Self-Supervised Representation Learning for Speech Processing

Abdelrahman Mohamed
Hung-yi Lee
Shinji Watanabe
Tara N. Sainath
Karen Livescu
Shang-Wen Li
Shu-wen Yang
Katrin Kirchhoff

May 23, 2022

Outline

- Historical Context of Representation Learning
- Speech Representation Learning Paradigms
- Benchmarks for Self-Supervised Learning Approaches
- Analysis of Self-Supervised Representations
- From Representation Learning to Zero Resources
- Topics beyond Accuracy
- Toolkits for Self-Supervised Speech Representation Learning

What is SSL?

Tara N. Sainath

Types of ‘Learning’

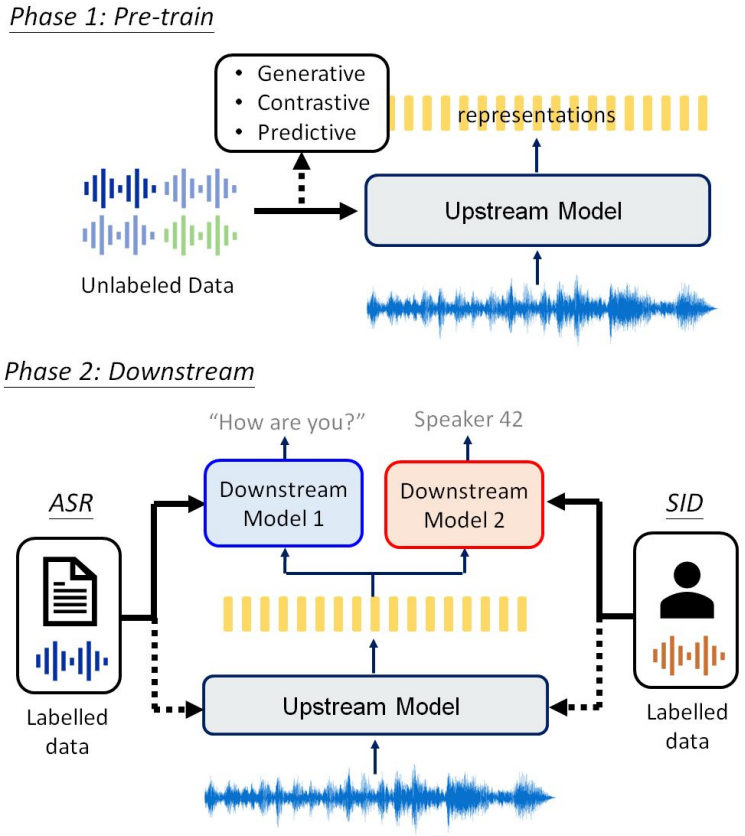
- **Supervised** learning: use labeled data
- **Representation Learning**
  - **Unsupervised** learning: discover patterns in data without pre-assigned labels
  - **Semi-supervised** learning: use a small number of labeled samples to guide learning with a larger amount of unlabeled data
  - **Self-supervised** learning (SSL): uses information from input data as the label to learn representations useful for downstream tasks

→ SSL is the focus of this tutorial.

SSL Framework

Two stages in the framework:

1. Use SSL to pre-train an upstream model
2. Downstream task uses the learned representation from a pre-trained model (frozen) or fine-tune the pre-trained model using supervised data



What Type of Representation?

We want a to learn a representation that is

- Disentangled
- Invariant
- Hierarchical

so it can be beneficial for downstream tasks.

Historical Development of SSL

Tara N. Sainath

First Wave: Clustering and Mixture Models

- Early works used simple clustering methods
- Gaussian Mixture Model / Hidden Markov models (GMMs/HMMs)
- Extracting features from generative models

Second Wave: Stacked Neural Models

- Neural models allow for more diverse modeling of input signals compared to GMMs.
- Techniques, often applied to vision and NLP and inspired speech, include:
  - restricted Boltzmann machines (RBM)
  - Denoising autoencoders
  - Noise contrastive estimation (NCE)
  - Sparse coding
- Higher capacity networks achieved by building ‘deep’ networks with multiple layers.

Third Wave: Learning Through Pre-text Task Optimization

- Learn networks to map the input to desired representations by solving a *pre-text task*, with the following characteristics:
  - All layers are trained end-to-end to optimize a single pre-text task
  - Deep networks with many layers are used
  - The representation model is evaluated on many tasks
- The third wave looks at designing a pre-text task, which allows the model to efficiently use knowledge from unlabeled data.
  - Generate an object from partial information
  - Use previous tokens in the sentence to predict the next token
  - Contrastive learning

Speech representation learning paradigms

Abdelrahman Mohamed

What makes speech representation learning unique?

- Speech inputs have a variable number of lexical units per sequence.

What makes speech representation learning unique?

- Speech inputs have a variable number of lexical units per sequence.
- Speech is a long sequence that doesn't have segment boundaries.

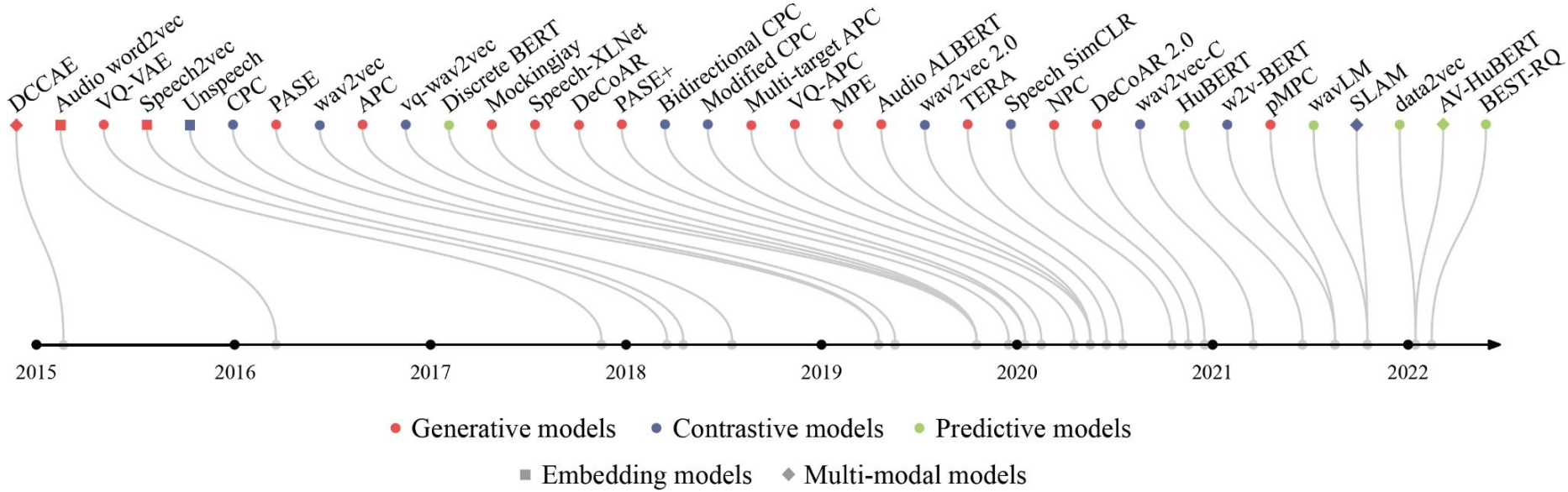
What makes speech representation learning unique?

- Speech inputs have a variable number of lexical units per sequence.
- Speech is a long sequence that doesn't have segment boundaries.
- Speech is continuous without a predefined dictionary of units to explicitly model in the self-supervised setting.

What makes speech representation learning unique?

- Speech inputs have a variable number of lexical units per sequence.
- Speech is a long sequence that doesn't have segment boundaries.
- Speech is continuous without a predefined dictionary of units to explicitly model in the self-supervised setting.
- Speech processing tasks might require orthogonal information, e.g., ASR and Speaker ID.

Speech representation learning methods



Speech representation learning methods

Contrastive approaches

Speech representation learning methods

Contrastive approaches

Predictive approaches

Speech representation learning methods

Contrastive approaches

Predictive approaches

Generative approaches

Speech representation learning methods

Contrastive approaches

Predictive approaches

Generative approaches

Contrastive Predictive Coding (CPC)

van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC

van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC

- The first successful representation learning approach for speech data.

van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC

- The first successful representation learning approach for speech data.
- It triggered lots of research in speech representation learning.

van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task

- Distinguish correct (positive) samples from wrong (negative) ones.

van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task

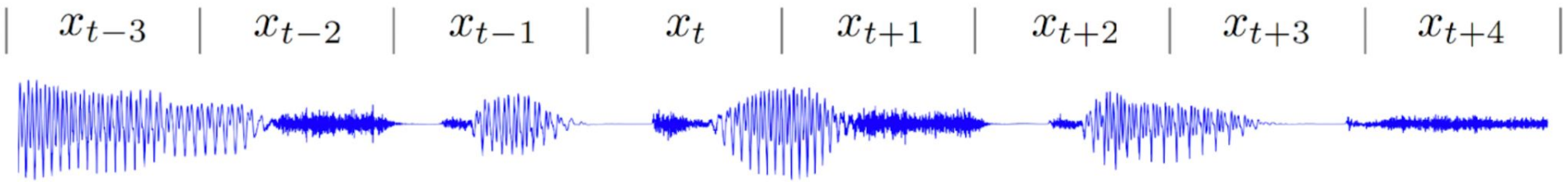
- Distinguish correct (positive) samples from wrong (negative) ones.
- **But, how do we choose positive and negative examples?**

van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task

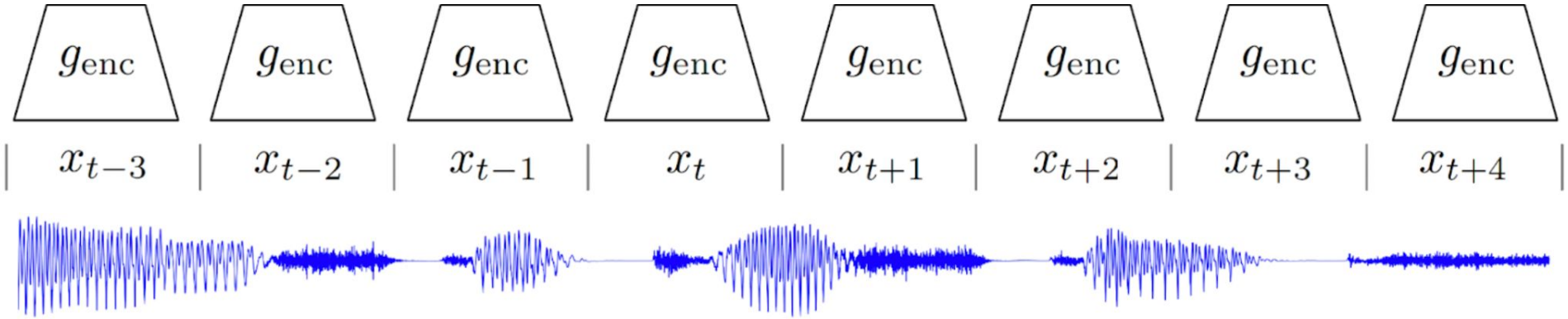
van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task



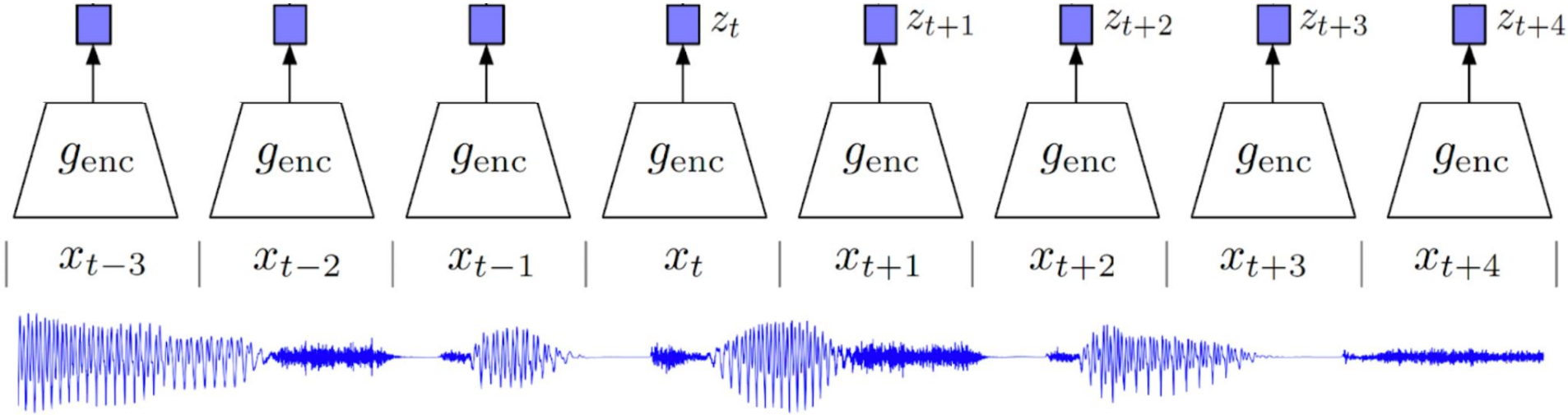
van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task



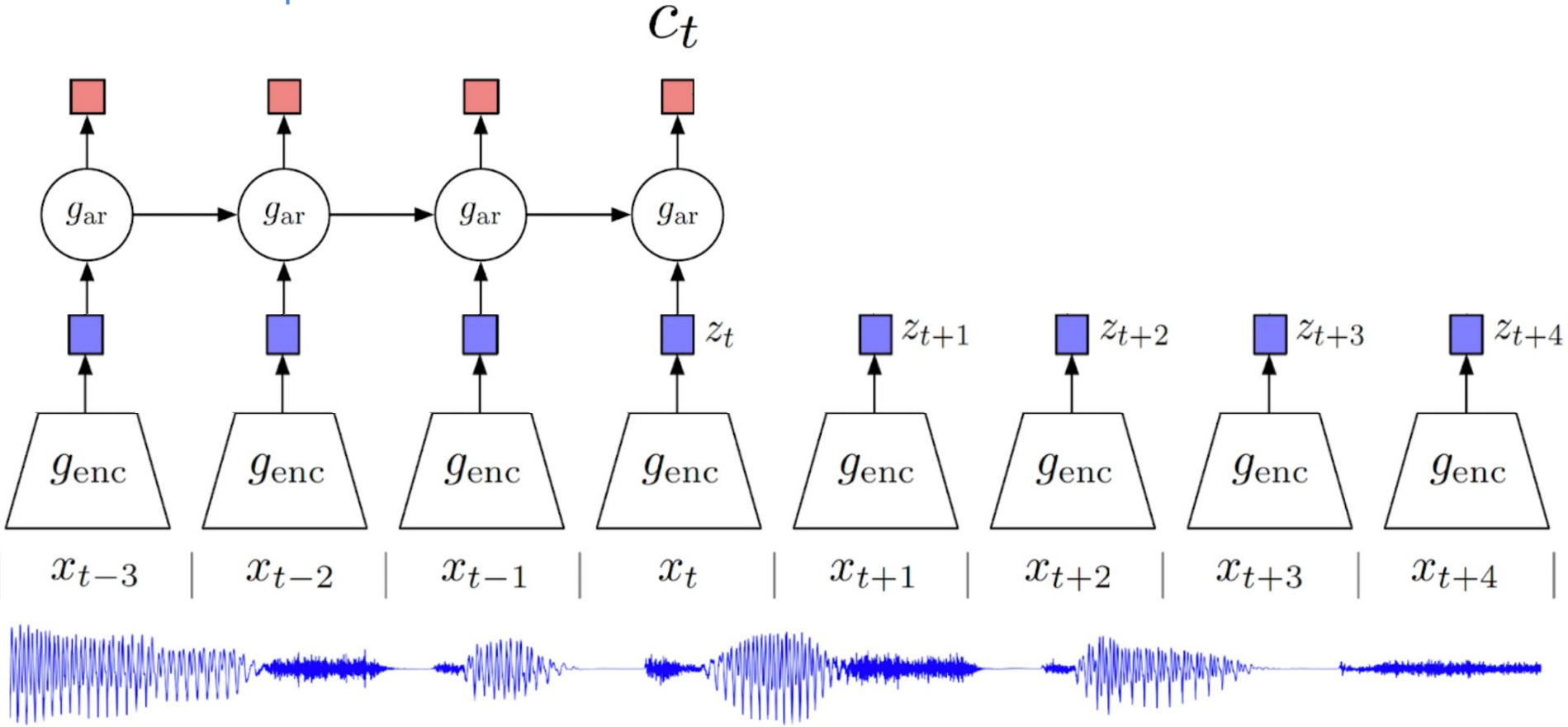
van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task



