how models draw conclusion

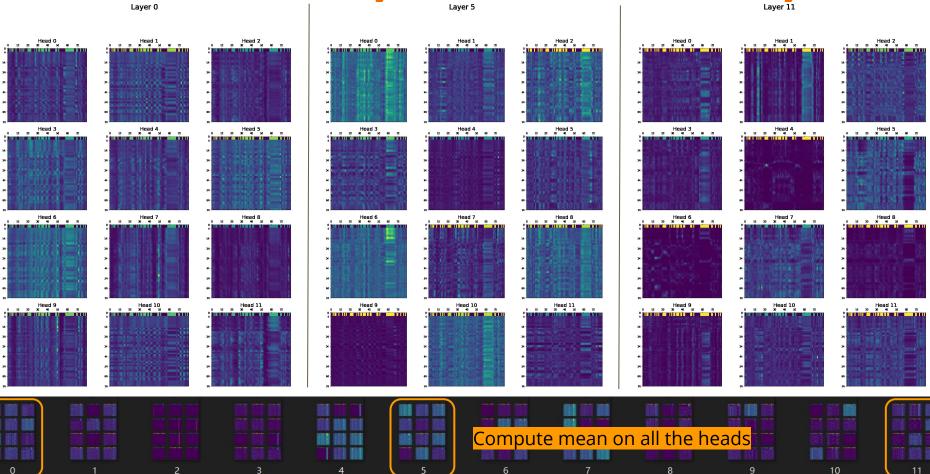
Recall: Model Architecture



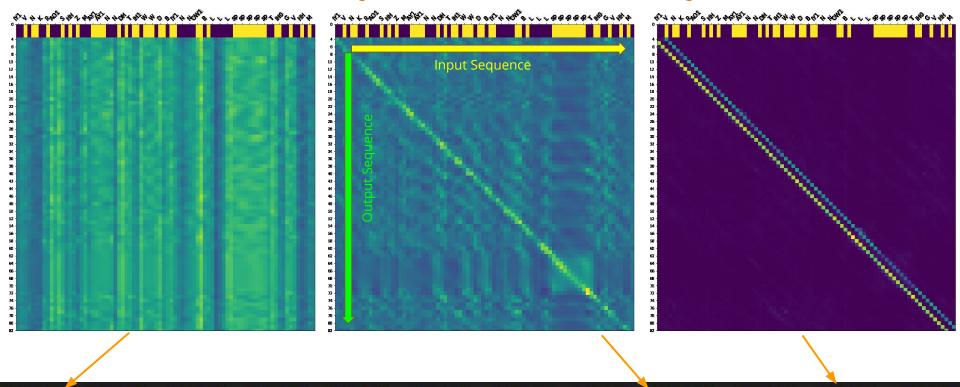
Attention of layers - 12 Heads Summary



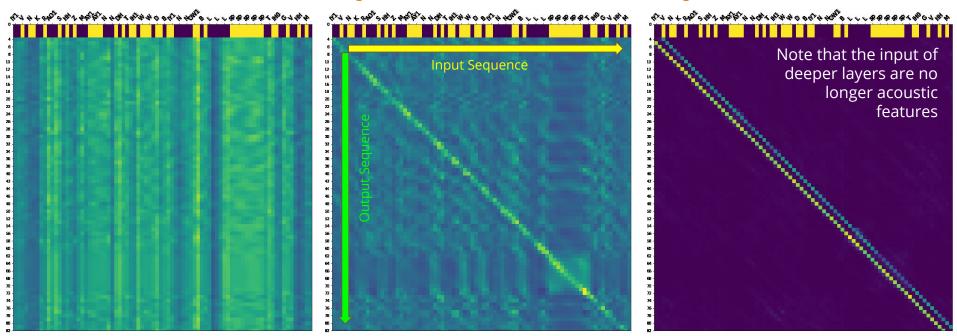
Attention of layers - 12 Heads Summary



Attention of layers - What Each Layer Does?



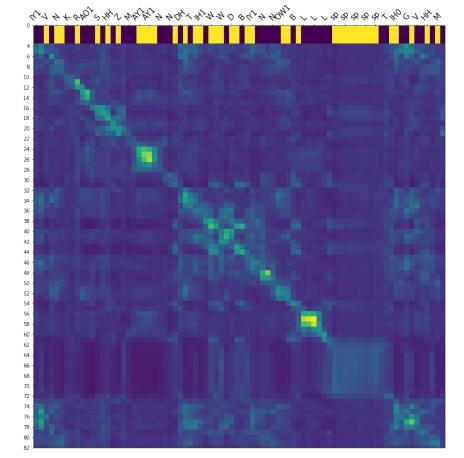
Attention of layers - What Each Layer Does?



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

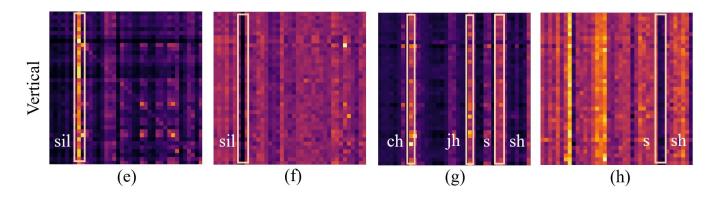
Diagonal - Observing Phoneme Boundaries



Diagonal - Observing Phoneme Boundaries

diagonal attentions are highly correlated with phoneme boundaries

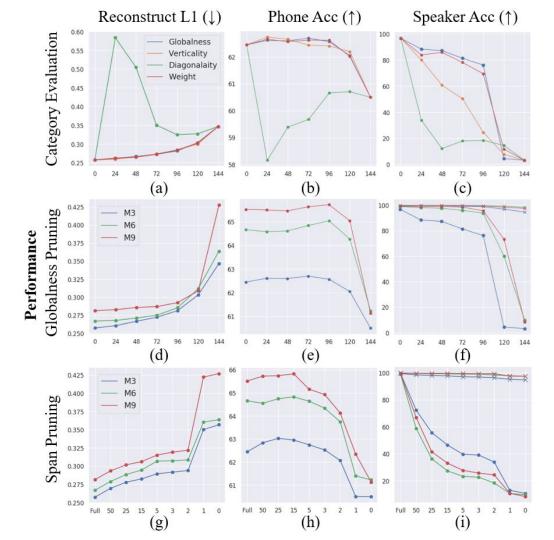
Vertical - Observing concentration



(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

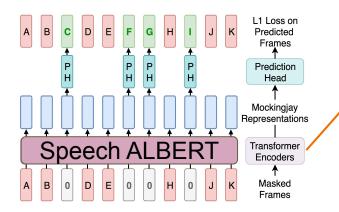
Recall that: Mockingjay = Speech BERT

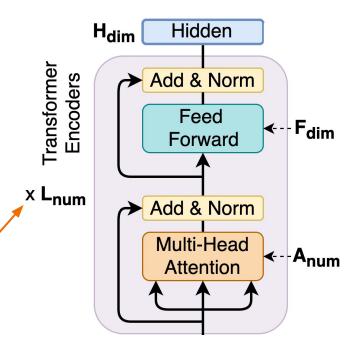
Speech ALBERT:

Share all the weights of each Transformer Layer!

Exps:

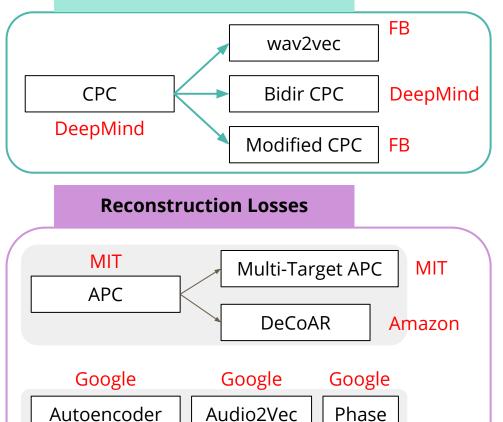
comparing with Mockingjay (uses less memory)