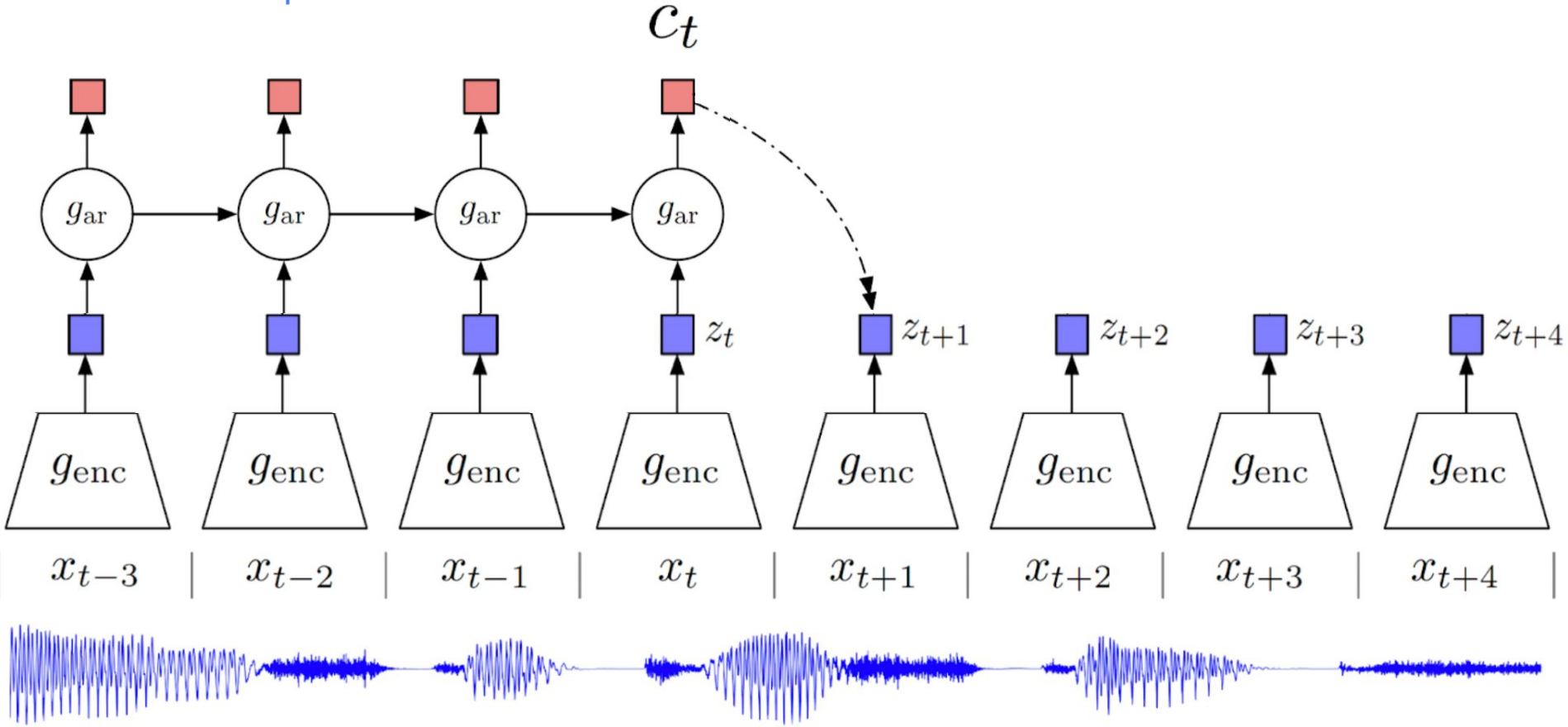
van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task



van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

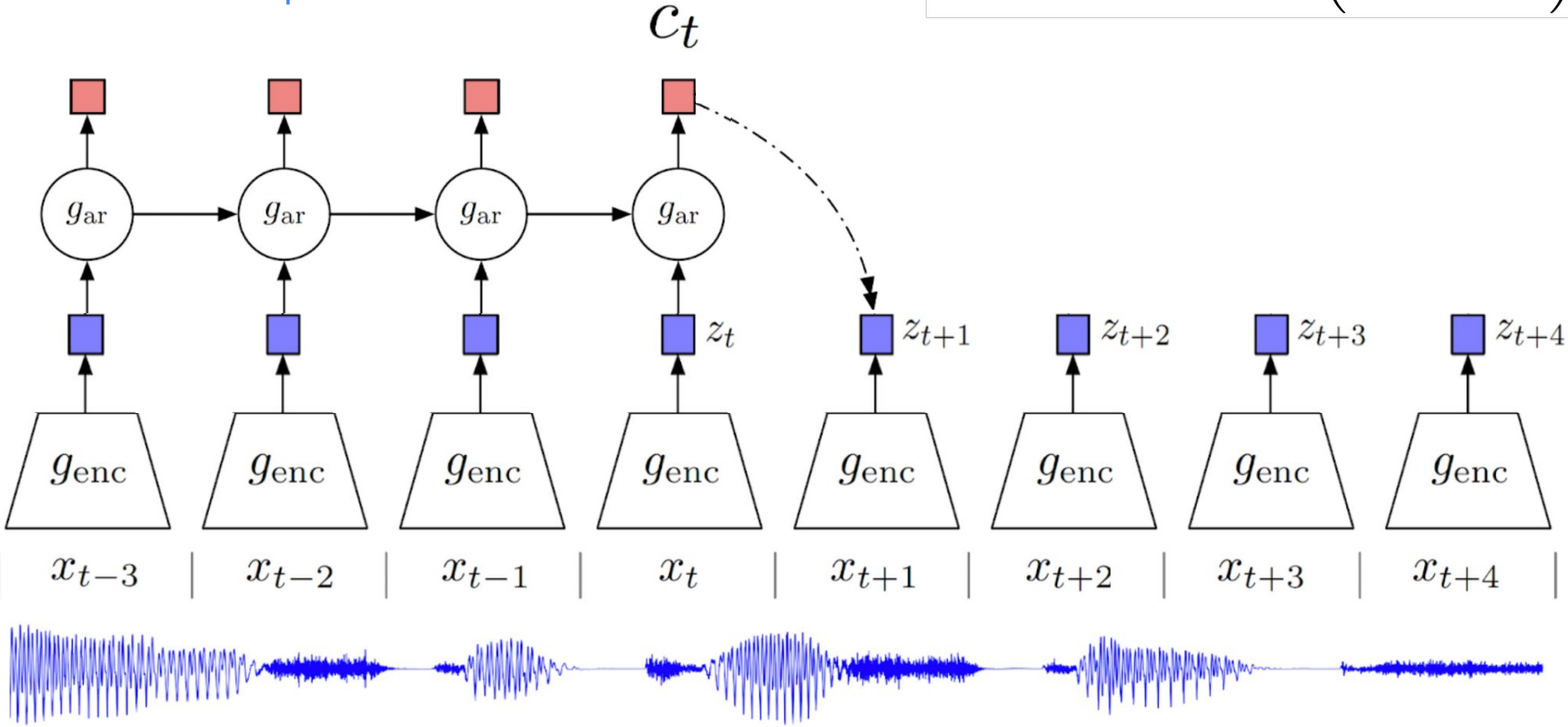
CPC: The pretext task



van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

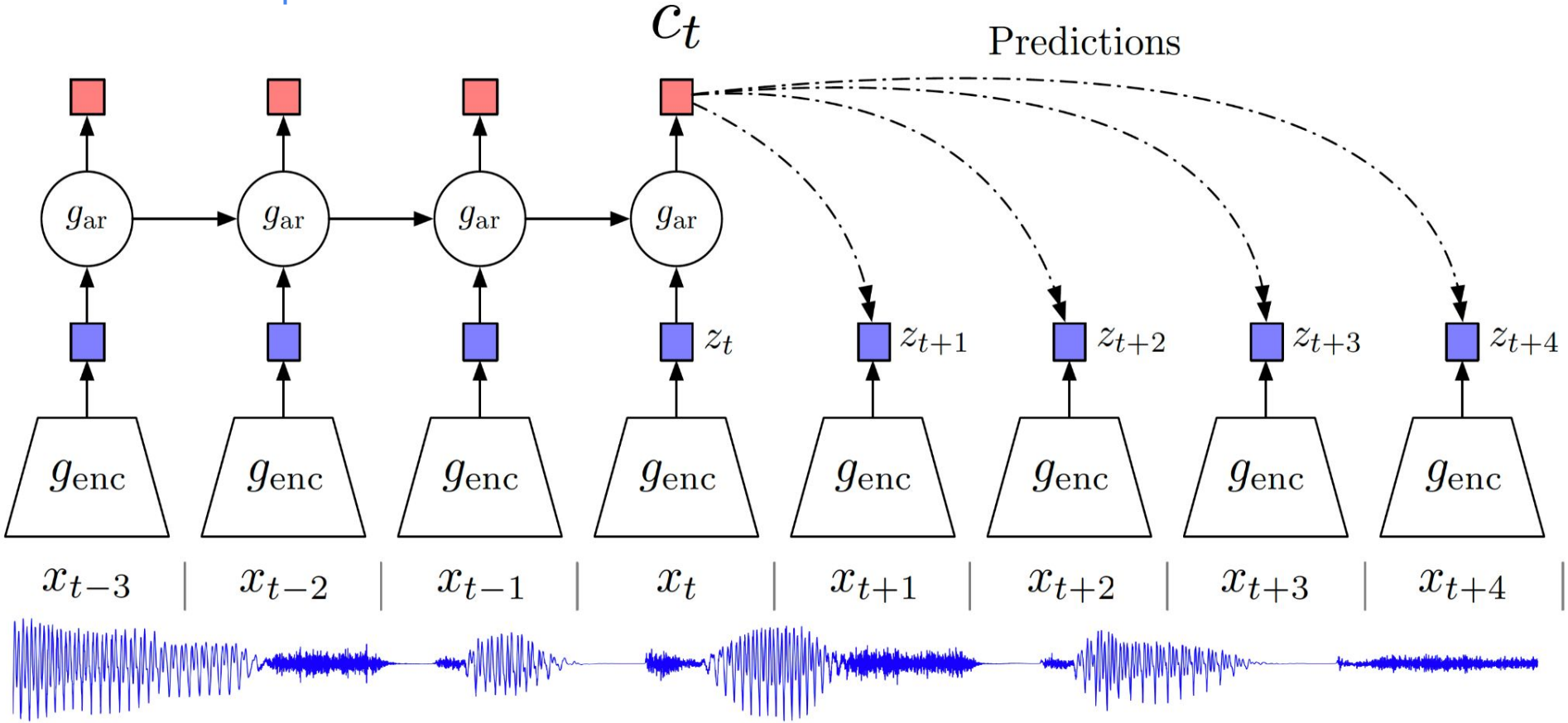
CPC: The pretext task

$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$



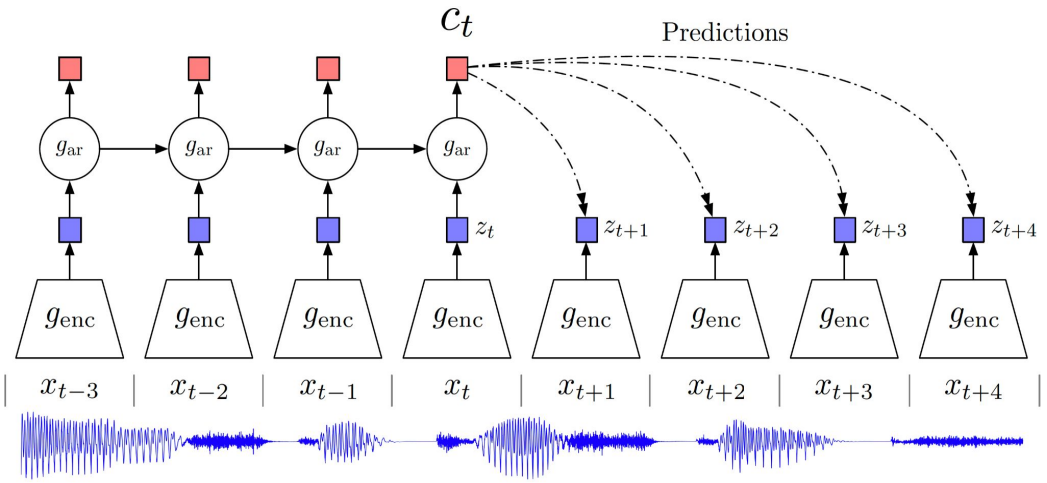
van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task



van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

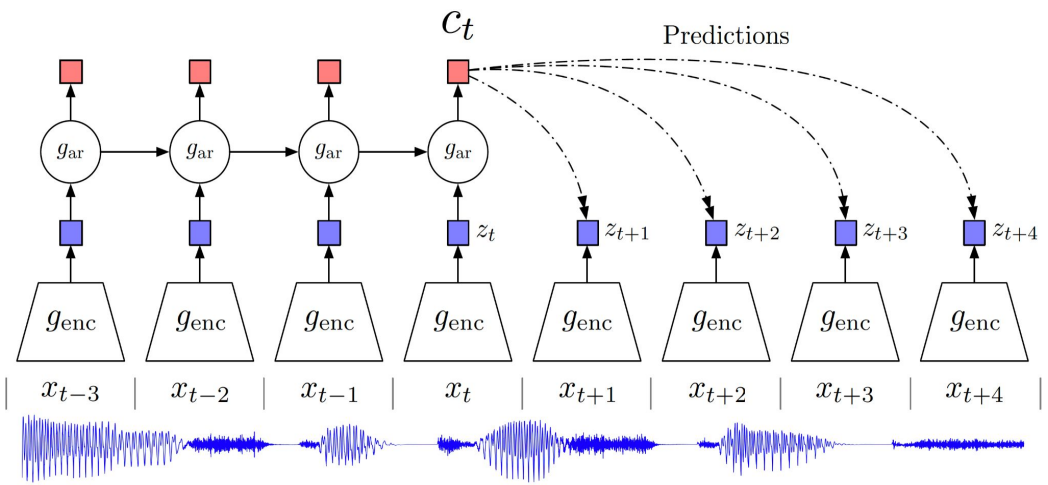
CPC: The pretext task



van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task

$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$

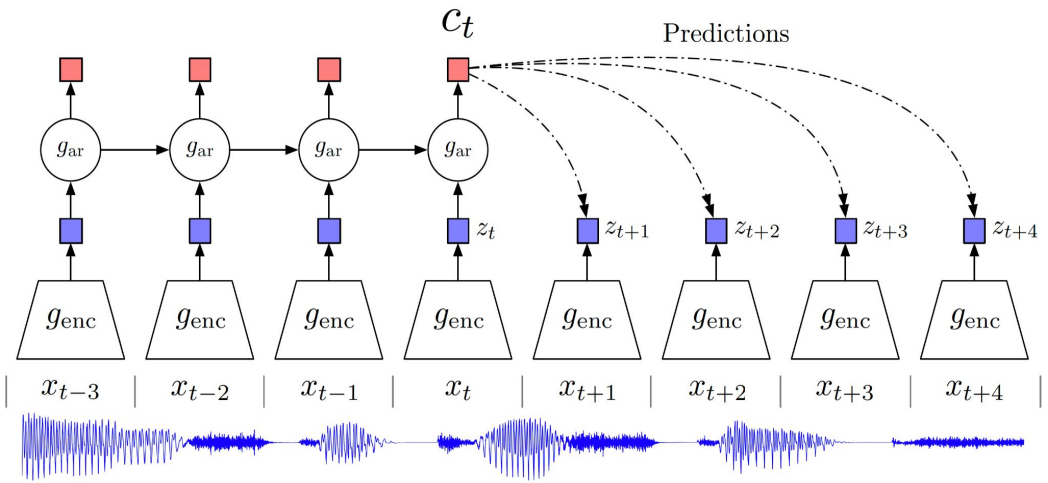


van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task

$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$



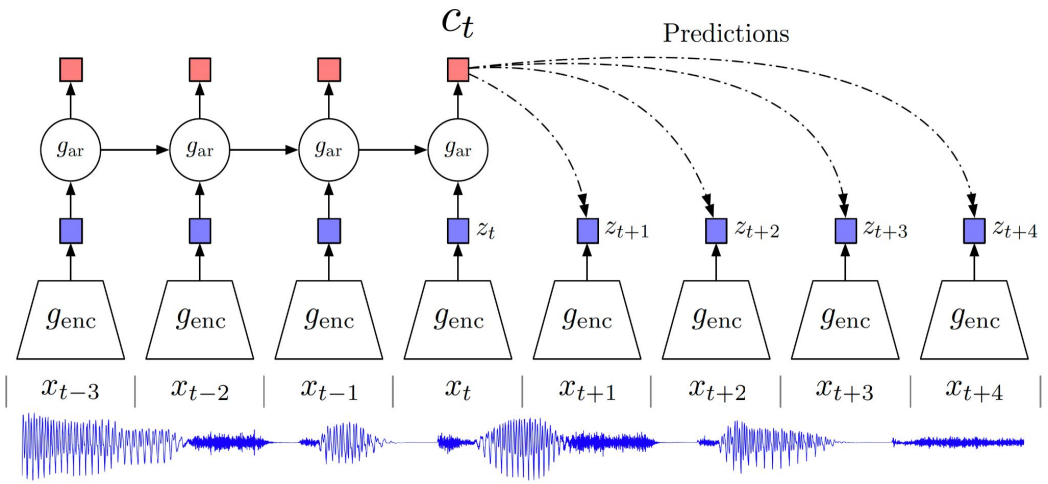
van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task

$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$

- InfoNCE maximizes the mutual information between the input signal and the learned latent variables C.

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$



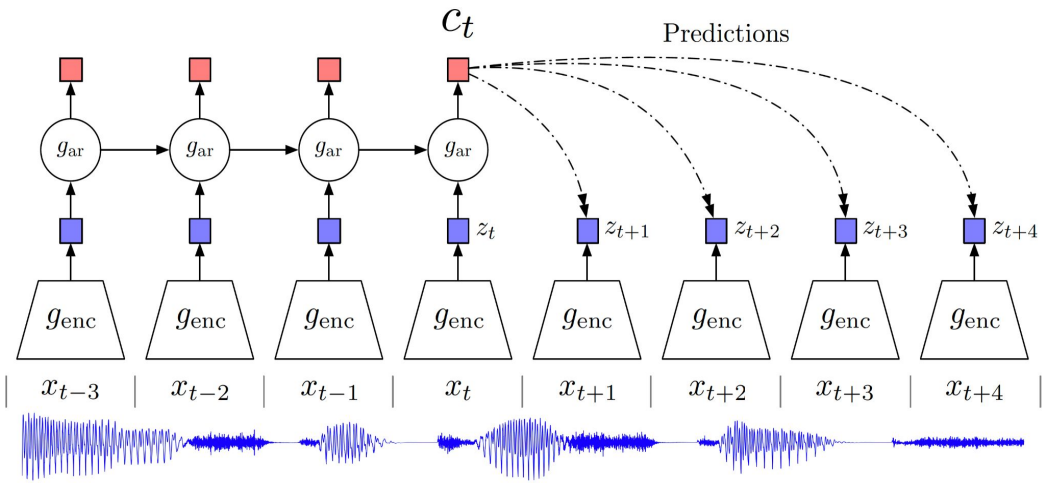
van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: The pretext task

$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$

- InfoNCE maximizes the mutual information between the input signal and the learned latent variables C.
- Strategies for sampling negative and positive examples determine the nature of representations, e.g., whether they are good for ASR or Speaker ID.

$$\mathcal{L}_N = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$



van den Oord et al, 2019 “Representation Learning with Contrastive Predictive Coding”

CPC: Extensions

- The CPC approach inspired many follow up work

CPC: Extensions

- The CPC approach inspired many follow up work:
  - With better architectures and normalization for stable training.

Schneider et. al., 2019 "wav2vec: Unsupervised Pre-training for Speech Recognition"
Kawakami et. al., 2020 "Learning Robust and Multilingual Speech Representations"
Rivi`ere et. al., 2020 "Unsupervised pretraining transfers well across languages"

CPC: Extensions

- The CPC approach inspired many follow up work:
  - With better architectures and normalization for stable training.
  - Bidirectional autoregressive components.

Schneider et. al., 2019 "wav2vec: Unsupervised Pre-training for Speech Recognition"
Kawakami et. al., 2020 "Learning Robust and Multilingual Speech Representations"
Rivi`ere et. al., 2020 "Unsupervised pretraining transfers well across languages"

CPC: Extensions

- The CPC approach inspired many follow up work:
  - With better architectures and normalization for stable training.
  - Bidirectional autoregressive components.
  - Which investigates the multilingual transfer of representations.
  - ….

Schneider et. al., 2019 "wav2vec: Unsupervised Pre-training for Speech Recognition"
Kawakami et. al., 2020 "Learning Robust and Multilingual Speech Representations"
Rivi`ere et. al., 2020 "Unsupervised pretraining transfers well across languages"

wav2vec 2.0

wav2vec 2.0

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0

- The first approach to show significant improvements for low-resource ASR.

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0

- The first approach to show significant improvements for low-resource ASR.
- Impressive results on multilingual representations.

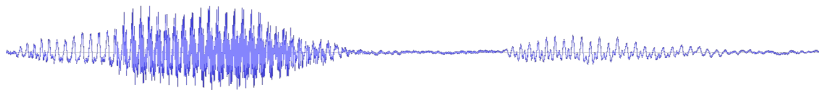
Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0

- The first approach to show significant improvements for low-resource ASR.
- Impressive results on multilingual representations.
- Strong performance on a wide range of downstream speech tasks.

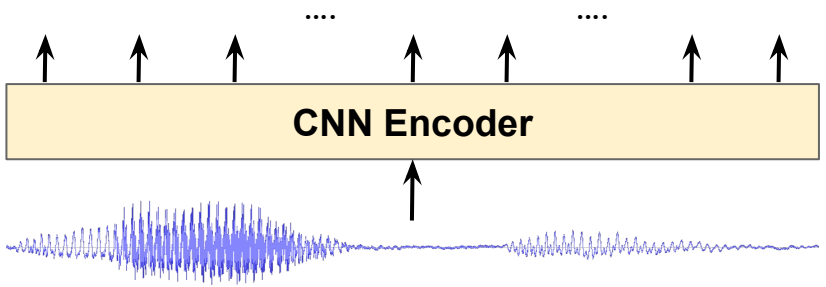
Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: The pretext task



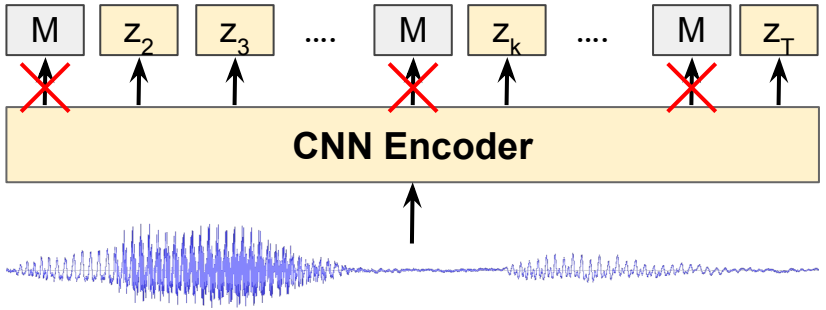
Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: The pretext task



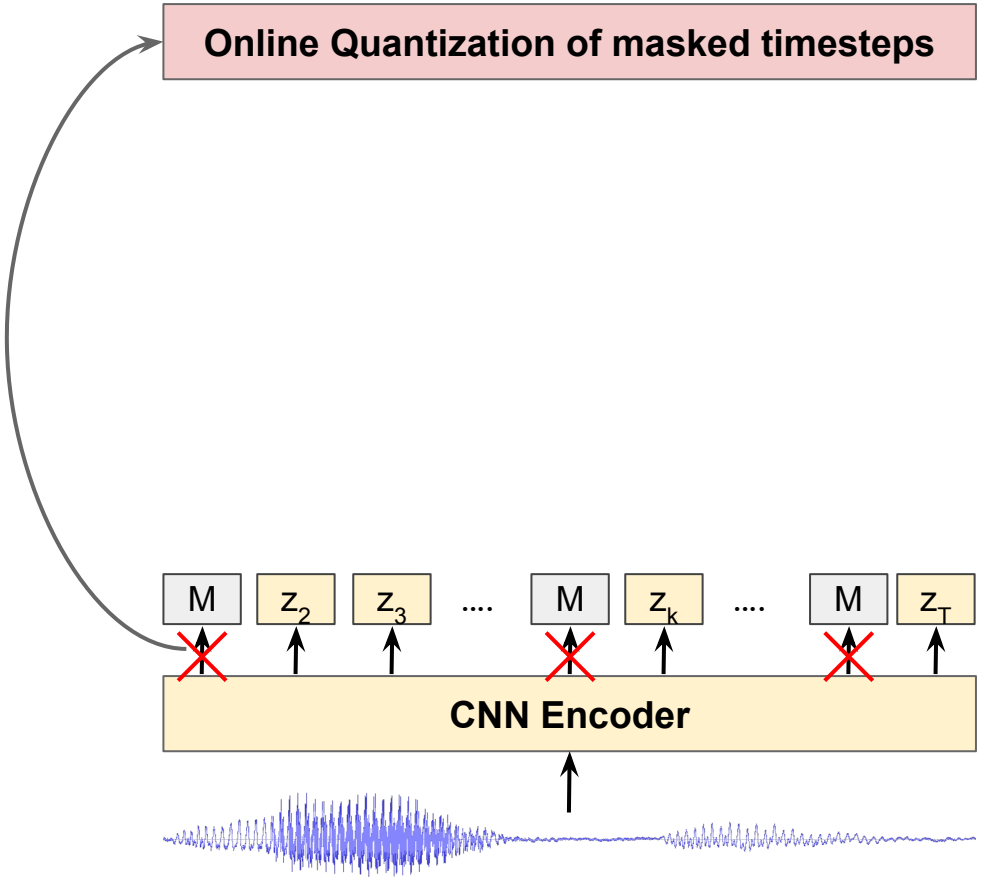
Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: The pretext task



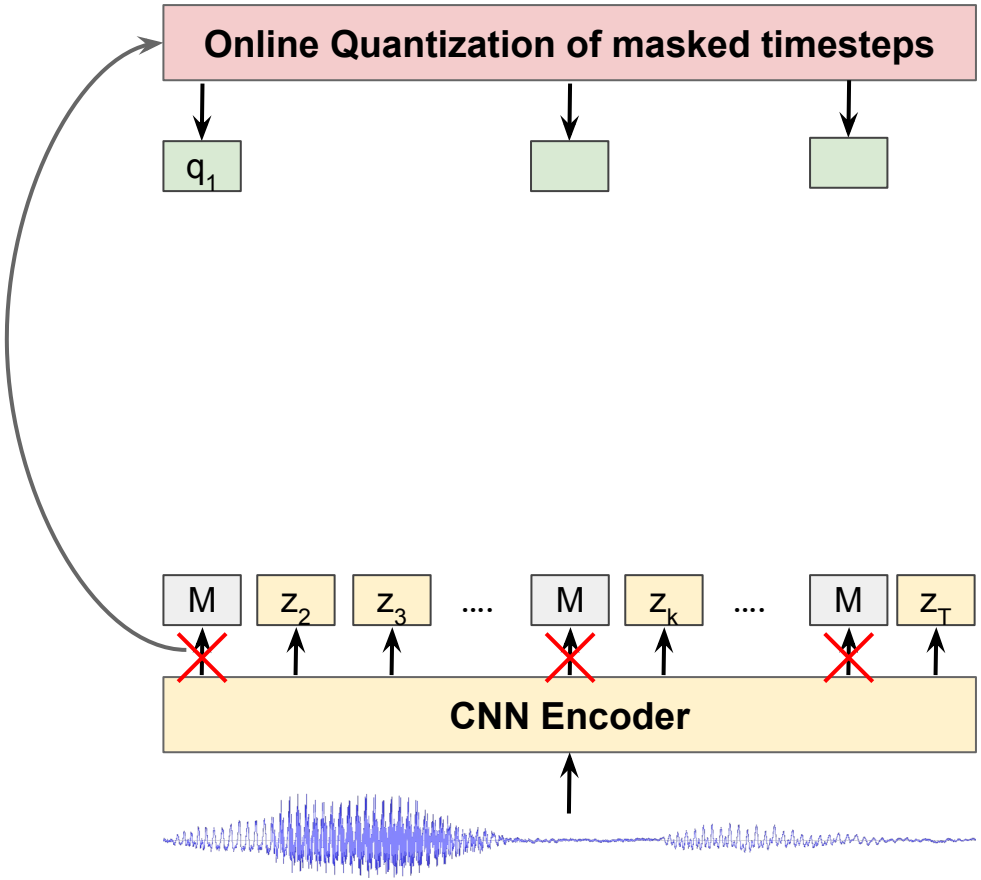
Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: The pretext task



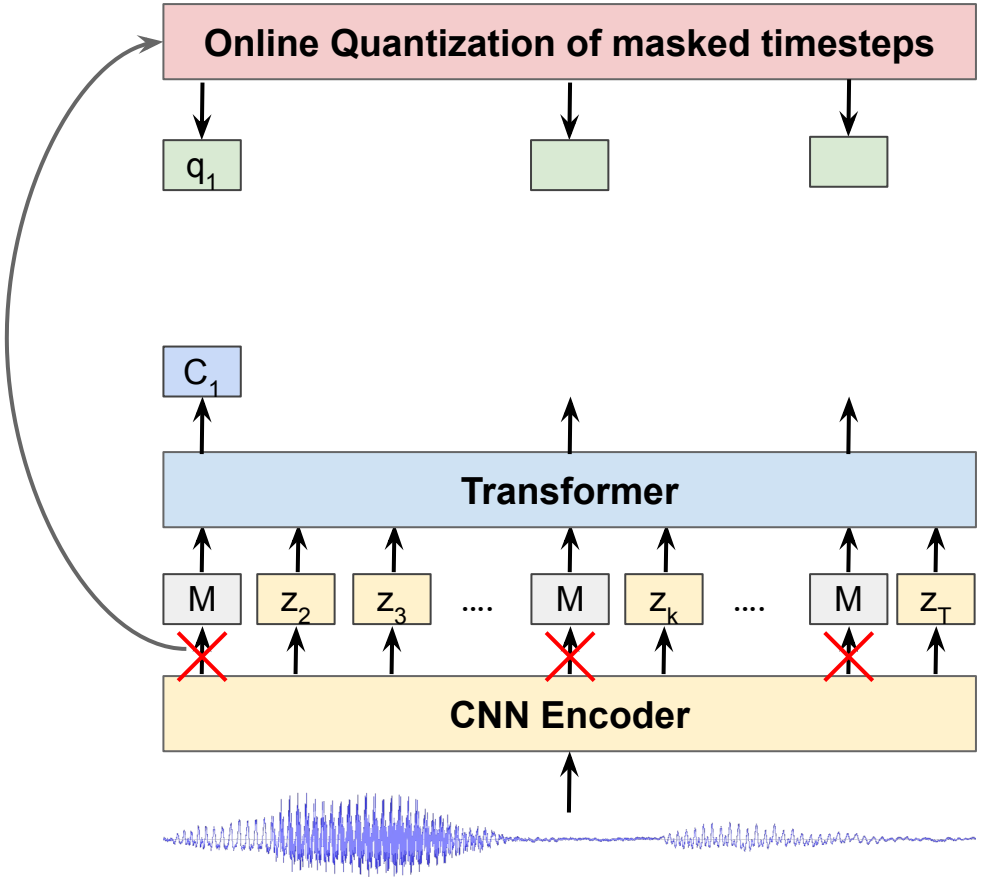
Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: The pretext task



Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

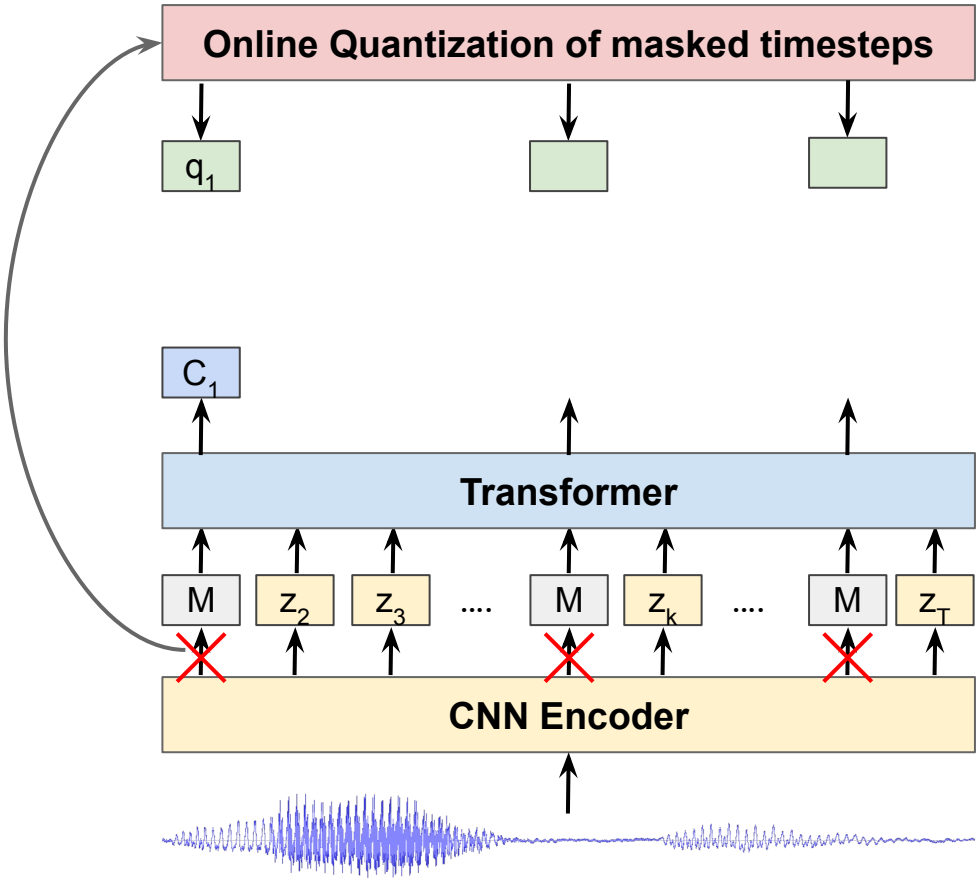
wav2vec 2.0: The pretext task



Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: The pretext task

- The goal is to maximize the similarity between the learned contextual representation and the quantized input features at the same position.

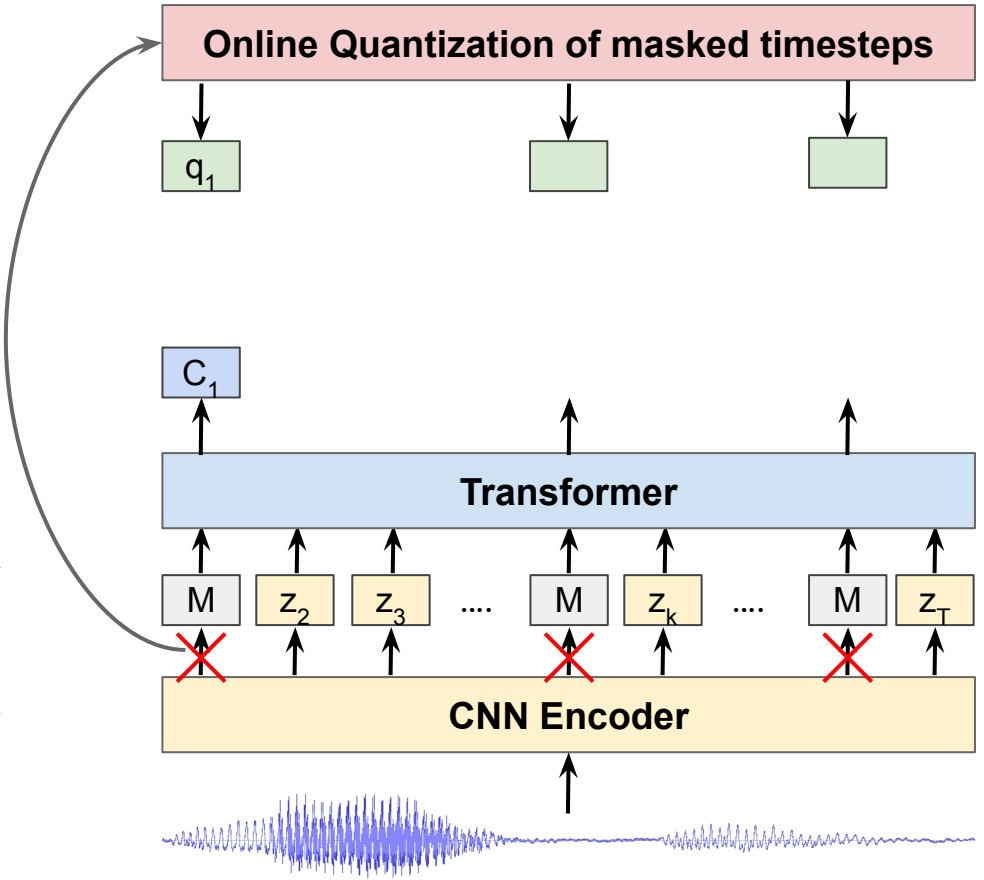


Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: The pretext task

- The goal is to maximize the similarity between the learned contextual representation and the quantized input features at the same position.