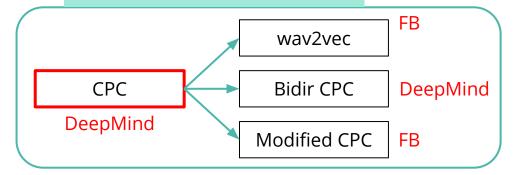


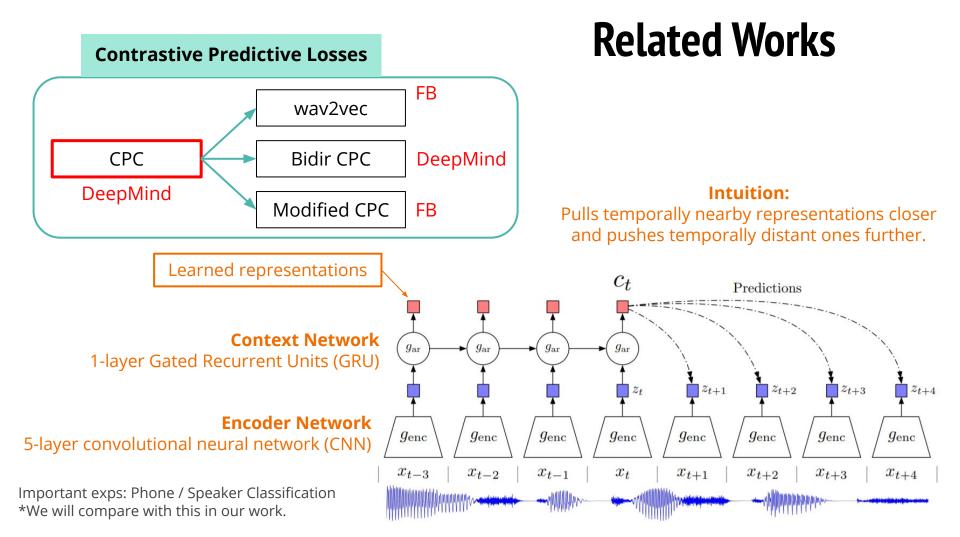


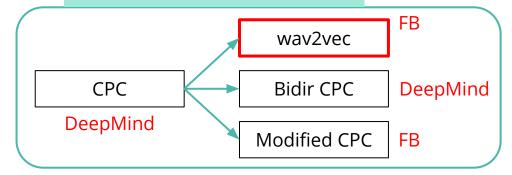
Related Works What else besides Mockingjay?

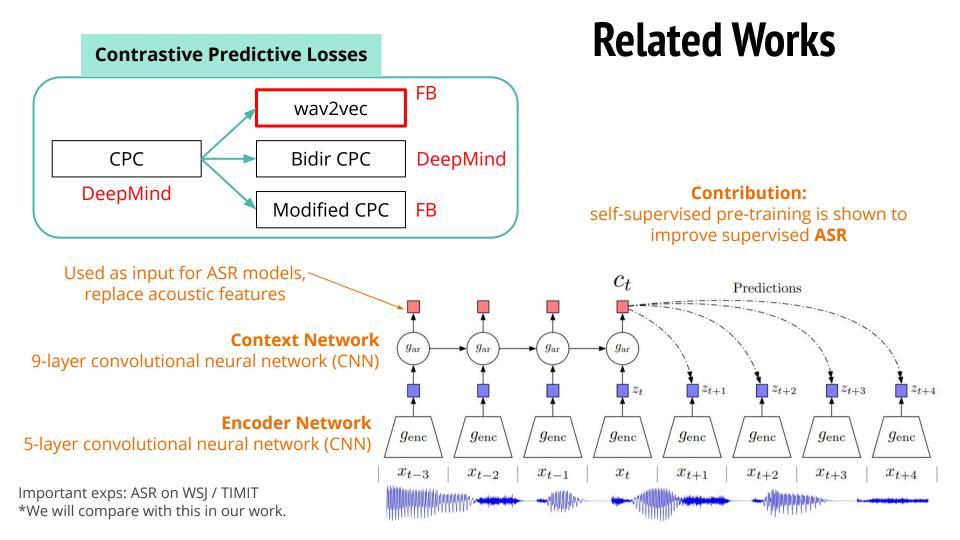
A top-down introduction to all recent related works.

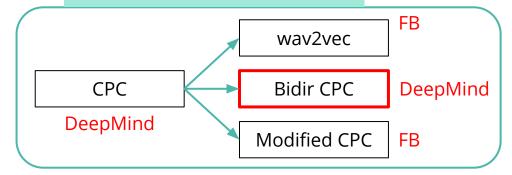


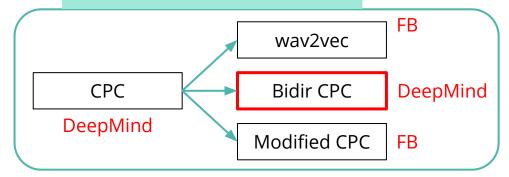








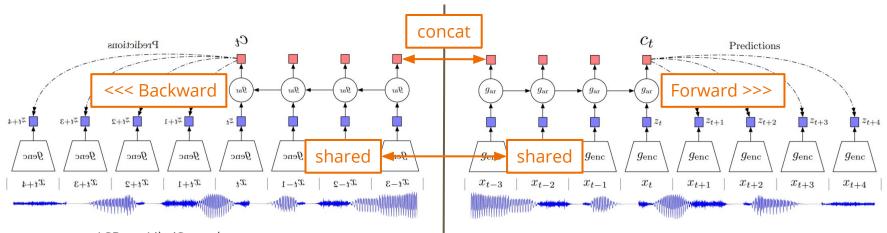




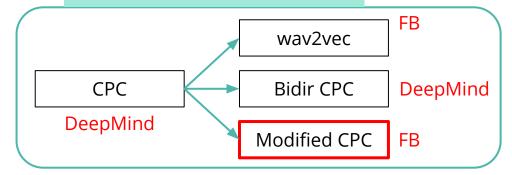
Related Works

Contribution: bidirectional context + ASR

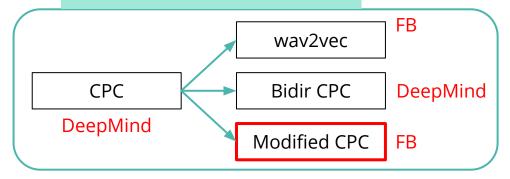
learning representations from large amount of unlabeled data (8000 hrs) can provide improvements for out-of-domain transfer (different datasets / cross-lingual).



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



Related Works



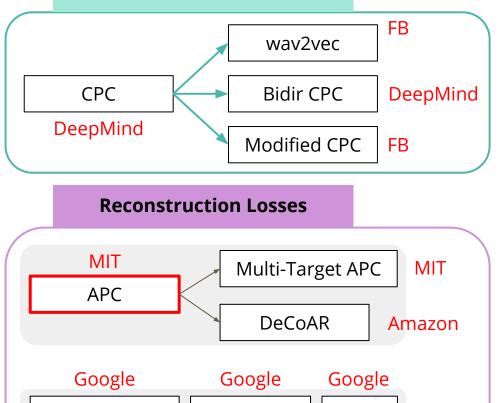
Contributions:

changing the batch normalization to channel-wise normalization
 replace the linear prediction layer to a Transformer layer
 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.

Autoencoder

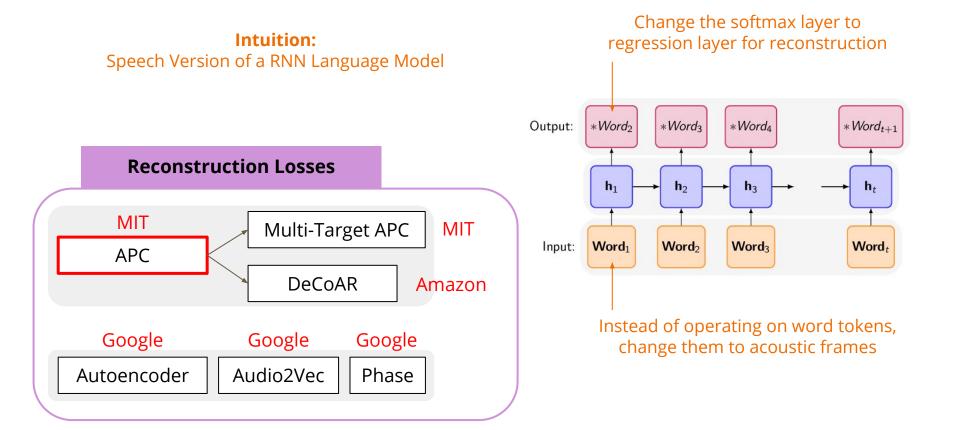
Related Works

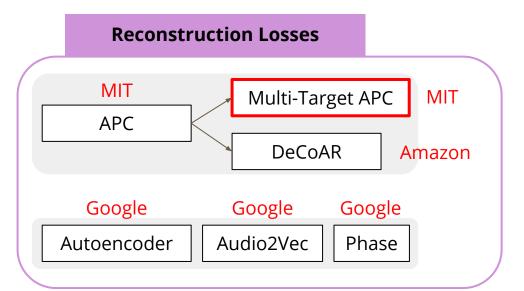


Audio2Vec

Phase

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

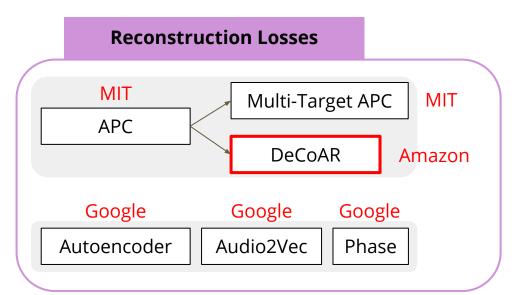




Important exps: Phone Classification, ASR on WSJ They use settings that are not conventional.

Related Works

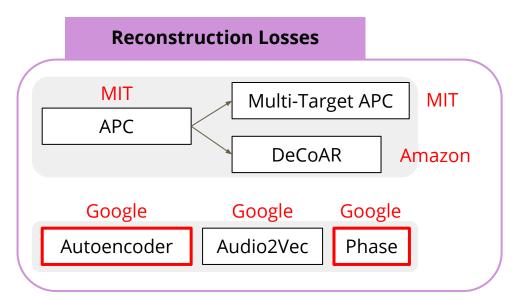
Intuition: The APC objective is extended to bidirectional. \mathbf{x}_{t-3} \mathbf{x}_{t-2} \mathbf{x}_{t-1} L_r target An auxiliary RNN is used to refresh current hidden states with the knowledge learned in L_f target x_{t-3} \mathbf{x}_{t-2} \mathbf{x}_{t-1} Xt \mathbf{x}_{t+1} the past, allowing the model to remember x_{t-4} \mathbf{x}_{t-3} x_{t-2} L_r input more from the past. h_t **Reconstruction Losses** L_f input x_{t-4} \mathbf{x}_{t-3} \mathbf{x}_{t-2} \mathbf{x}_{t-1} \mathbf{X}_t MIT Multi-Target APC MIT predicts the future frame APC conditioning on previous context, DeCoAR Amazon but also predicts the past Google Google memory through reconstruction. Google Autoencoder Audio2Vec Phase

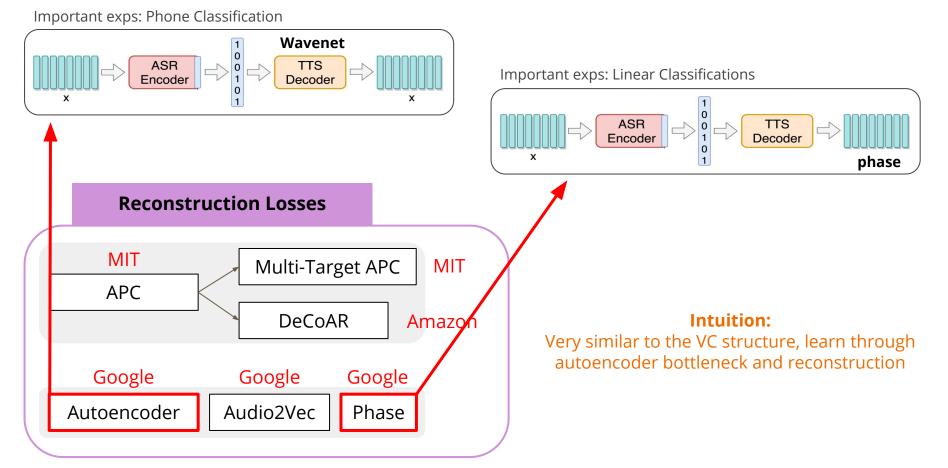


Important exps: ASR on WSJ / LibriSpeech *We will compare with this in our work.

Related Works

Intuition: <u>Deep Contextualized</u> <u>Acoustic Representations</u> **Representation Learning with Unlabeled Data** APC Combining the bidirectionality of ELMo and the reconstruction objective of APC. Reconstruction loss is Reconstruction summed over all possible slices in the entire sequence. FFN FFN **ELMO** \overline{z}_{t+K} **Reconstruction Losses** Encoder forward LSTM MIT Multi-Target APC MIT ... APC DeCoAR Amazon forward LSTM Google Google Google Filterbank Feature Autoencoder Audio2Vec Phase \boldsymbol{x}_{t} x_{t+K}

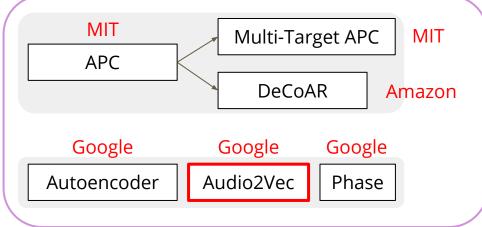


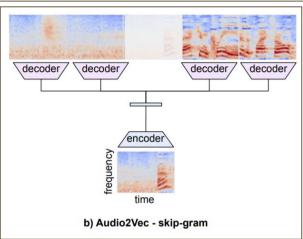


Important exps: Linear Classifications

frequency

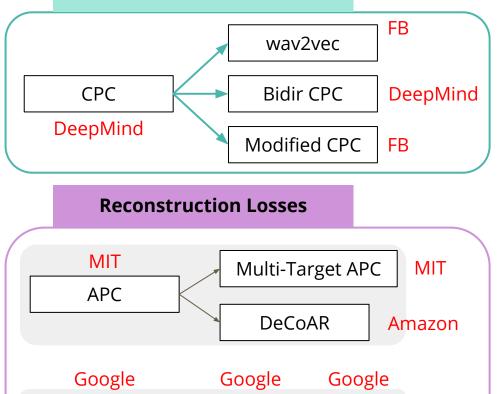
Related Works encoder encoder decoder ? Intuition: time Audio version of Word2Vec encoder encoder encoder encoder c) Temporal gap **frequency** time **Reconstruction Losses** a) Audio2Vec - CBoW