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$



Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Gumbel softmax

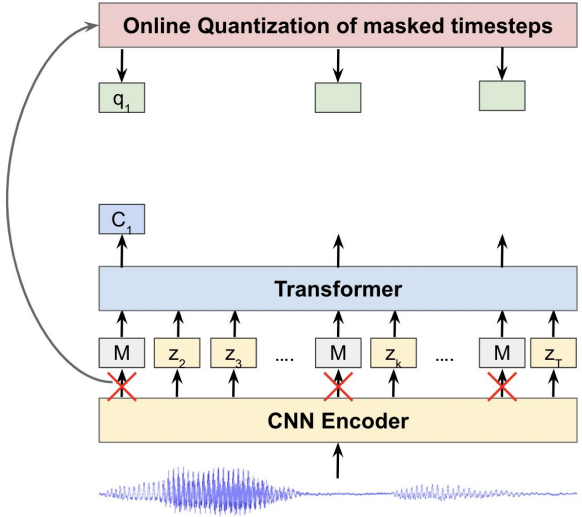
- The model learns an online quantization of audio representations using a Gumbel softmax.

```
gumbels = (logits + gumbels) / tau  # ~Gumbel(logits,tau)
y_soft = gumbels.softmax(dim)

if hard:
    # Straight through.
    index = y_soft.max(dim, keepdim=True)[1]
    y_hard = torch.zeros_like(logits, memory_format=torch.legacy_contiguous_format).scatter_(dim, index, 1.0)
    ret = y_hard - y_soft.detach() + y_soft
else:
    # Reparametrization trick.
    ret = y_soft
return ret
```

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Implementation details

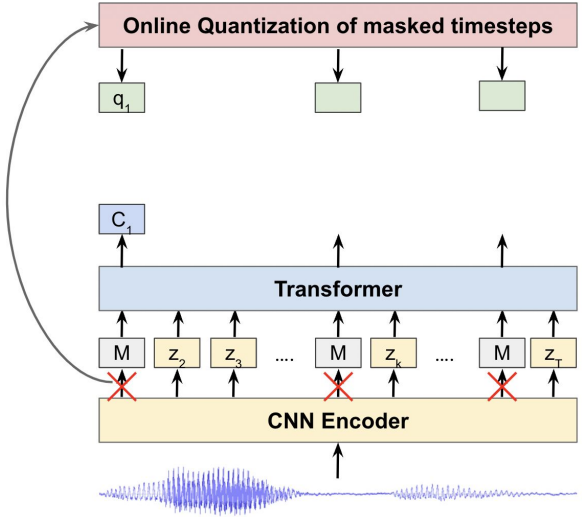
- Product quantization with more than one codebook yields better results.



Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Implementation details

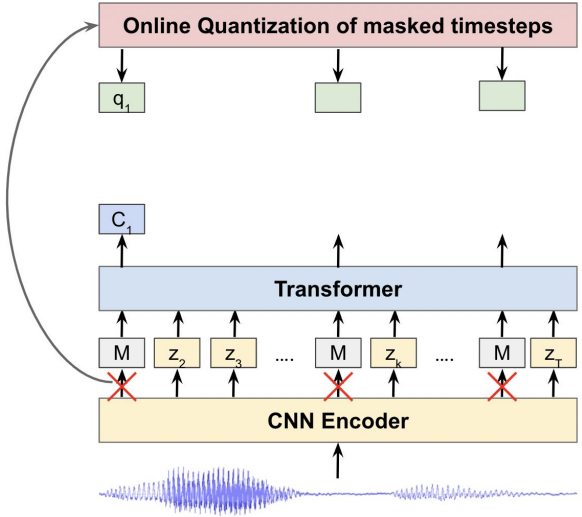
- Product quantization with more than one codebook yields better results.
- An entropy loss is added to the Gumbel softmax distribution to maximize codebook diversity.



Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Implementation details

- Product quantization with more than one codebook yields better results.
- An entropy loss is added to the Gumbel softmax distribution to maximize codebook diversity.
- Negative examples are chosen from masked segments in the same utterance that don't belong to the same codeword.



Baevski et al, 2020 "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations"

wav2vec 2.0: Results

- The first approach to get into single-digit WER on Librispeech test-other using only 10 mins of labels.

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Results

- The first approach to get into single-digit WER on Librispeech test-other using only 10 mins of labels.

Model	Unlabeled data	LM	dev		test	
			clean	other	clean	other
10 min labeled						
Discrete BERT [4]	LS-960	4-gram	15.7	24.1	16.3	25.2
BASE	LS-960	4-gram	8.9	15.7	9.1	15.6
		Transf.	6.6	13.2	6.9	12.9
LARGE	LS-960	Transf.	6.6	10.6	6.8	10.8
	LV-60k	Transf.	4.6	7.9	4.8	8.2

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Results

- It is the first self-supervised approach to produce competitive results compared to semi-supervised learning approaches.

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Results

- It is th
  comp

Model	Unlabeled data	LM	dev		test	
			clean	other	clean	other
Supervised						
CTC Transf [51]	-	CLM+Transf.	2.20	4.94	2.47	5.45
S2S Transf. [51]	-	CLM+Transf.	2.10	4.79	2.33	5.17
Transf. Transducer [60]	-	Transf.	-	-	2.0	4.6
ContextNet [17]	-	LSTM	1.9	3.9	1.9	4.1
Conformer [15]	-	LSTM	2.1	4.3	1.9	3.9
Semi-supervised						
CTC Transf. + PL [51]	LV-60k	CLM+Transf.	2.10	4.79	2.33	4.54
S2S Transf. + PL [51]	LV-60k	CLM+Transf.	2.00	3.65	2.09	4.11
Iter. pseudo-labeling [58]	LV-60k	4-gram+Transf.	1.85	3.26	2.10	4.01
Noisy student [42]	LV-60k	LSTM	1.6	3.4	1.7	3.4
This work						
LARGE - from scratch	-	Transf.	1.7	4.3	2.1	4.6
BASE	LS-960	Transf.	1.8	4.7	2.1	4.8
LARGE	LS-960	Transf.	1.7	3.9	2.0	4.1
	LV-60k	Transf.	1.6	3.0	1.8	3.3

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Results

- It is th
  comp

Model	Unlabeled data	LM	dev		test	
			clean	other	clean	other
Supervised						
CTC Transf [51]	-	CLM+Transf.	2.20	4.94	2.47	5.45
S2S Transf. [51]	-	CLM+Transf.	2.10	4.79	2.33	5.17
Transf. Transducer [60]	-	Transf.	-	-	2.0	4.6
ContextNet [17]	-	LSTM	1.9	3.9	1.9	4.1
Conformer [15]	-	LSTM	2.1	4.3	1.9	3.9
Semi-supervised						
CTC Transf. + PL [51]	LV-60k	CLM+Transf.	2.10	4.79	2.33	4.54
S2S Transf. + PL [51]	LV-60k	CLM+Transf.	2.00	3.65	2.09	4.11
Iter. pseudo-labeling [58]	LV-60k	4-gram+Transf.	1.85	3.26	2.10	4.01
Noisy student [42]	LV-60k	LSTM	1.6	3.4	1.7	3.4
This work						
LARGE - from scratch	-	Transf.	1.7	4.3	2.1	4.6
BASE	LS-960	Transf.	1.8	4.7	2.1	4.8
LARGE	LS-960	Transf.	1.7	3.9	2.0	4.1
	LV-60k	Transf.	1.6	3.0	1.8	3.3

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Extensions

- wav2vec 2.0 inspired many follow up work:

Conneau et. al., 2020 "Unsupervised Cross-lingual Representation Learning for Speech Recognition"
Sadhu et. al., 2021 "Wav2vec-C: A Self-supervised Model for Speech Representation Learning"
Chung et. al., 2021 "W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training"
Babu et. al., 2021 "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale"

wav2vec 2.0: Extensions

- wav2vec 2.0 inspired many follow up work:
  - Multilingual pretraining.

Conneau et. al., 2020 "Unsupervised Cross-lingual Representation Learning for Speech Recognition"
Sadhu et. al., 2021 "Wav2vec-C: A Self-supervised Model for Speech Representation Learning"
Chung et. al., 2021 "W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training"
Babu et. al., 2021 "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale"

wav2vec 2.0: Extensions

- wav2vec 2.0 inspired many follow up work:
  - Multilingual pretraining.
  - With more effective quantization.

Conneau et. al., 2020 "Unsupervised Cross-lingual Representation Learning for Speech Recognition"
Sadhu et. al., 2021 "Wav2vec-C: A Self-supervised Model for Speech Representation Learning"
Chung et. al., 2021 "W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training"
Babu et. al., 2021 "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale"

wav2vec 2.0: Extensions

- wav2vec 2.0 inspired many follow up work:
  - Multilingual pretraining.
  - With more effective quantization.
  - Combining contrastive and predictive losses.
  - …

Conneau et. al., 2020 "Unsupervised Cross-lingual Representation Learning for Speech Recognition"
Sadhu et. al., 2021 "Wav2vec-C: A Self-supervised Model for Speech Representation Learning"
Chung et. al., 2021 "W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training"
Babu et. al., 2021 "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale"

Speech representation learning methods

Contrastive approaches

Predictive approaches

Generative approaches

Hidden Unit BERT (HuBERT)

HuBERT

Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT

- A simple method to apply BERT style representation learning for speech.

Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT

- A simple method to apply BERT style representation learning for speech.
- Matched or beat the SOTA on ASR while being the best for many speech tasks.

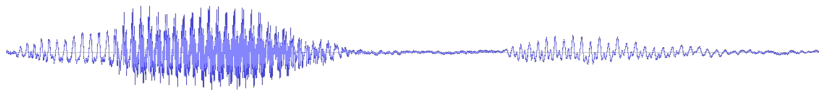
Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT

- A simple method to apply BERT style representation learning for speech.
- Matched or beat the SOTA on ASR while being the best for many speech tasks.
- With its high-quality discrete units, HuBERT facilitated Textless NLP research.

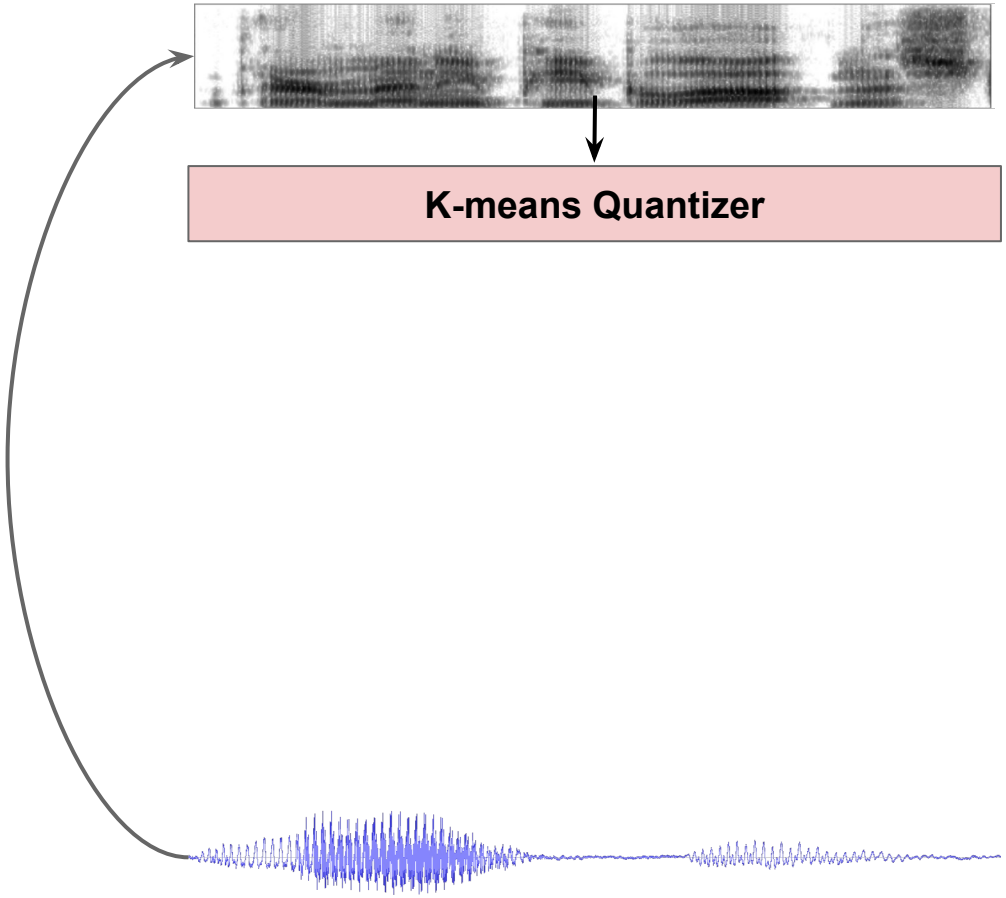
Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

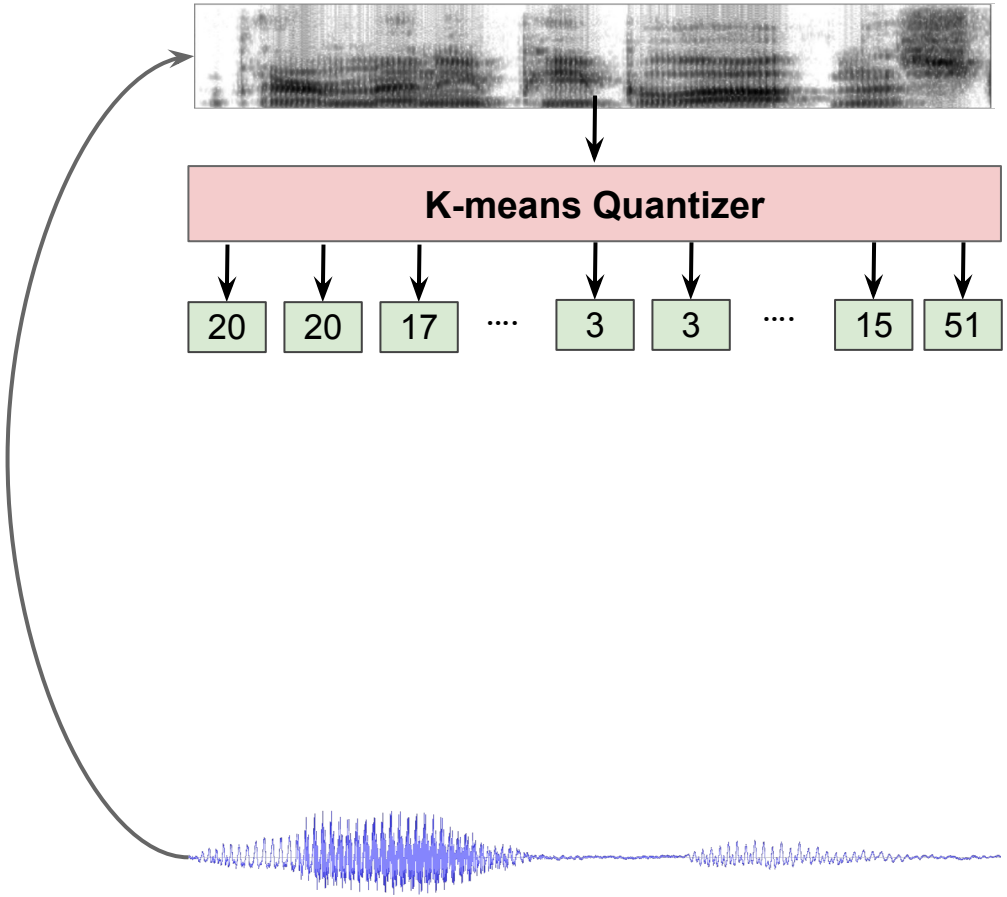
HuBERT: The pretext task



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

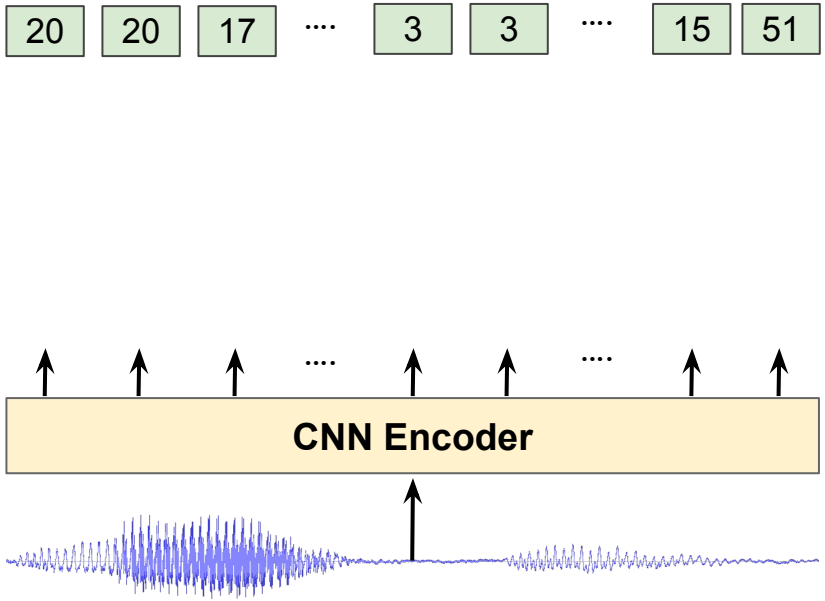
HuBERT: The pretext task

- The K-means quantizer produces frame-level labels.