Autoencoder

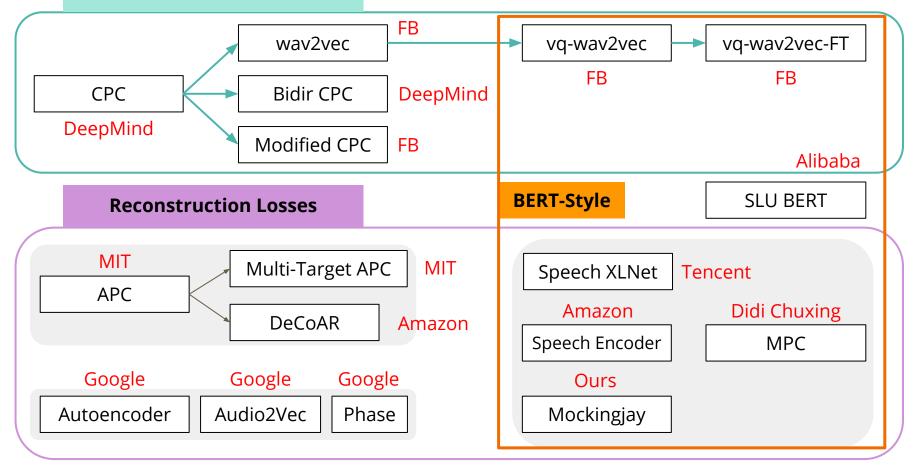
Related Works

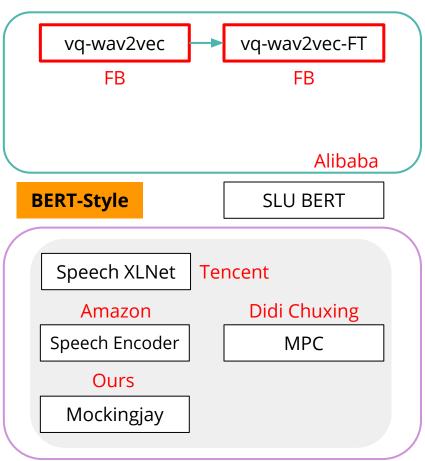


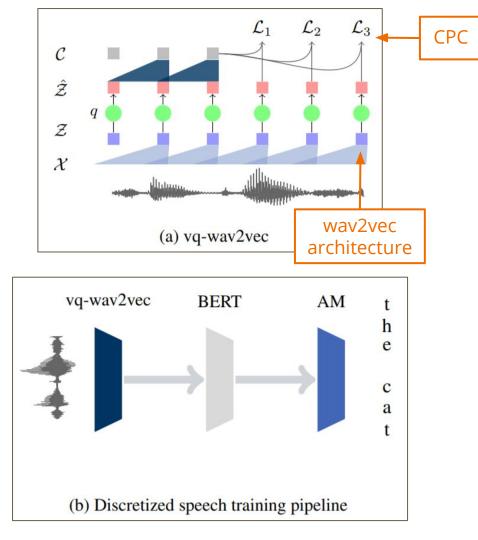
Audio2Vec

Phase

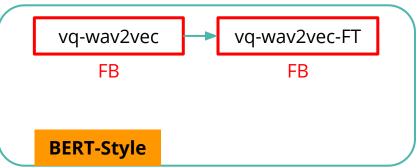
Contrastive Predictive Losses

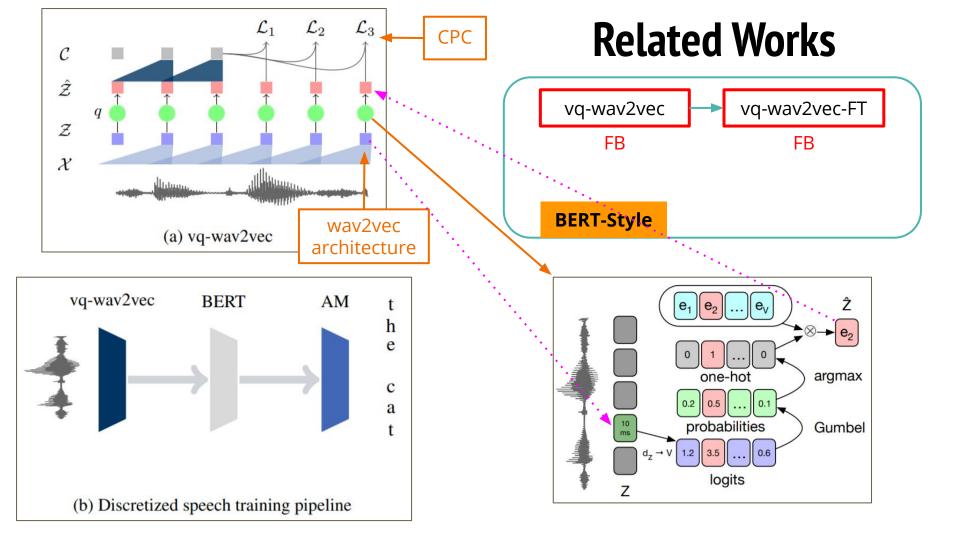


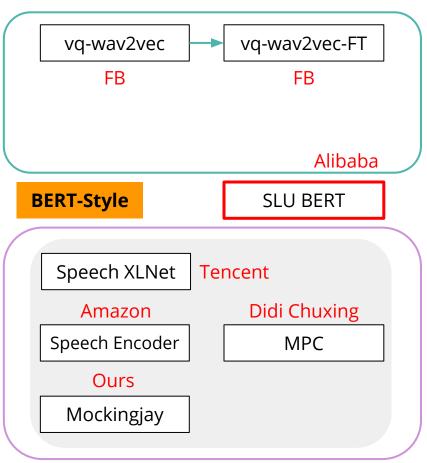


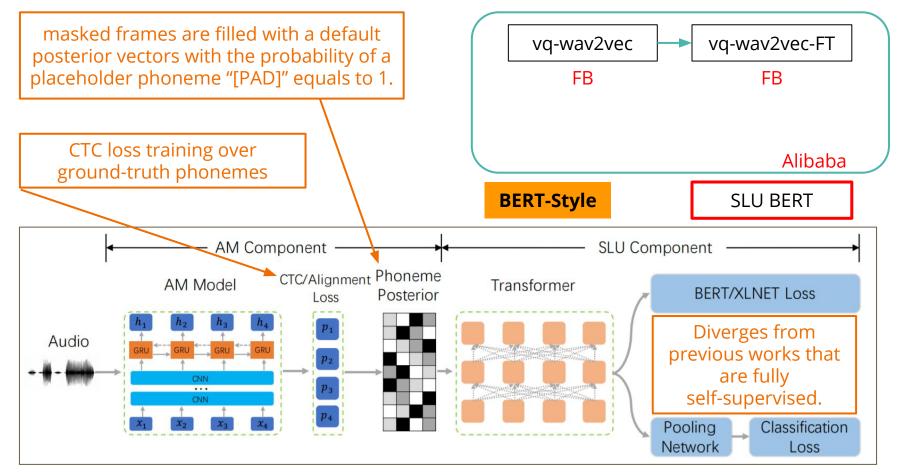


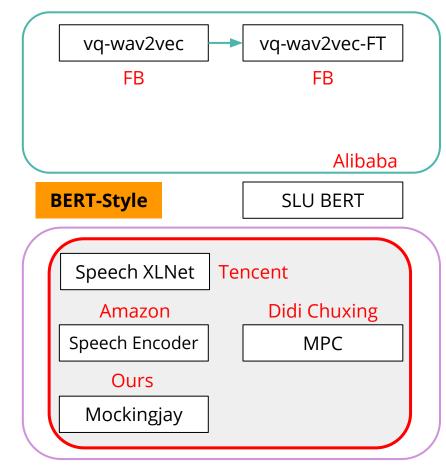






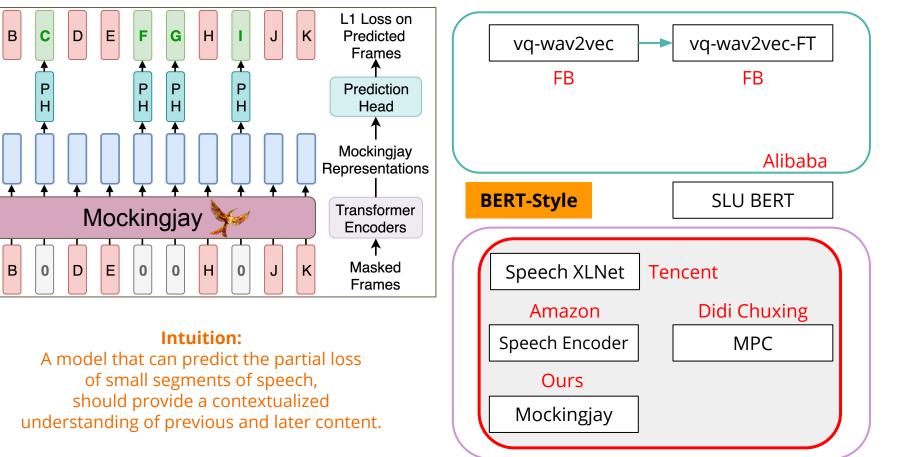




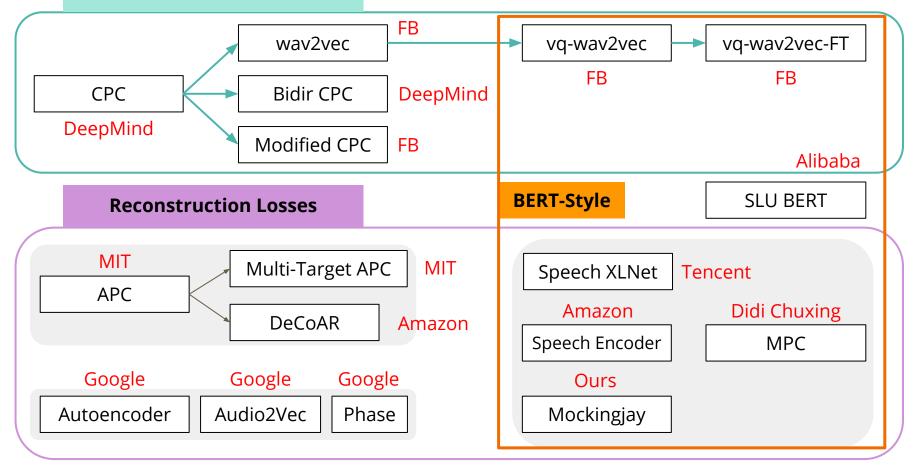


The Trend: All of these works emerges around October, 2019. All submitted to ICASSP 2020

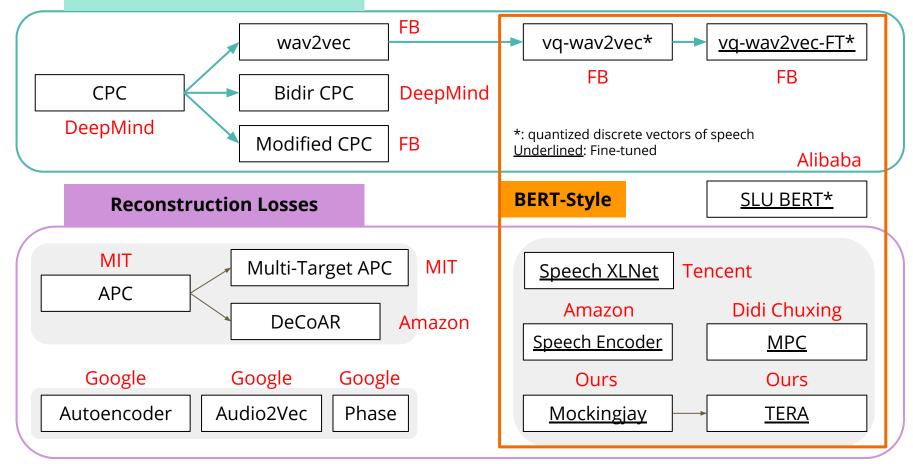
(Speech XLNet and MPC did not make it)



Contrastive Predictive Losses

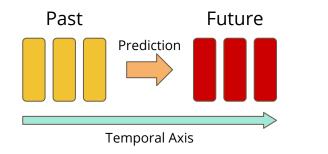


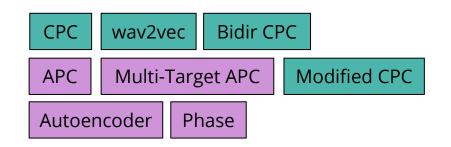
Contrastive Predictive Losses

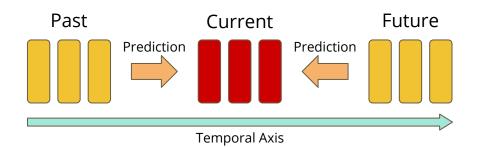


Related Work Summary

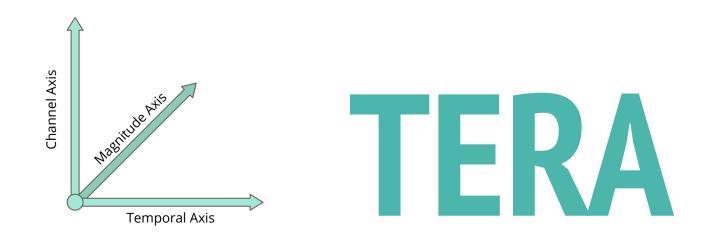
The design of auxiliary task fundamentally defines what the model learns!







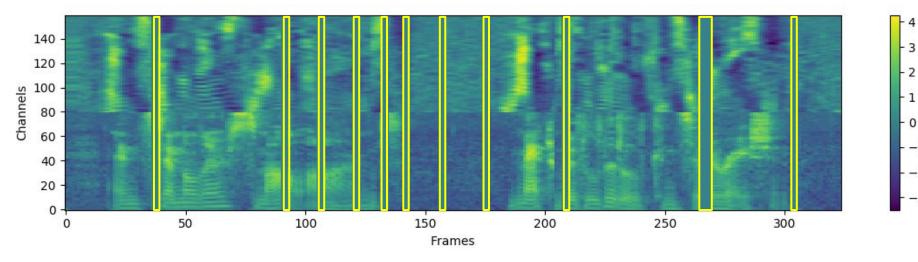




Transformer Encoder Representations from Alteration

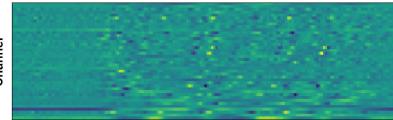
Extending Mockingjay to multi-target learning on three dimensions

Recall: we mask mel spectrogram on time axis



Consider fMLLR on 3 Axis:

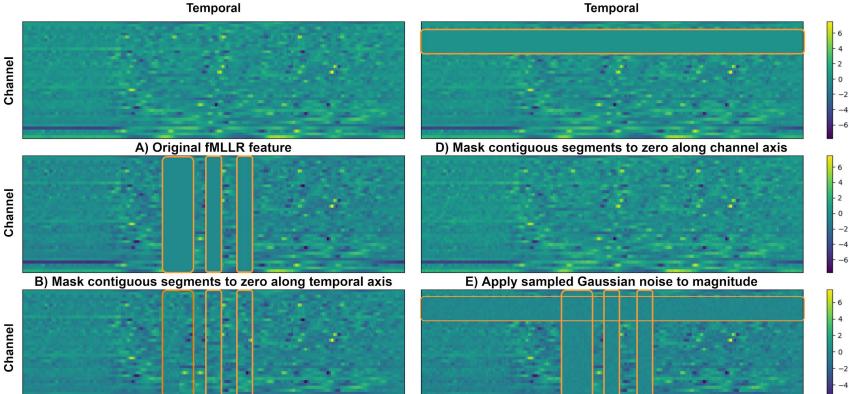
Channel



Temporal

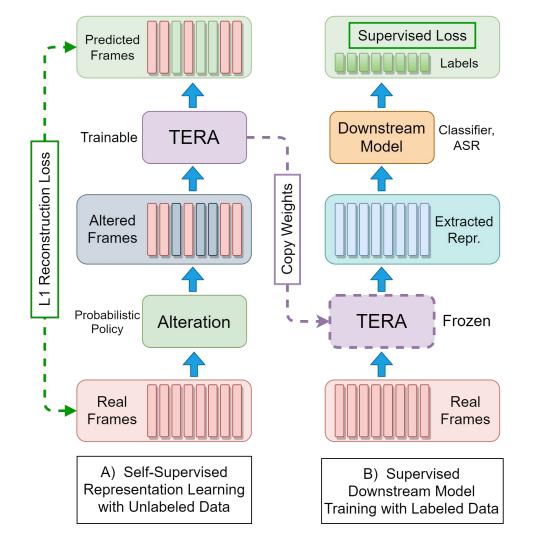
Multi-target Pre-training

Temporal

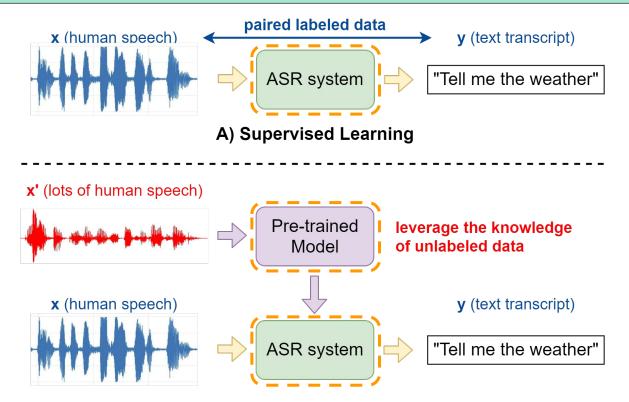


C) Replace contiguous segments with random segments

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

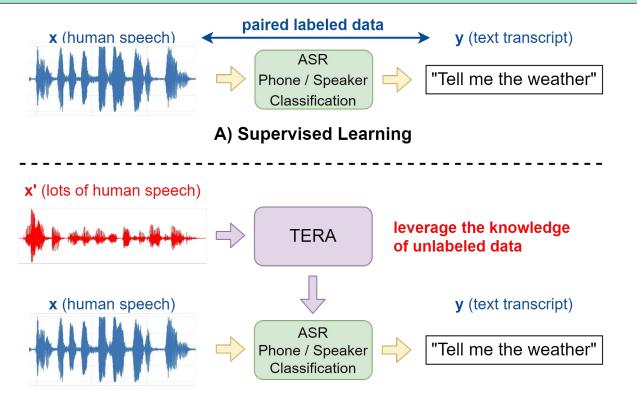


Recall: Self-Supervised Learning for Speech



B) Self-Supervised Learning for Improving Supervised Systems

Self-Supervised Learning: TERA



B) Self-Supervised Learning for Improving Supervised Systems

Progress: 10 pages, 90% Done

nais. 27 CPC with ASP: In the wardwar [9] paper, the CPC [8]

VQ speech representations. 5) Bidiractional CPC: The CPC loss has also been ex-

on mixture of auxiliary objectives that provide bio

are an improve to earning. A second secon

All of the baseline features (MPCC and fMLLR) use same ASR funnersork as the proposed features, hence the features of the second second

VI. CONCLUSION The conclusion over here

APPENDIX A PROOF OF THE PIEST ZONALAR EQUATION

E. Comparison with Other Representations

loss is used to pre-tails speech sepresentations E pose of speech recognition. Where a 5-layer co (CNN) encoder network and a 9-layer context in pre-trained with CPC loss on large amounts of

Extraction Factor

mining and downstream damants by evaluating pro-training profession. In the first stage, sport are solide with ASR on the TMHT [2] dataset, and show that by everyone [3] there and everyone the proposed method can help address the domain minimation to:

which out the other DMT (2) taken 1 and the data of the orbit-left (2) to an at subtracts. Therefore, programmers are subtracted to the orbit-left (2) to an at subtracts. The orbit-left (2) to an at subtracts the constraints of the orbit-left (2) to an at subtracts. The orbit-left (2) to an at subtracts the constraints of the orbit-left (2) to an at subtracts. The orbit-left (2) to an at subtracts the constraints of the orbit-left (2) to an at subtracts. The orbit-left (2) to an at subtracts the constraints of the orbit-left (2) to an at subtracts the constraints of the orbit-left (2) to an at subtracts. The orbit-left (2) to an at subtracts the constraints of the orbit-left (2) to an at subtract the constraints of the orbit-left (2) to an at subtract the constraints of the orbit-left (2) to an at subtract the constraints of the orbit-left (2) to an at subtract the constraints of the orbit-left (2) to an at subtract the constraints of the orbit-left (2) to an at subtract the constraints of the orbit-left (2) to an at subtract the orbit-left (2) to an at at subtract the orbit-le