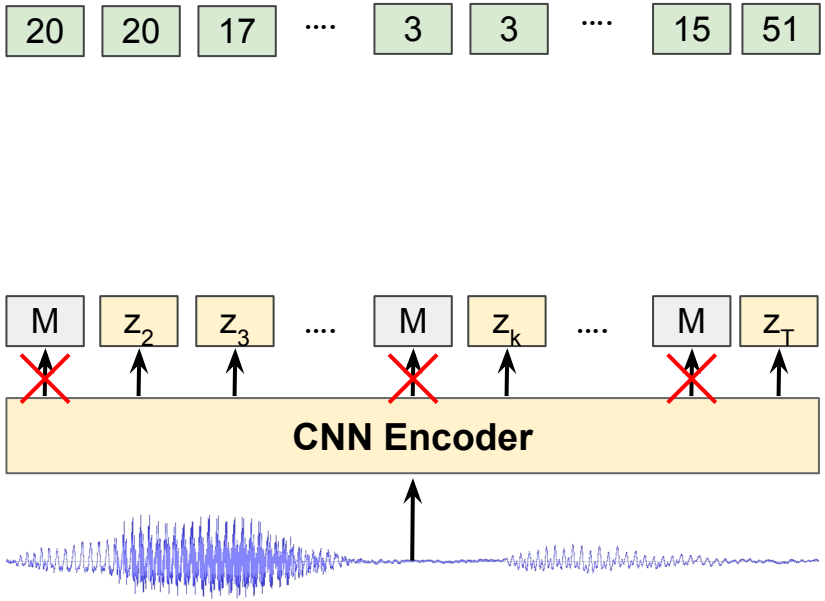
Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task



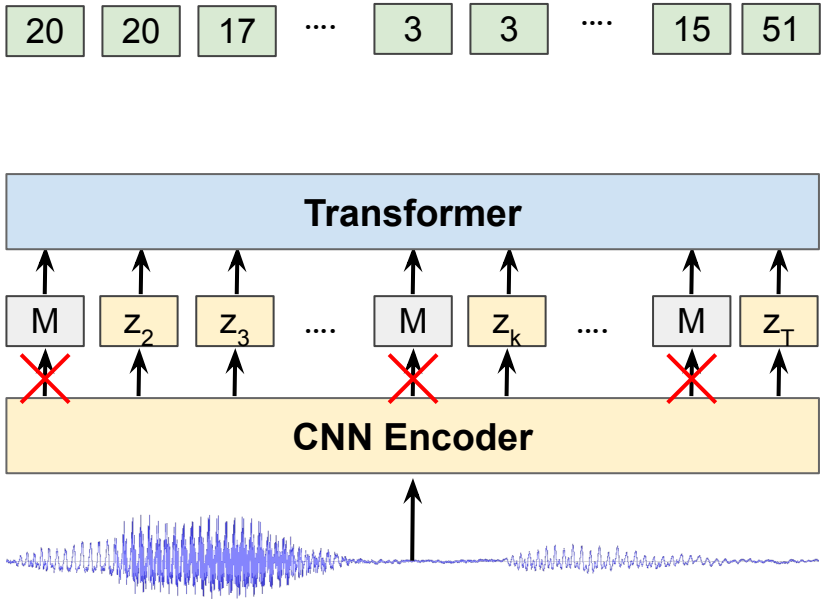
Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

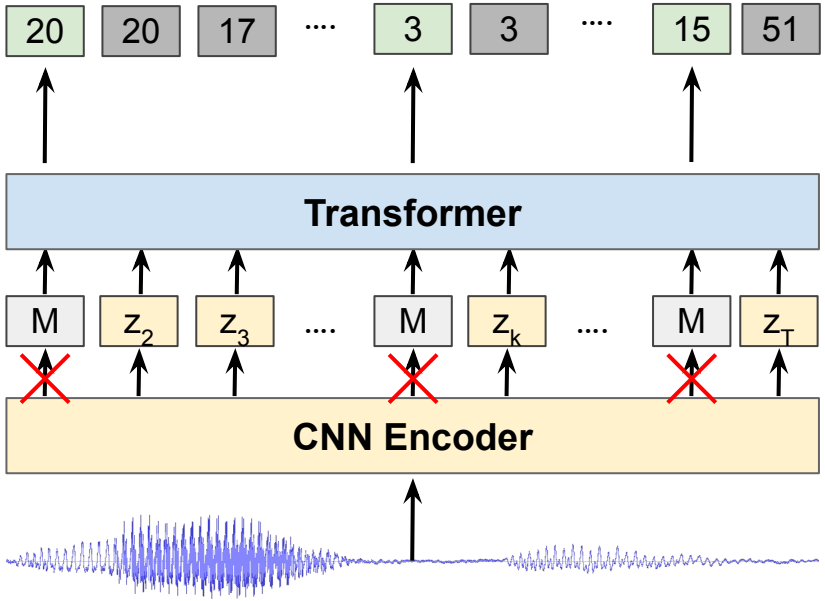
HuBERT: The pretext task



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task

- Although the frame labels are imperfect, their consistency is more important!

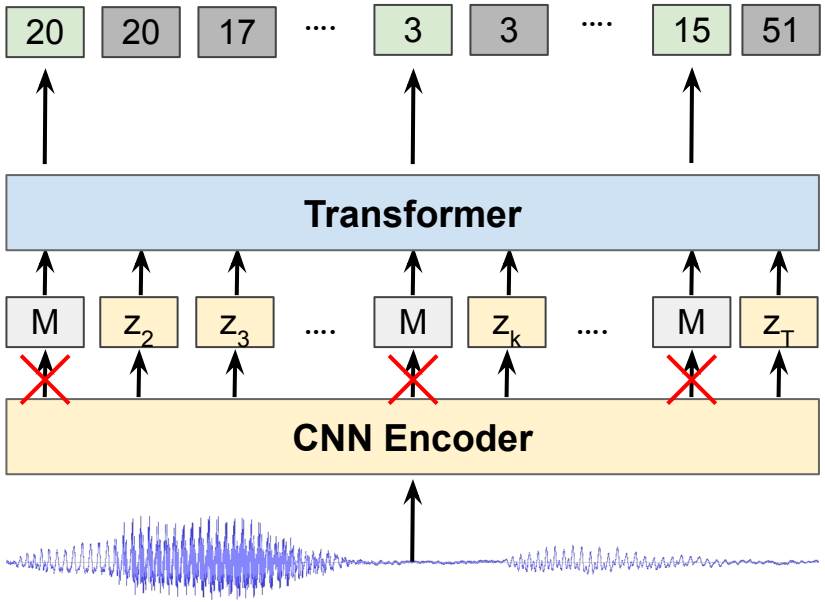


Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task

- Although the frame labels are imperfect, their consistency is more important!
- The model is trained using masked prediction:

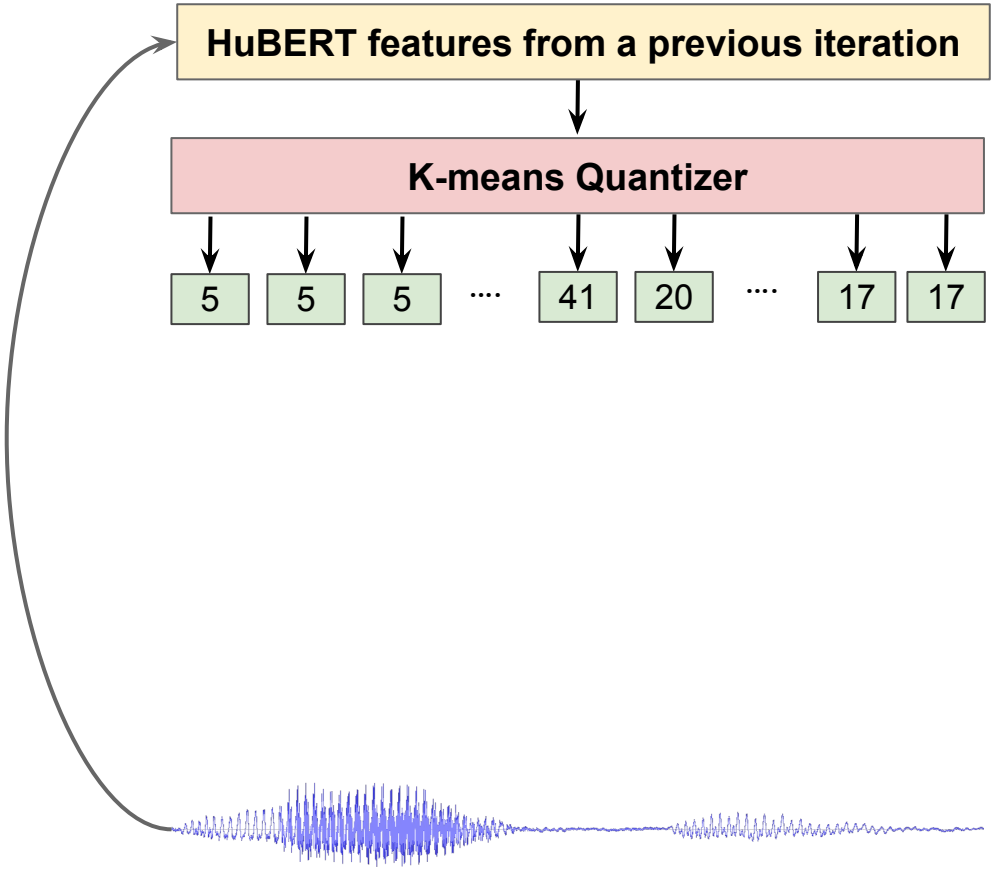
$$\mathcal{L}_m = \sum_{t \in M} -\log p(y_t \mid X)$$



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task

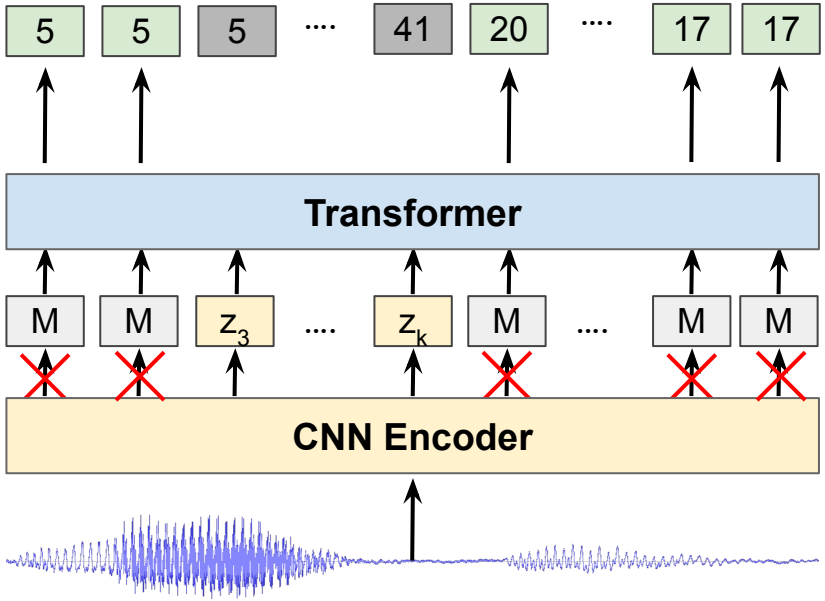
- Then the process can be repeated using learned HuBERT features from a previous iteration.



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task

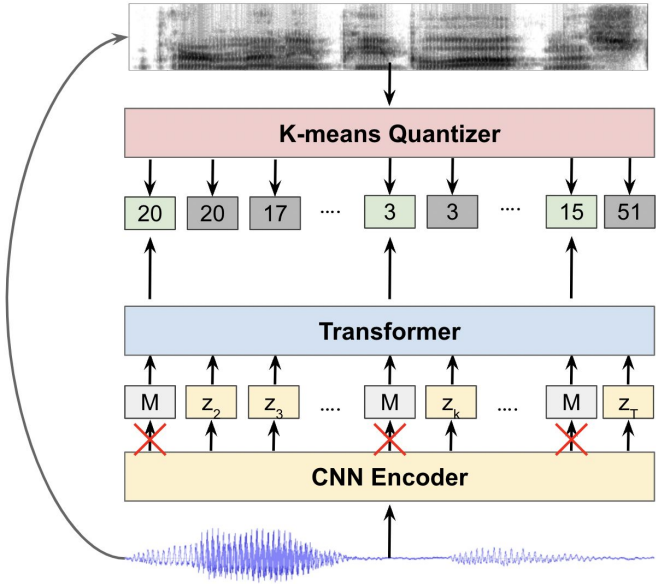
- Then the process can be repeated using learned HuBERT features from a previous iteration.



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: Implementation details

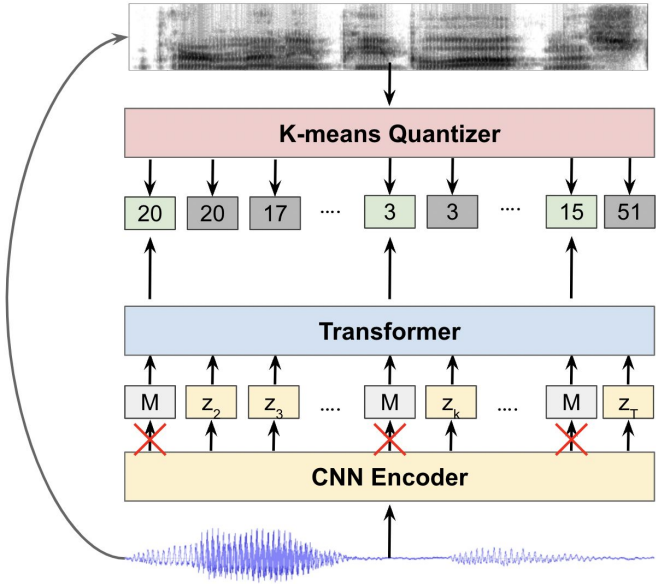
- A small codebook size, e.g., 50, 100, is used for the initial training iteration to focus on phonetic differences rather than speaker and style.



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: Implementation details

- A small codebook size, e.g., 50, 100, is used for the initial training iteration to focus on phonetic differences rather than speaker and style.
- For the subsequent two iterations, layers 6 and 9 of the base architecture (12 layers) are used for the clustering steps. They found empirically to contain higher quality features over many speech tasks.



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

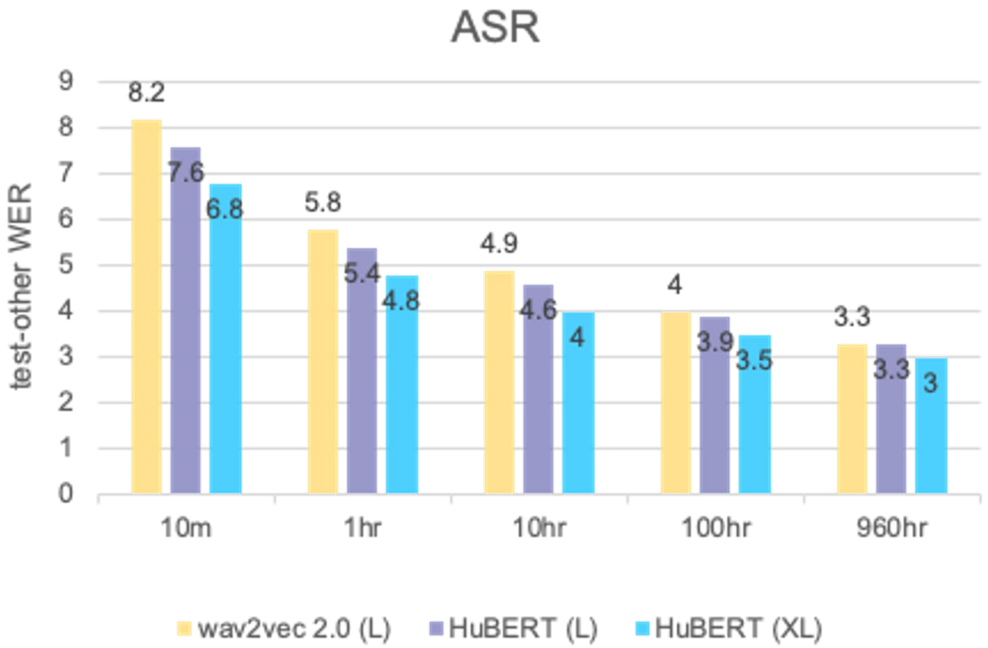
HuBERT: Results

- Matched or beat the SOTA on ASR.

Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: Results

- Matched or beat the SOTA on ASR.



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: Results

- Matched or beat the SOTA on ASR.
- The best representations for multiple downstream tasks.