IL EXAMPS VERSE IL EXAMPS VERSE wet proceed is the paper to use to example sent proceed is the paper to use to example the USE (Vision Vision) CPC (Vision the allow the table (Vision Vision) CPC (Vision the allow the Vision the Visio

A Representation Learning with Corectairle Learning A najor hearch of speech representation learning method pages, relative bandles for protectairle foreases on the Contractive Posterior Contexpondent and the series of the foreases on the Contractive Posterior Contexpondent in (B), and which we depresent the series of the series of the series of the series of the foreases on the Contexpondent in (B), and series of the Contexpondent in (B), and series of the se

a V-N we first study the phonetic context launed pre-training context. In other words, we use available combination that includes the time alteration object where $[0,\infty]$ is the richely the phononic counter largend $[0,\infty]$ combustions the hierarchite the time alteration of remains with fister and solution (assisters, Note, in 2) Learning Querker Entry with Channel Al-(20)), in core the dwe important space of types $[0,\infty]$, and $[0,\infty]$ and $[0,\infty]$ constraints of the largend Al-sel phonon largenda in the important space of types $[0,\infty]$. All Learing with Entry with the entry the OLLER frame all querker meganisms in these calculators. The large and the phonon-largenda in MDCC results in wave per-distribution of the entry o

climinates most of the pitch

Abrad Fumen

A Learnine Phonetic Content As expected, IMLLR [51] features superformed MFCC in one separability, under both fateur and 1-hidden classifiers.

R. Learning Speaker Identity

In evaluation with ASR, we not only not representations t tional Courses with Time Admention

C. DeciseSter yest ASS

Submitting to the IEEE/ACM Journal of TASL around summer

TERA: Self-Supervised Speech Representation Learning with Transformer Encoders Andy T. Liu, and Hang-yi Lee

radies a wilf-supervised speech representa-radie TEX, which stands for Transference base from Jhermann, hereit approximate offer hermafields. Record approximate offer hermafields for contractive house that con-tains and another the house that con-tains and another the house that con-

the original associal: Issues: from its altered rematespart. We use perhabilistic alteration periods to alter the input associate frame- ty altering configurous segments in both the temporal and channel issues which asymptotic potence magnitudes. The key insight is that The model can period it the original frames theory encoupled logat.	tations through tol' expected training videout flow use likels. In TRAN, we training the model is required to errol frames: from correspond frames. We insteaders to the enabling objective video's constraint (1) is not allo observed to the enabling objective video's constraint alteration, reconstraining frames. We applied in themesis, 'n magnitude, thrustone models which we ob- ferance magnitude. These antickey objectives can be all shown to the enabling encourse, the applied in all shown to the enabling encourse, the applied in all shown to the enabling encourse, the applied in all shown to the enabling encourse that a statistic resour- ting the enabling encourse of the enablishest of the chemically sampling formed a probabilistic of
Index Terms-Self-supervised, speech representation, represen-	policy to create random alterations.

The leg size behind all the anniusy objectives to it model can produce the original inners frame compty then it should perside a good segrenations of the inner many sector of the sector of the sector of the inner many sector of the sector of the sector of the inner many sector of the real kinetic sector of the sector of the sector of the real kinetic sector of the sector of the sector of the real kinetic sector of the sector of the sector of the real kinetic sector of the sector of the sector of the real kinetic sector of the sector of the sector of the real kinetic sector of the sector of the sector of the real kinetic sector of the sector of the sector of the real kinetic sector of the sector of the sector of the real kinetic sector of the sector of the sector of the sector of the real kinetic sector of the sector of the sector of the sector of the real kinetic sector of the sector of the sector of the sector of the real kinetic sector of the sector of the sector of the sector of the real kinetic sector of the sector of the sector of the sector of the real kinetic sector of the sector of the sector of the sector of the real kinetic sector of the sector of the sector of the sector of the real kinetic sector of the sector of the sector of the sector of the real kinetic sector of the real kinetic sector of the s

In the formation of the anathury introduction account proof and quark processing of the anathur proof and quark quark proof and quark quark

Representation Learning with Reconstruction construction Another recently enough branch of speech representia in lineary approach is do speech representiation to the section of the secti

actains to improve guarantization of the complex means the indiversionality of an adult neural has been added by the means and the mean added by the means and the mean added by the means the means the means added by the means the means the means added by the means added by the means the means added by the means added by the means the means added by the means added by the means added by the means the means added by the means added by

100 hr 400 hr 960 hr

auto representations with sch-supervision Leners vol. 27, pp. 686–604, 2020. [14] W. Wang, O. Tang, and K. Darawa, "U bidentinal speech encoders via made

 446
 223
 .

 443
 303
 644
 713
 678

 441
 721
 645
 723
 647

 442
 774
 660
 361
 659

 451
 773
 647
 664
 664
 1.5.1 Frame-wise phone classification results on LibriSpeech. There are 41 possible phonene classes, with 15 p sets all identical to the areas used in the CPC [B] Intentions. We present instage set accuracy (10) from lines in inde blidder have character accuracy areas varies and first voltacions and moved or areas variants.

AILE II: Frame-wise phone linear classification results

REFERENCES

KEPUBLINES K. Kewdomi, L. Nog, C. Dyr, P. Binnern, and A. van des Gerd, "Learning robust and multilegad speech representation," 2020. M. BYRD, A. Dords, 194, Marzir, and E. Dapon, "Unspected 7

a well across languages," 2021 I. Glass, "Generative pre-basising for speech with Inter-order," in (CAUP, 2023.

1. S. Schwidz, and M. Auli, "sp-warbur: Self-terning of disense speech spectrations," arXiv preprior FINLEALD, 2019. Well Burgssenamme in water up Chang, W.-N. Ens, H. Tang, and J. Ghan, "An ansupervised provide model for second appropriation legitims," in American de Woll Supervised and Science Ling, T. Lin, J. Infanor, and K. Kreddolff, "Deep contextualized preparations for sent-septroned speech recognition," 2019. "Site: Generative International Internation In addition, all of the above approaches demonstrate strong that extracting representations from a pre-trained model co-sistently improves performance over smilla features. n addition, all of the above approaches dem his section is organized as follow: In Section III-N describe our model architecture and its designs. Next, forcion III-II we describe the proposed multi-target pre-ing orburns. In which we introduce how multi-fold data $P_1 = (P_T \times L_a + W_c)$ h index location i, in J_T, w

We me multi-layer Transformer Decodes and multi-bead self-attention [25] to learn under the proposed TERA pre-training scheme. at IMLLR features supprised MPCCs ions for DNNHMM [5] hybrid speech mainly due to the speaker adaptation in documents behalf INMLIMM the rad increase larger in Figure [2], we begin with white an predict the partial larger of small segment of $M = 10^{-10}$ (small section with the rad in the size of small segment (small section set) and predict the partial larger of small section set (small section set) and predict section set (small section set) and set (small section set) are set (small section set) and set (small section set) are set (small secti abset. In this paper, we denote to learn hidd

Criters that will operate on those raid MLLR frames 7. It comparer vision [26], "Speci-Three Alveration: For the first objective, man situation, from Alveration: For the first objective, man situation, to pot and frame count, through the inconstruction of the 21 form 4 potential potential potential potential potential for 21 form 4 potential potential potential potential potential for 21 form 4 potential potential potential potential of a substantially, a certain processing or 4 (spec Tunes are bless to zero for a manimum of a substantial potential potential potential potential for a potential potential potential potential potential for a manimum of the potential potential potential potential for a manimum of the potential pot

Roperso Mattern	here	Average	Panine	Arrison	Fialse	Average	
CPC IE	10.1		1				
TERA busy time							
TEMA-base; time + channel							
			+- traditional features with linear classifier				

TABLE III: Steaker linear classification re-

edels		Erscore 6.12		Hencore 6.10	
CRU + TERADARE THE		611		6.00	
KORU = TERA-base: Eme = channel	8.22		8.27		
KRU + TERA-base: time + channel + mag					

					1.06
BOX BOX understimal understimal understimal understimal bidestimal bidestimal bidestimal bidestimal bidestimal	ion 2017 ion 21-5 ion 52-5 ion 52	Line 1 Fraze Soli 22.5 58.4 Boli 12.5 4.4 Boli 15.5 4.5 uniderstrongi 9.7 45.3 uniderstrongi 9.7 45.3 uniderstrongi 46.2 75.4 uniderstrongi 46.2 75.1 uniderstrongi 46.2 75.1 Materstrongi 46.3 75.3 Materstrongi 46.4 75.3 Materstrongi 46.5 72.3 Materstrongi 46.4 72.3 Materstrongi 46.5 72.3 Materstrongi 46.5 72.3 Materstrongi 46.5 72.3 Materstrongi 46.5 72.4	Law 1 (Trans.) Plane Bits 125 94 0.4 Bits 125 9.4 0.4 Bits 125 9.4 0.4 Bits 15.5 9.4 0.4 Bits 17.0 8.4.3 0.3 anthronous 9.7 8.3 0.5 milleproteini 6.2 2.5.1 6.0 Milleproteini 6.2 2.5.1 6.0 Milleproteini 6.1 7.1 73.8 Mileproteini 6.1 7.1 73.6 Mileproteini 6.1 7.2 6.5 Mileproteini 6.1 7.2 6.5 Mileproteini 6.1 7.2 6.5	December 2000 Tenant Tenant Tenant Tenant max 19.3 80.4 19.4 18.4 max 19.3 80.4 19.4 18.3 max 19.3 8.4 1.6 10.7 max 19.7 8.4 1.6 10.7 max 19.4 19.4 10.7 10.4 max 19.7 19.4 10.7 10.4 max 19.7 19.4 10.1 10.4 max 19.7 19.4 10.3 10.4 max 19.7 19.4 10.4 10.4 max 19.7 19.4 10.4 10.4 max 19.7 19.4 10.4 10.4	Definition Table Table <thtable< th=""> Table Table</thtable<>

2008 - 2 Signal Pi Mi M. Revo Reverse Enclose

Menne, Andreas Barger, Kary M. (2007) 2017 (2017) and comparison compared and an impact on C. E. M. 19 and comparison compared and an impact on C. E. M. 19 C. M. (2017) 2017 (2017) 20	and Super Processing, vol. 23, no. 11, pp. 166–164, 113, Republice, Vol. House, L. Kalano, T. Solma, and Super Proceedings of the Super Science of Super- information, 2014 IEEE International Conference on and Super Proceedings (2016), pp. 62, pp. 61, 10 and Super Science on Superior Science on Super- International Conference on Superior Science, Science Science, 2014, Science, Science, Science, Science, Science, Science, Science
N.O. Phon, T.S. Npayer, J. Nadoux, M. Miller, X. Miller, and	

DESIGN TRANSACTION OF ADDR. OTHER AND LANGUAG PROCESSIN	G, TOL, NOL NI, MERITIN VIEW 5
Temporal	Temoral
A Count MLR hours	C Mark contracts in the second second sec
A: Degrad MCCR Houses	C) Mask computer segments to zero arong channel zon
	a de la
(it. Mask contiguous segments to zero along temporal axis	Et Apply sampled Gaussian noise to magnitude
C) Replace contiguous segments with random segments	F) Combining the alterations in (E), D(, and E)
is highlighted in comps.	n applied for the proposed auxiliary objective. The altered part
a makes channel alteration which Π_{i} from $\{0,1,\ldots,0\},\ldots,0\}$ from the analysis of the fractions is the matrix $(0,1,\ldots,0),\ldots,0$ for the fraction of the matrix $(0,1,\ldots,0),\ldots,0$ for the second	ner men and 62 vielane. Is wich 7 hm distantion $f_{\rm eff}$ is a bound to get 7, where $f_{\rm eff}$ is a bound to get 7, where $f_{\rm eff}$ is a bound to get 7, where $f_{\rm eff}$ is a bound to get 7, where $f_{\rm eff}$ is a bound to get 7, where $f_{\rm eff}$ is a bound of the first the first firs
fixed amount of bins R_{\pm} . Hence following the work of [18] that also uses accountic features with 40 frequency bins, we set the maximum channel alteration width W_C to 8 channels.	three autiliary objectives can be used together as a miniture. We use different combinations of auxiliary objectives to learn meaningful speech representations, as shown in Figure 20.

ormine: Additionally, we introduce the Mapshafe Attention: Additionally, we success use ad objective of magnitude alteration by applying sampled assess noise to argument the magnitude of input sequences. we do not always after the magnitude, but initial diplorees use summum is fixed percentage of time. We denote this percentage of time. We denote this percent and the model, and to prove to the model, and to $model' = \frac{1}{2}$ from the model distribution N with conditi the original frame

most important underbring factor of speech is mole P, amount of random index end for the if v > -0.8 and v < 0.9 then for all h in f_T do $k \leftarrow nplace Wy number of consensive fus$ $in <math>\mathcal{T}$ with with-in uticance randomly samp pply on top of provious alterations, T +-In addition to measuring phone linear separability abasted with classifiers with a single hidden layer if apply magnitude distance then if apply time or channel alteration then Apply on top of previous alterations, $\vec{X} = \vec{x}$ and \vec{x} value v from uniform distribution 22(0, 3 % them Sample magnitude vector \overrightarrow{T} from N(0,0.2)Apply noise $\vec{x} \leftarrow \overrightarrow{T} + \overrightarrow{T}$ to alter magnitude Do nothing, 2 +- 2 O. Hubble DENTROY & OF Street Joid DNNIBH AM. A.S. A.S. $L_{max}(\theta_{trave},\theta_{max}) = \left\| \vec{\mathcal{T}} - P_{max}(\mathbf{T}_{max}(x)) \right\|_{1}$

All inference time, meaningful representations, asize where all inference time, meaningful representations, asize where well MLIR features are used as ispace (Figure 7), while the first Dataset X, and making investing Transition from the time of the time and is in image.

Alexeridan 1 The TERA Multi-tarent Pre-tunit

contage Py, tin width Hey may

C. Incornerative with Downspeare Test-

Marina Marina	here	Average	Panen	Arrison	Fialse	Average
	10.1					
		951				
sections - channel		98.5	99.4	98.8		95.8
tore; time + channel + mag				98.5		
	17.6	16.5		strend droke	or which the	or classifier

eeck. There are 251 possible classes, we use identical training	TABLE V: Comparison of different a
ane wise linear classification accuracy (%), and attenuor-wise	presented, with the same phone set, trai
does neer time.	auxiliary objective of time + channel + s

	Edit-OC () Edit-OC ()	550 hr	25 14	14.96	9.AL A.29	CON + TO TA Balls
	wwiten lags C	963 la	100 14		4.90	IGRU - PRANE (1) IGRU - MILLE (1)
		Non	138.74	X.56		
rates	BORU + OALR BORU + TRAJER	Conc.	100 14	A-53	425	wardnes of wardnes of
and of	BORD + TERA medium BORD + TERA (seal	953 kr 953 kr	100 hr	ATT 835	4.05	ICHU - TEXA Sunc
	BORU + TERA-timer BORU + TERA-timer	900 Br 900 br	100 14	8.40	600	ICHU + TERA-Indu ICHU + TERA-Indu
	TABLE VE: Comparis	n d ne	and some	ch mre	owninking.	ICHU - TERA slaps
1 1	approaches for ASR.	VI needla	are from	n trainis	IN AN ASR	TABLE IX: Comparison

som in op to foreign spreadures, transit and tamp pro-trained robot We report ALS word error entry (WER). THMIT (2). At the pre-training data are from LibdSpeech () at WER after rescoring on the Libridynech (6) are clean. If not specified otherwise. We report scalar for phone shows in services, Build or (CC), (1), work-clean.



 Inter
 100 ls
 560 ls
 960 ls

 Linear
 I Hidden
 Linear
 Concut
 Linear
 I Hidden

 64.6
 72.5
 65.5

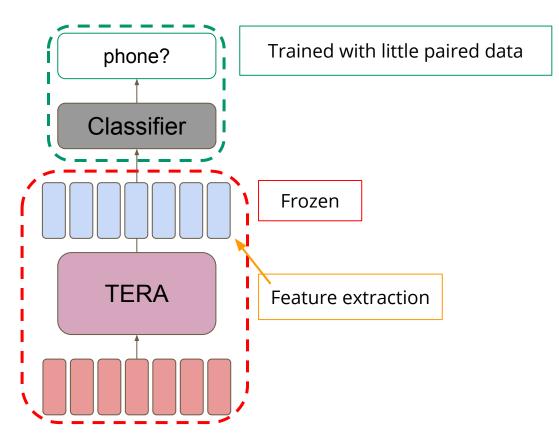
etwork depth and size. LibriSpeech (6) frame-wise phone classification moults as a set, and test set as in Table [] and []]. All of the TERA models use the combine