	PR	KS	IC	SID	ER	ASR (WER)		QbE	SF		ASV	SD
	PER ↓	Acc ↑	Acc ↑	Acc ↑	Acc ↑	w/o ↓	w/ LM ↓	MTWV ↑	F1 ↑	CER ↓	EER ↓	DER ↓
FBANK	82.01	8.63	9.10	8.5E-4	35.39	23.18	15.21	0.0058	69.64	52.94	9.56	10.05
PASE+ [16]	58.95	82.54	29.82	37.99	57.86	24.92	16.61	0.0072	62.14	60.17	11.61	8.68
APC [7]	42.21	91.01	74.69	60.42	59.33	21.61	15.09	0.0310	70.46	50.89	8.56	10.53
VQ-APC [32]	41.49	91.11	74.48	60.15	59.66	21.72	15.37	0.0251	68.53	52.91	8.72	10.45
NPC [33]	43.69	88.96	69.44	55.92	59.08	20.94	14.69	0.0246	72.79	48.44	9.4	9.34
Mockingjay [8]	70.84	83.67	34.33	32.29	50.28	23.72	15.94	6.6E-04	61.59	58.89	11.66	10.54
TERA [9]	49.17	89.48	57.90	57.57	56.27	18.45	12.44	0.0013	67.50	54.17	15.89	9.96
modified CPC [34]	42.54	91.88	64.09	39.63	60.96	20.02	13.57	0.0326	71.19	49.91	12.86	10.38
wav2vec [12]	32.24	95.59	84.92	56.56	59.79	16.40	11.30	0.0485	76.37	43.71	7.99	9.9
vq-wav2vec [13]	34.24	93.38	85.68	38.80	58.24	18.70	12.69	0.0410	77.68	41.54	10.38	9.93
wav2vec 2.0 Base [14]	5.56	96.23	92.35	75.18	63.43	9.57	6.32	0.0233	88.30	24.77	6.02	6.08
wav2vec 2.0 Large [14]	4.75	**96.66**	95.28	86.14	65.64	3.75	3.10	0.0489	86.94	27.80	5.65	**5.62**
HuBERT Base [35]	5.05	96.30	98.34	81.42	64.92	6.74	4.93	**0.0736**	88.53	25.20	**5.11**	5.88
HuBERT Large [35]	**3.28**	95.29	**98.76**	**90.33**	**67.62**	**3.67**	**2.91**	0.0353	**89.81**	**21.76**	5.98	5.75

Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: Extensions

- HuBERT inspired many follow up work:

Chen et. al., 2021 "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing"
Chiu et. al. 2022 "Self-supervised Learning with Random-projection Quantizer for Speech Recognition"
Shi et. al. 2022 "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction"
Lakhotia et. al., 2021 "Generative Spoken Language Modeling from Raw Audio"

HuBERT: Extensions

- HuBERT inspired many follow up work:
  - Combined masked prediction and denoising pre-training.

Chen et. al., 2021 "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing"
Chiu et. al. 2022 "Self-supervised Learning with Random-projection Quantizer for Speech Recognition"
Shi et. al. 2022 "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction"
Lakhotia et. al., 2021 "Generative Spoken Language Modeling from Raw Audio"

HuBERT: Extensions

- HuBERT inspired many follow up work:
  - Combined masked prediction and denoising pre-training.
  - Random clustering is as effective in learning representations as k-means.

Chen et. al., 2021 "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing"
Chiu et. al. 2022 "Self-supervised Learning with Random-projection Quantizer for Speech Recognition"
Shi et. al. 2022 "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction"
Lakhotia et. al., 2021 "Generative Spoken Language Modeling from Raw Audio"

HuBERT: Extensions

- HuBERT inspired many follow up work:
  - Combined masked prediction and denoising pre-training.
  - Random clustering is as effective in learning representations as k-means.
  - Multimodal clustering for audio-visual ASR.

Chen et. al., 2021 "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing"
Chiu et. al. 2022 "Self-supervised Learning with Random-projection Quantizer for Speech Recognition"
Shi et. al. 2022 "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction"
Lakhotia et. al., 2021 "Generative Spoken Language Modeling from Raw Audio"

HuBERT: Extensions

- HuBERT inspired many follow up work:
  - Combined masked prediction and denoising pre-training.
  - Random clustering is as effective in learning representations as k-means.
  - Multimodal clustering for audio-visual ASR.
  - High-quality discrete units facilitated textless NLP research for speech generation.
  - ….

Chen et. al., 2021 "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing"
Chiu et. al. 2022 "Self-supervised Learning with Random-projection Quantizer for Speech Recognition"
Shi et. al. 2022 "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction"
Lakhotia et. al., 2021 "Generative Spoken Language Modeling from Raw Audio"

data2vec

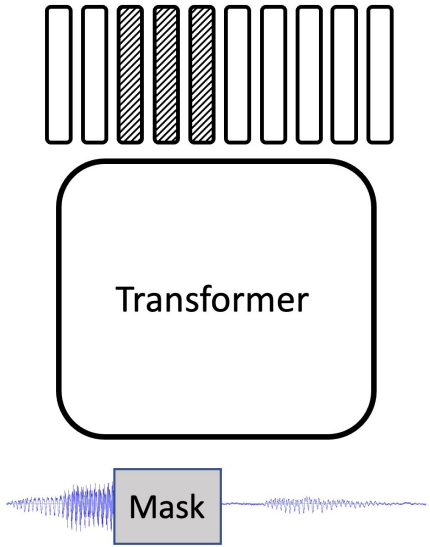
data2vec

- A single algorithm can work with entirely different input types, e.g., Speech, Vision, and Text.

Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: The pretext task

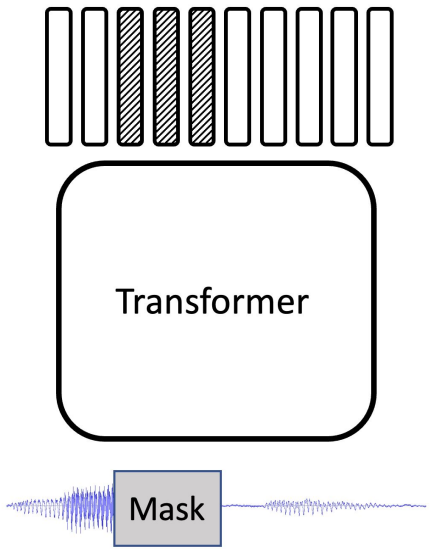
- How can we get high-quality targets for masked positions?



Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: The pretext task

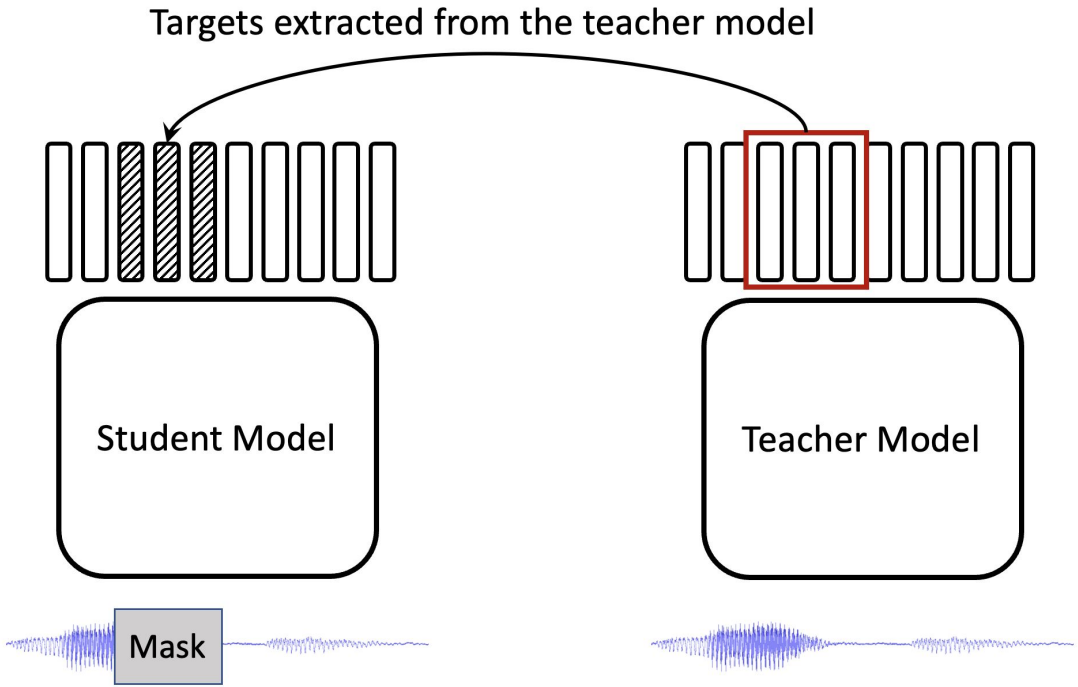
- How can we get high-quality targets for masked positions?
- What about using the same transformer model but with unmasked input?



Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: The pretext task

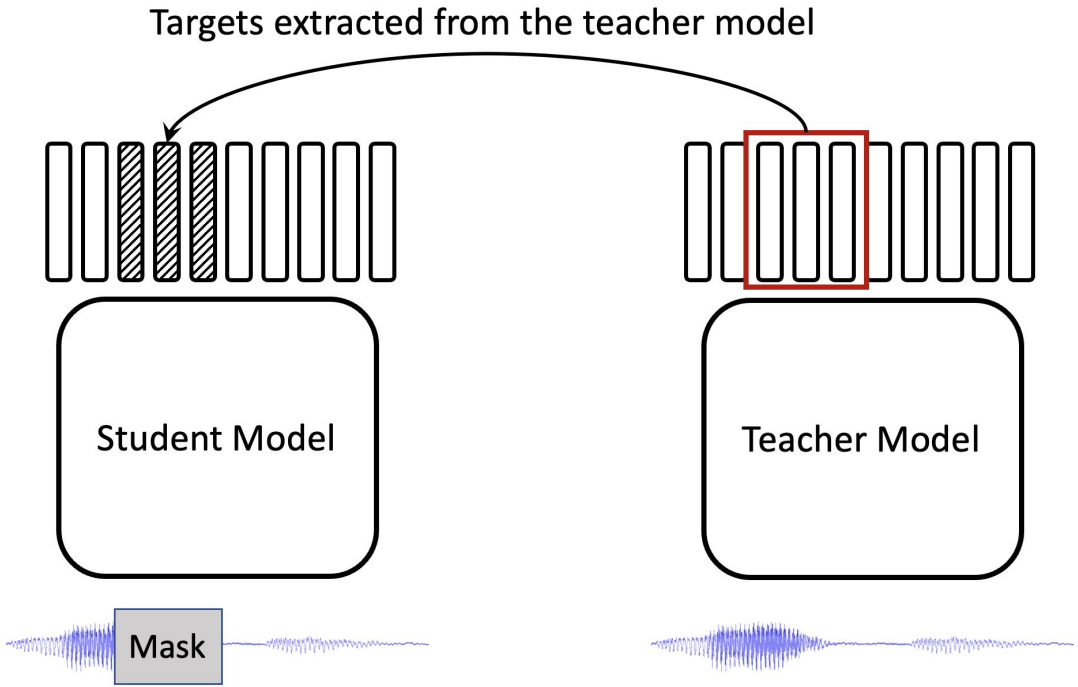
- How can we get high-quality targets for masked positions?
- What about using the same transformer model but with unmasked input?



Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: The pretext task

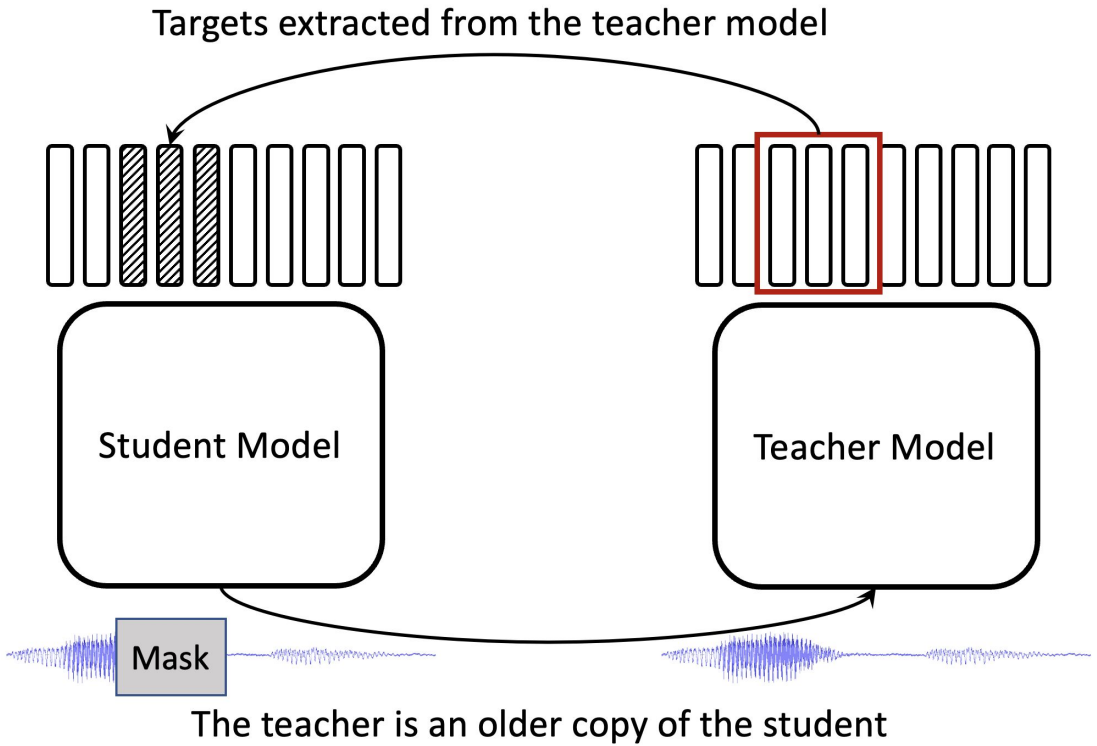
- How do we avoid trivial solutions? i.e., **representation collapse** with constant outputs for all masked data.



Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: The pretext task

- How do we avoid trivial solutions? i.e., **representation collapse** with constant outputs for all masked data.
- **Solution**: the teacher model is updated as an exponential moving average (EMA) of the student model.

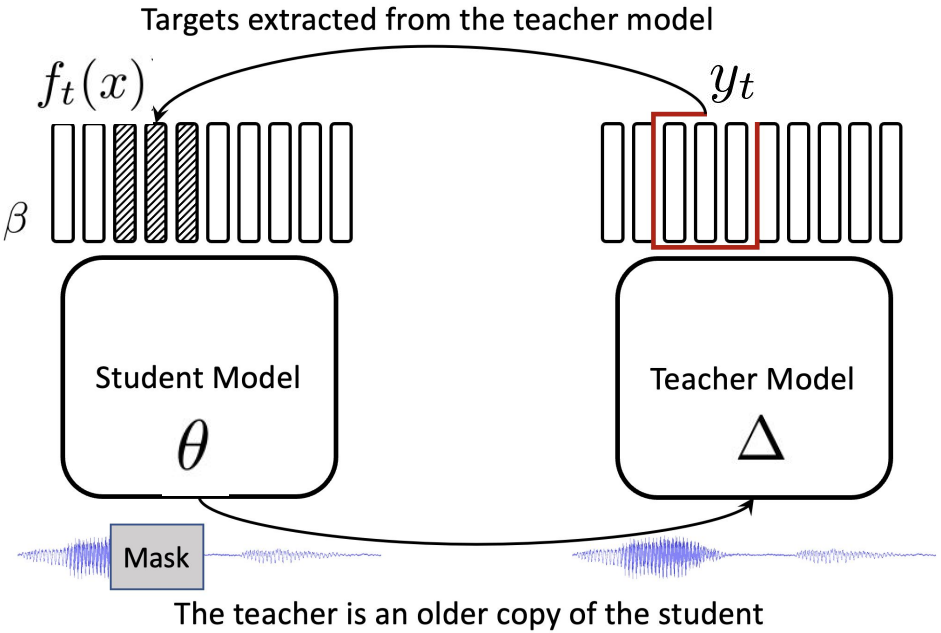


Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: The pretext task

$$\mathcal{L}(y_t, f_t(x)) = \begin{cases} \frac{1}{2}(y_t - f_t(x))^2/\beta & |y_t - f_t(x)| \leq \beta \\ (|y_t - f_t(x)| - \frac{1}{2}\beta) & \text{otherwise} \end{cases}$$

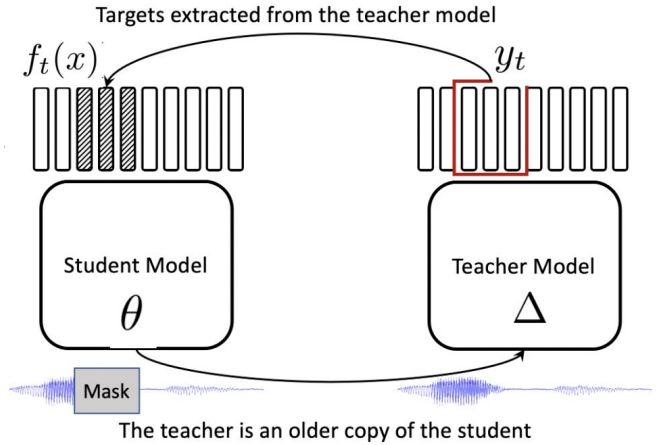
$$\Delta \leftarrow \tau\Delta + (1-\tau)\,\theta$$



Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: Implementation details

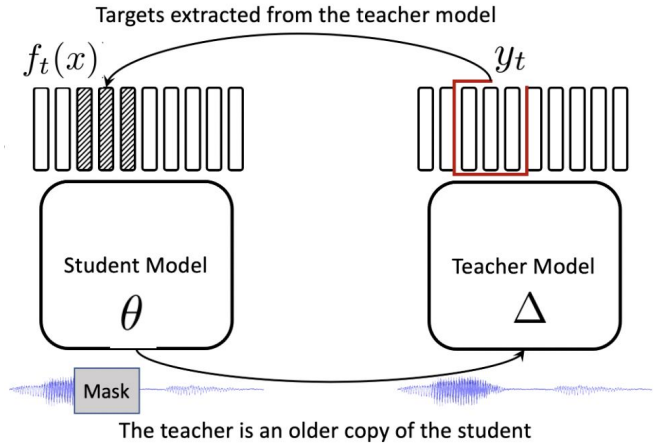
- To avoid collapse: Carefully tune the momentum factor, and don't use high max learning rates or short LR warm-up.



Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: Implementation details

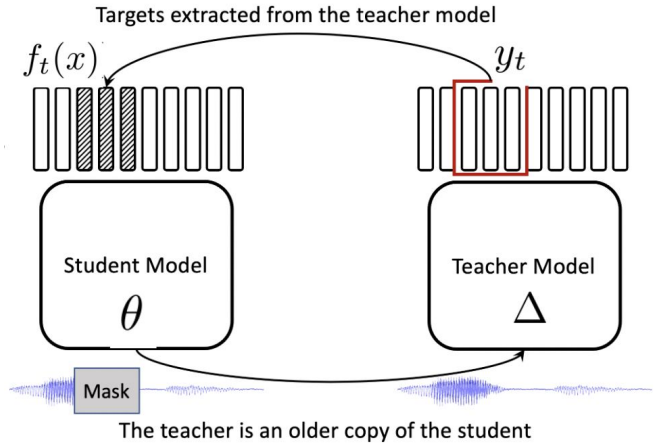
- To avoid collapse: Carefully tune the momentum factor, and don't use high max learning rates or short LR warm-up.
- Targets of the masked timesteps are the average of instance-normalized outputs of the top K blocks of the teacher network.



Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: Results

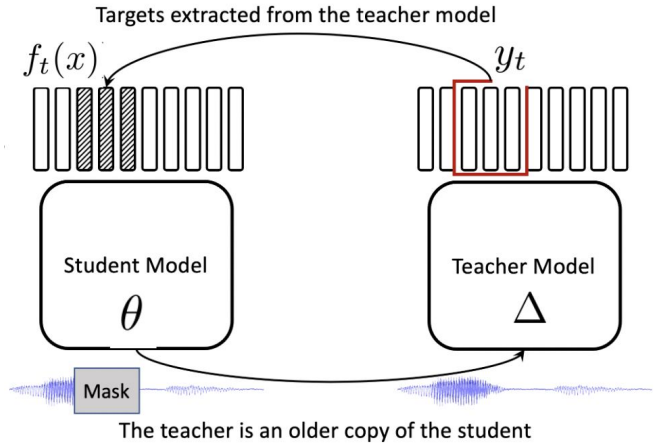
- A single algorithm can work with completely different input types, e.g., Speech, Vision, and Text.



Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: Results

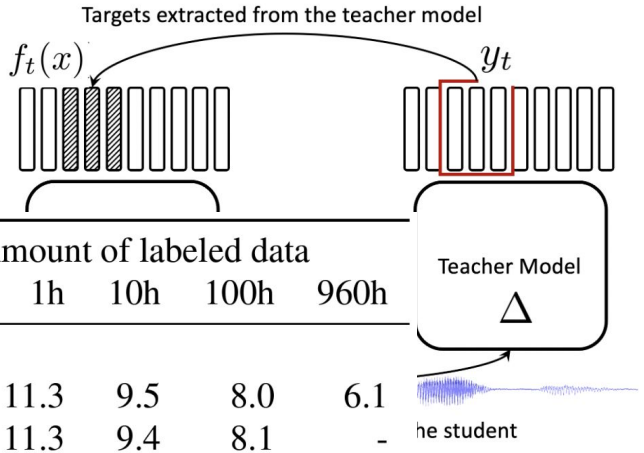
- A single algorithm can work with completely different input types, e.g., Speech, Vision, and Text.
- Strong representations are driven by the model's ability to predict complete contextualized representations from incomplete ones.



Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

data2vec: Results

- A si
diff
- Strc
abil
repr



	Unlabeled data	LM	Amount of labeled data 10m	1h	10h	100h	960h
Base models							
wav2vec 2.0 (Baevski et al., 2020b)	LS-960	4-gram	15.6	11.3	9.5	8.0	6.1
HuBERT (Hsu et al., 2021)	LS-960	4-gram	15.3	11.3	9.4	8.1	-
WavLM (Chen et al., 2021a)	LS-960	4-gram	-	10.8	9.2	7.7	-
data2vec	LS-960	4-gram	12.3	9.1	8.1	6.8	5.5
Large models							
wav2vec 2.0 (Baevski et al., 2020b)	LS-960	4-gram	10.3	7.1	5.8	4.6	3.6
HuBERT (Hsu et al., 2021)	LS-960	4-gram	10.1	6.8	5.5	4.5	3.7
WavLM (Chen et al., 2021a)	LS-960	4-gram	-	6.6	5.5	4.6	-
data2vec	LS-960	4-gram	8.4	6.3	5.3	4.6	3.7

Baevski et. al, 2022 "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language"

Speech representation learning methods

Contrastive approaches

Predictive approaches

Generative approaches

Vector Quantised Variational AutoEncoder (VQ-VAE)

VQ-VAE

- A generative model with discrete latent variables.

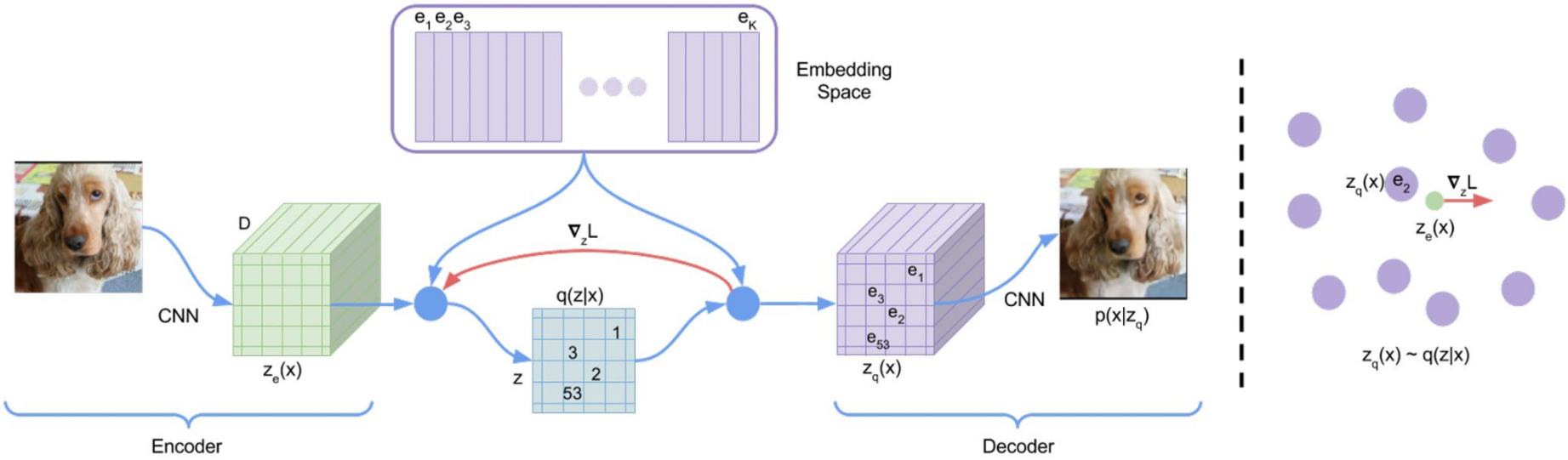
van den Oord et. al., 2017 "Neural Discrete Representation Learning"

VQ-VAE

- A generative model with discrete latent variables.
- Influenced lots of subsequent research work

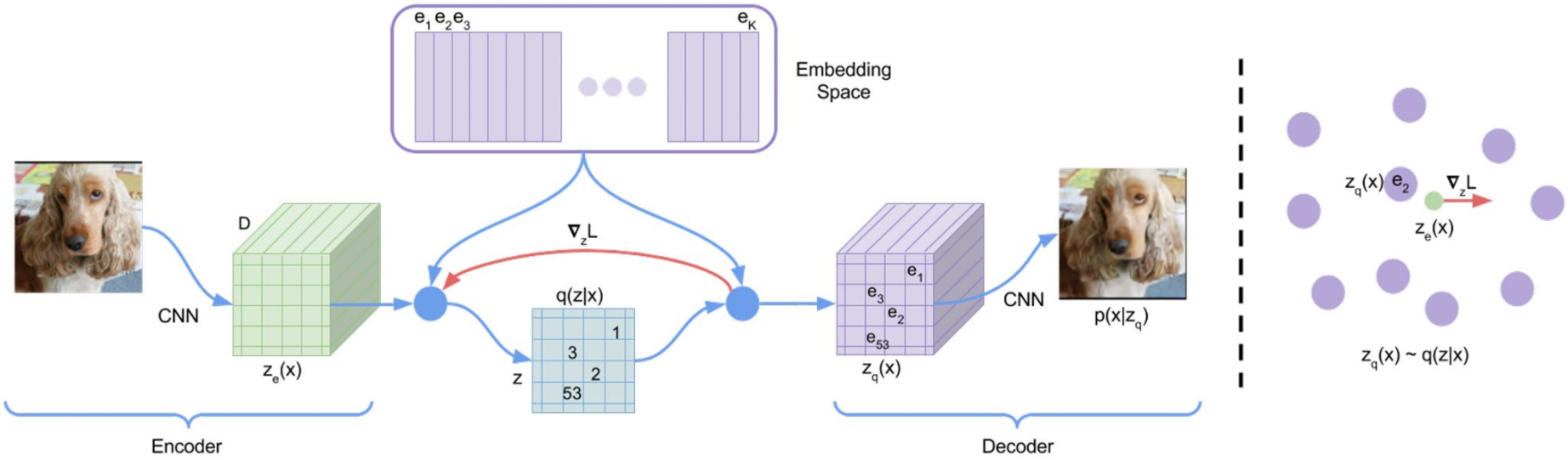
van den Oord et. al., 2017 "Neural Discrete Representation Learning"

VQ-VAE



van den Oord et. al., 2017 "Neural Discrete Representation Learning"

VQ-VAE



$$L = \log p(x|z_q(x)) + \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \text{sg}[e]\|_2^2$$

sg[ ] is the stop gradient operator

van den Oord et. al., 2017 "Neural Discrete Representation Learning"

Autoregressive Predictive Coding (APC)

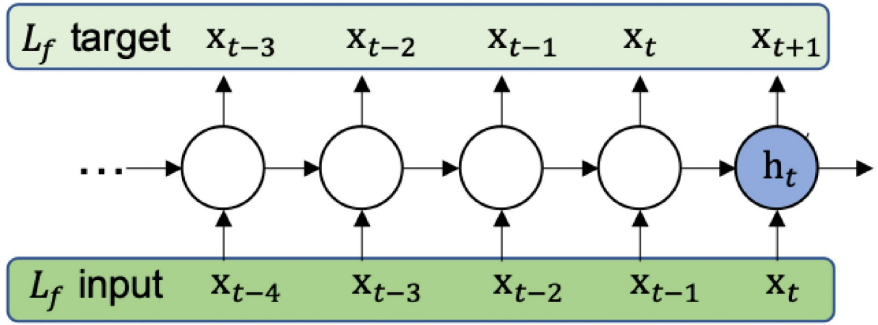
APC

- It is inspired by the next-step prediction loss for language model training.

Chung et. al., 2019 "An Unsupervised Autoregressive Model for Speech Representation Learning"

APC

- It is inspired by the next-step prediction loss for language model training.
- The model is pre-trained to predict a frame **n** steps in the future.



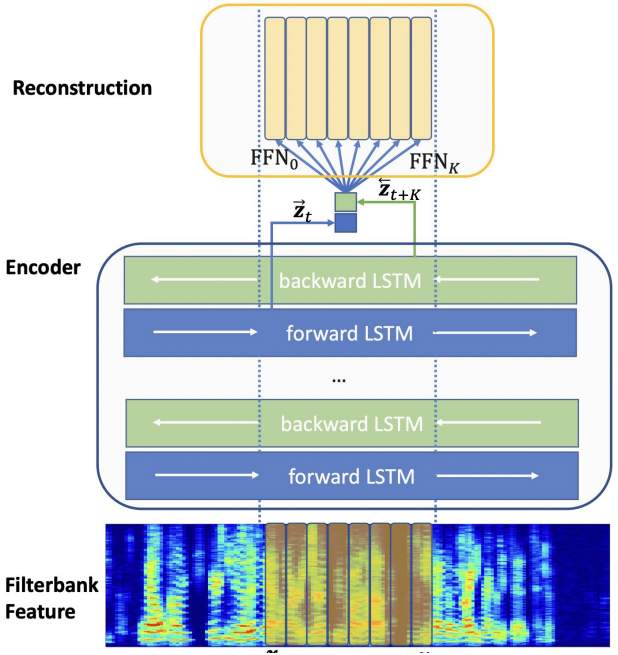
Chung et. al., 2019 "An Unsupervised Autoregressive Model for Speech Representation Learning"

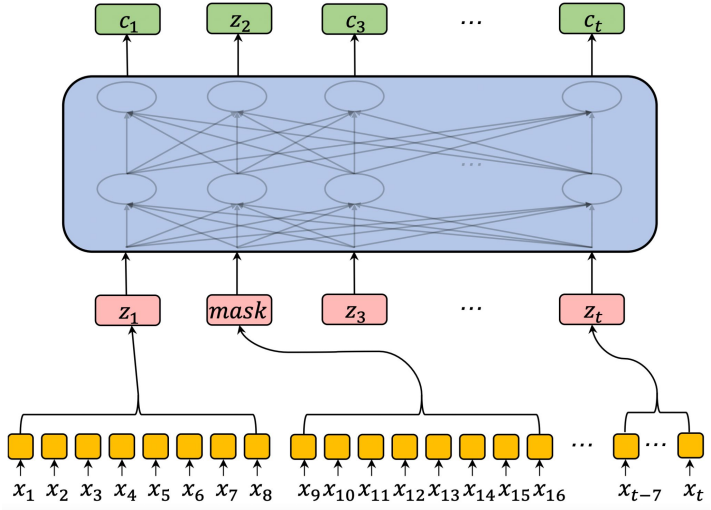
Masked Reconstruction

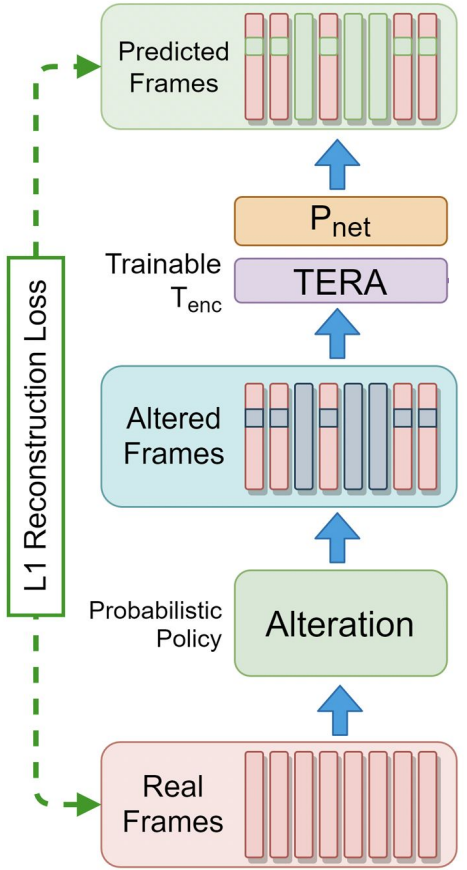
Masked Reconstruction

- Inspired by the success of BERT, many approaches were proposed to reconstruct continuous input from a masked view of the input.