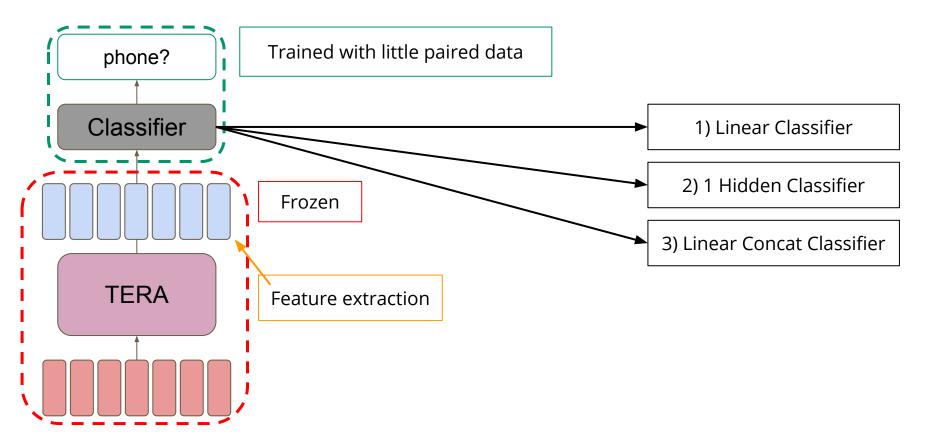
rs in part of the ASR system. Internation furgine this (WER) and WER after

F	1.00.0		ajii 744	in Arm	ope Plain	- Ave.	100		Ray
	10.4	1 .	-	-			_		100
	68.4								
	78.8								
	93.6	- 98.							
	95.9	95		0 98	5 99.4	- 98			
	17.6	14							
	9.4								
h	1 E R	el Sace	A Dest	am 151 m	ouble class	-	in Mandalah	mining	TABLE
					cution new				provention
			TRA OVER 1						endiny
-									
	1		10 hr	1	60 br		60 hr		19692
		ALE	Encore	GER	Исколе	WER	Rescore.		Table C
-		K.18.	6.12	1.8	6.00	1.8.1	8.99		Balle C
									10.00
									DICIO
		0.8	4.01	8.29	6.00	8.9	6.00	1	
									- 608U
					or formance				1000

	Links	THORE			938	
TERA-base time						
	651		95.9	99.2	8.32	6.61

Invariage (2017). 2020, pp. 7999-2005. Barris T, Barlos M, Wang L, Salta J, Linda J, and Barris T, Barlos M, Wang L, Salta J, Linda J, and Barris T, Barlos M, Wang L, Salta J, and Barris T, Barlos M, Salta J,	hers are in %.	
emann, F. Mohlardi, Y. Qian, F. Solwann, J. Solwoler, G. Saron, K. Vardis, "The kind reconstruent codilit," in UEEE 2011. HELL, 2005, "Here reconstruent web focustions web focustion of the	manalay (PLLIP), 2020, pj: 798-8000. WH: 28 Polici, N. Osnopo, and Y. Kunyin, "Light pared ands for payers to cognition," <i>IEEE Transactions on Payarging Depaymental Depayers vol.</i> 2, no. 1, pp. 302–302, pp. 302– 000, pp. 302–302, pp. 302–303, pp. 302–303, pp. 302– n., "Antoneous Bieldhood linear transformations (of Ruma Annel Organiza," Company, queriel & Bayangie, no. 12, no. 1, pp. 988. A. Ghoshad, G. Bondame, L. Burger, O. Gurshel, N. Ganes, Verlag, "The Role Royal Rospingue and Salesanger, C. Gurshel, N. Ganes, P. Veller, "The Role Royal Rospingue and Role and Depayer, Salesanger, Venge, "The Role Royal Rospingue and Role and Role and C. Salesa, Venge, "The Role Royal Rospingue and Role and Rel 1002, 2017).	[41] X. Zighidow, N. Usanier, L. Kokkowa, T. Schutz, G. Symanez, and E. Dapon, "Learning Mathemis from one spools for phone recognition," 2019 IEEE Astronautional Conference on Annalistic Sported and Space Proceeding (COSIN), Apr. 2018, [10] [2018]. and Mathematical Astronomy (COSIN), Apr. 2018, [10] [2018]. [41] L. Solt, "Prime Recognition with NeuralIstical Conference on Apr. proceeding astrocky," (EVDAVIE Journal on Audo, Species, and Marc Proceeding)





Pre-train	1()0 hr
Representation	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

Pre-train	100 hr		
Representation	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	

Outperformed CPC

Using more objectives also improves performance!

Baseline feature was outperformed by TERA features.

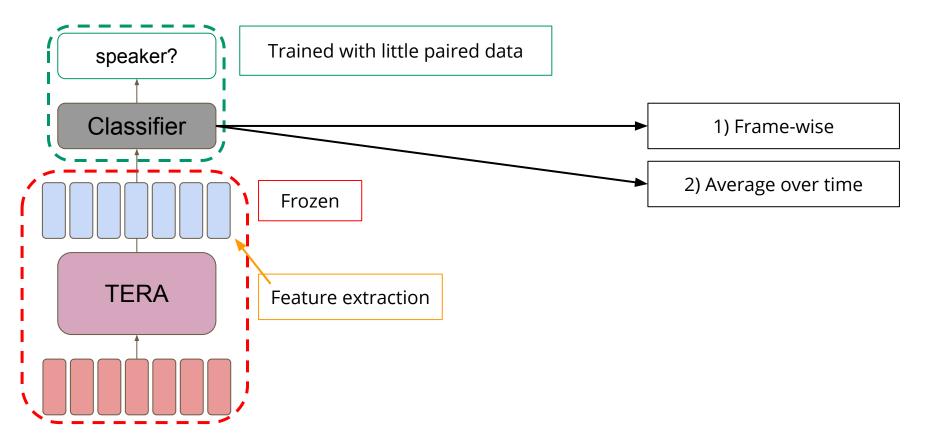
Pre-train	100 hr		46	50 hr	960 hr			
Representation	Linear	1 Hidden	Linear	1 Hidden	Linear	1 Hidden		
CPC [8]	64.6	64.6 72.5 -		-				
TERA-base: time	64.3	76.8	64.4	77.0	67.0	79.1		
TERA-base: time + mag	64.1	77.1	64.5	77.3	64.7	77.8		
TERA-base: time + channel	65.2	77.4	66.0	78.1	65.9	78.5		
TERA-base: time + channel + mag	65.1	77.3	66.3	78.3	66.4	78.9		
MFCC	39.7	59.9	\leftarrow traditional features with linear classifier					
fMLLR	52.6	68.4						

More pre-training data increases performance.

Having more alteration is like having more data.

Pre-train	1(00 hr	46	50 hr	960 hr			
Representation	Linear	1 Hidden	Linear	Linear 1 Hidden		1 Hidden		
CPC [8]	64.6	72.5			2			
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	\leftarrow 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.



Pre-train	10	0 hr
Representation	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

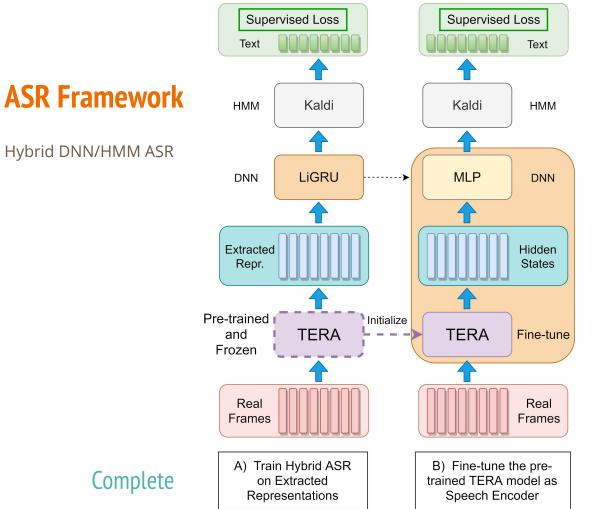
Baseline feature fails to encode speaker information.

Pre-train	100 hr]
Representation	Frame	Average	
CPC [8]	97.4	-	Outperformed CPC
TERA-base: time	68.4	96.1	
TERA-base: time + mag	70.8	96.1	Using more objectives also
TERA-base: time + channel	93.6	98.5	improves performance!
TERA-base: time + channel + mag	98.9	99.2	
MFCC	17.6	10.8	
fMLLR	0.4	2.6	

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

Pre-train	10	0 hr	46	0 hr	9	60 hr			
Representation	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 factures with linear classifier						
fMLLR	0.4	2.6	\leftarrow traditional features with linear classifier						

Pre-training on more data also gives benefit!



In Progress

Mixtures of auxiliary objectives - ASR

Pre-train	10)0 hr 📕
Models	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	100 hr		46	60 hr	960 hr	
Models	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

Pre-train	100 hr		46	60 hr	960 hr	
Models	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	100 hr		46	60 hr	960 hr	
Models	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

Pre-train	100 hr		46	60 hr	960 hr	
Models	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 C	Phone Classification		Speaker Recognition		Recognition
Representation	Tie-train Context	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	
Representation	Tie-train Context	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	
Representation	Tie-train Context	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	
Representation	Tie-train Context	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	
Representation	Tie-train Context	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	10	100 hr		360 hr		60 hr
Representation		#Param	Linear	1 Hidden	Linear	Concat	Linear	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	00.0	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!



Q&A