Masked Reconstruction: DeCoAR, MPC, TERA







Ling et. al., 2019 "Deep Contextualized Acoustic Representations For Semi-Supervised Speech Recognition"
Jiang et. al., 2019 "Improving Transformer-based Speech Recognition Using Unsupervised Pre-training"
Liu et. al., 2020 "TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech"

Multi-modal SSL

Karen Livescu

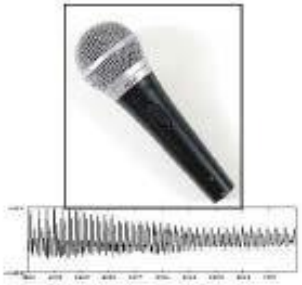
“Human speech is multi-modal”

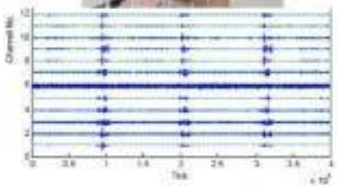
- Seeing a speaker’s face
  - Improves the effective SNR by ~22 dB [Sumby & Pollack 1954]
  - Improves understanding of complex or accented speech [Reisberg et al. 1987]
- McGurk effect: Listeners cannot ignore visual cues even if they try
  - Neither adults… [McGurk & MacDonald 1976]
  - Nor infants [Rosenblum et al. 1997]
- Visual cues facilitate learning phonetic segments in children [Mills 1987, Legerstee 1990]

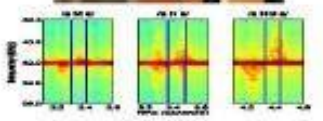
“Human speech is multi-modal”

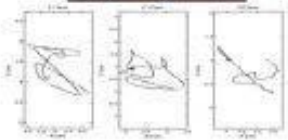
- It’s not just the speaker’s face…
  - The visual scene informs about the content
- It’s not just the visual modality…
  - Physical environment/touch
  - Situational context
  - Listener’s internal state

Types of multi-modal speech data







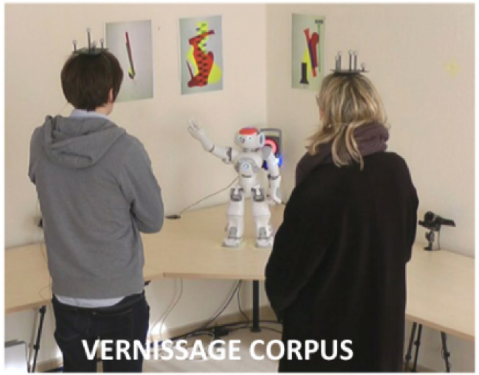


Type 1: “Intrinsic”

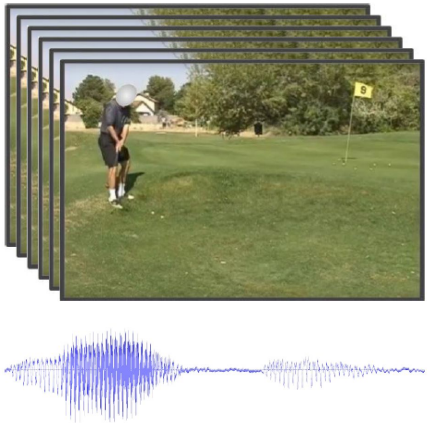
- Multiple modalities associated with the speech itself
- Typically used to improve robustness to noise of a representation or recognition model

Figure credits:
[Schultz & Wand *Sp. Comm.* 2009, Zhu et al. Interspeech 2007, Lingala et al. *Mag. Res. Med.* 2016, Saenko et al. *PAMI* 2009, Paula West]

Types of multi-modal speech data



http://vernissage.humavips.eu/



https://srvk.github.io/how2-dataset/

https://groups.csail.mit.edu/sls/downloads/placesaudio/

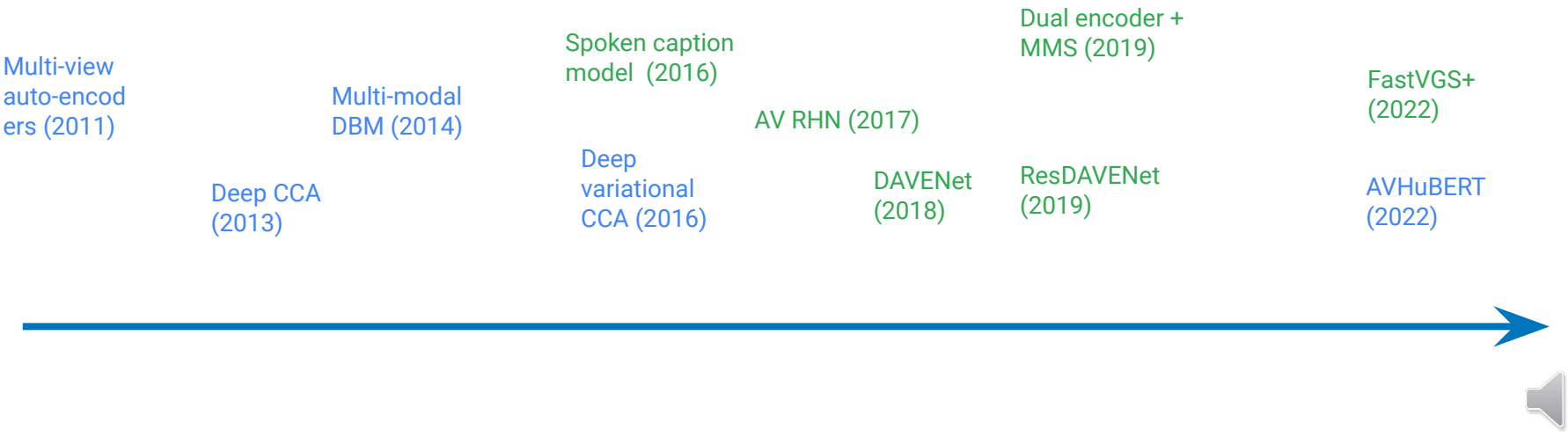
Type 2:

- "Contextual": Additional modalities provide context beyond the speech signal
- Typically used to learn a more semantic model, or as complementary input

Multi-modal representation learning timeline

Green = Learning with contextual modalities

Blue = Learning with intrinsic modalities



Learning with intrinsic modalities

Learning representations from multiple intrinsic modalities

Often referred to as multi-view/multi-modal representation learning

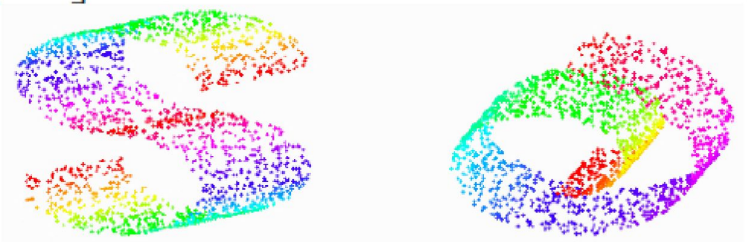
Training data: samples of a *d*-dimensional random vector that has a natural split into two sub-vectors (e.g. x = speech audio, y = corresponding image)

Goal: Find a representation of each view that is predictive of the other, or a joint representation of both

Intuition: If the noise in the views is independent, then the shared information must be signal

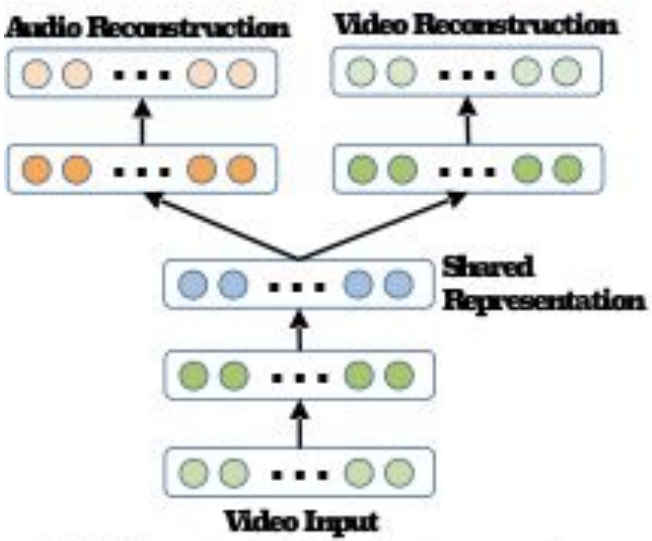
At test time, each view may be available alone, or they may be available together

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \ \mathbf{x} \in \mathbb{R}^{d_x}, \ \mathbf{y} \in \mathbb{R}^{d_y}, \ d_x + d_y = d$$

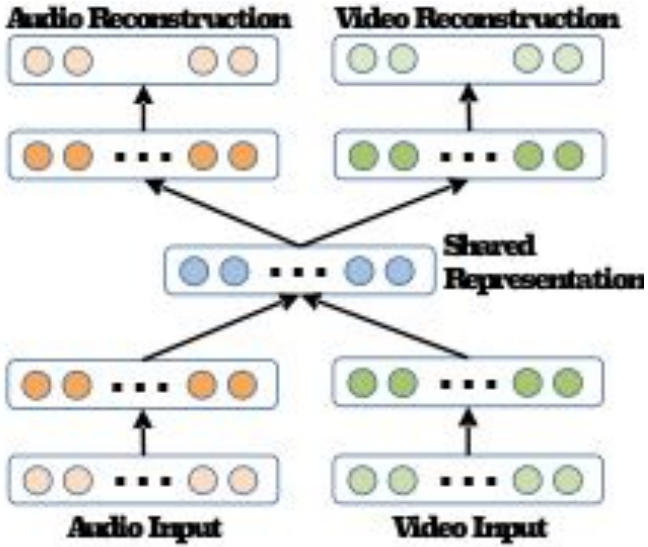


Approaches for learning from multiple intrinsic modalities

Multi-view autoencoders



(a) Video-Only Deep Autoencoder

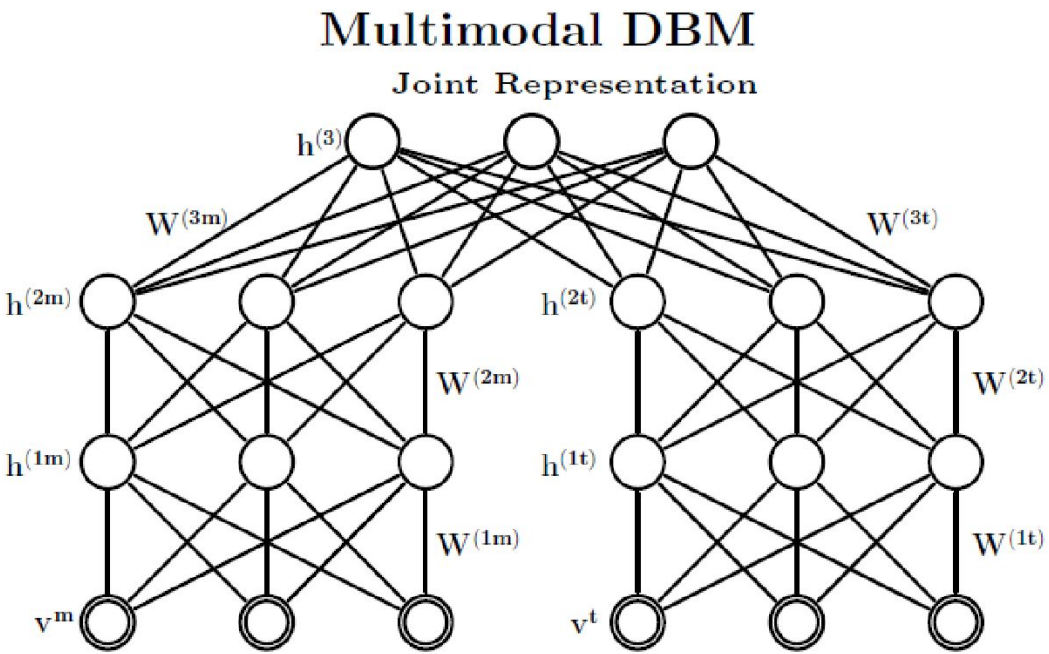


(b) Bimodal Deep Autoencoder

Ngiam et al., "Multimodal deep learning," ICML 2011.

Approaches for learning from multiple intrinsic modalities

Multi-modal deep Boltzmann machines



Srivastava & Salakhutdinov, "Multimodal Learning with Deep Boltzmann Machines," JMLR 2014.

Approaches for learning from multiple intrinsic modalities

Canonical correlation analysis (CCA) [Hotelling 1936]

- One of the oldest and most popular multi-view techniques
- Finds linear projections of each view that maximize the correlation between projected views

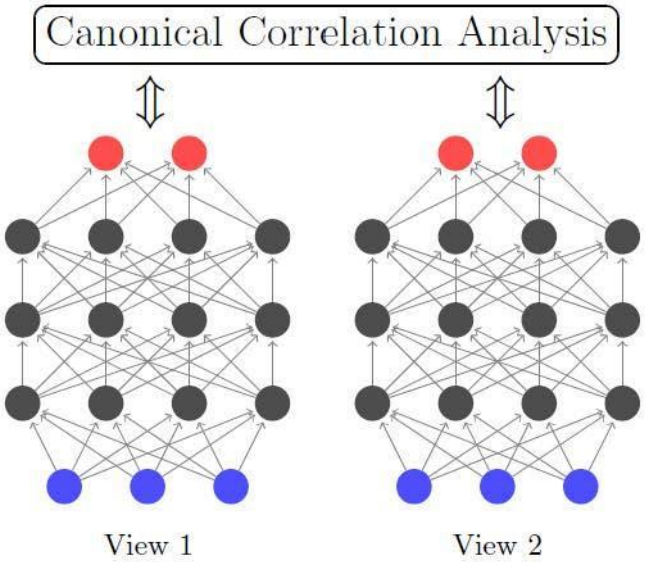
$$v_j, w_j = \arg\max_{v,w} \mathrm{corr}(v^T X, w^T Y)$$

$$\text{s.t. } \mathrm{corr}(v_j^T X, v_k^T X) = 0, \ \mathrm{corr}(w_j^T Y, w_k^T Y) = 0 \quad \text{for } j \neq k$$

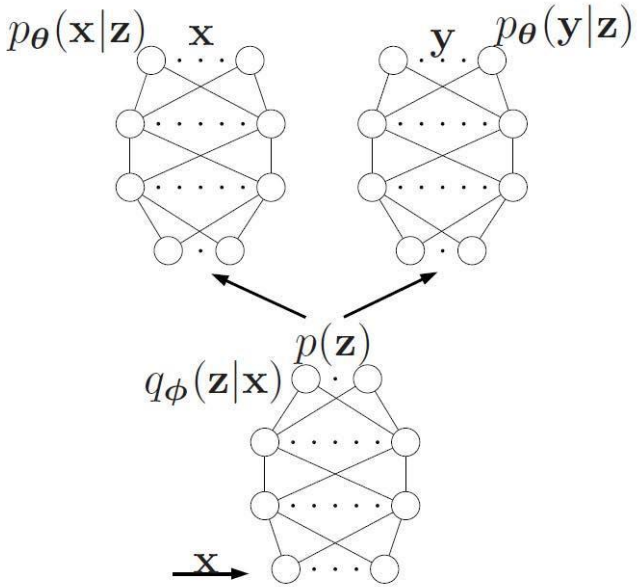
- Like PCA, can be solved via an eigenproblem, and can be interpreted as constrained reconstruction
- Unlike PCA, scale-invariant
- Has “nice” theoretical properties if the noise is uncorrelated across the views
- Has recently been used also for model analysis (see Part 2 of this tutorial)

Approaches for learning from multiple intrinsic modalities

Deep extensions of CCA



Andrew et al., “Deep canonical correlation analysis,” ICML 2013.



Wang et al., “Deep variational canonical correlation analysis,” arXiv:1610.03454, 2016.

Approaches for learning from multiple intrinsic modalities

Multi-view contrastive loss

$$\min_{f,g} \sum_i \max\{\, 0, m + \mathrm{dist}\big(f(x_i^+), g(y_i^+)\big) - \mathrm{dist}\big(f(x_i^+), g(y_i^-)\big) \,\}$$

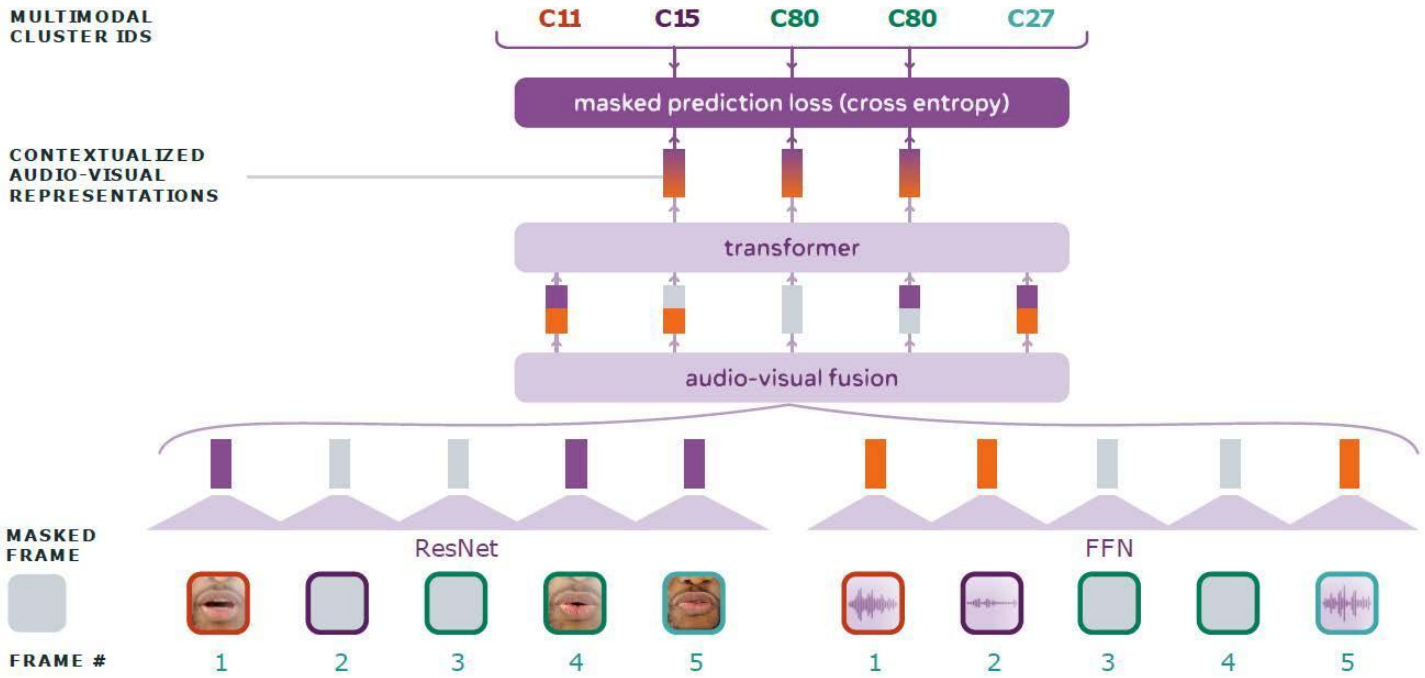
- Where *f* (*x*), *g*(*y*) are arbitrary (e.g., deep neural) functions of inputs in the two modalities, and dist() is some distance function
- Goal: Bring paired examples closer together, while keeping random unpaired examples farther apart by some margin *m*
- Works well when it is easy to find negative examples
- Used by Hermann & Blunsom for learning text representations from multilingual data; has since been applied to a variety of multi-modal settings

Hermann & Blunsom, "Multilingual models for compositional distributed semantics," ACL 2014.

Approaches for learning from multiple intrinsic modalities

AV-HuBERT: An extension of HuBERT with multimodal clusters

- Improves both lip-reading and ASR performance



Shi et al., “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction,” ICLR 2022

Learning with contextual modalities

Learning with contextual modalities

Goal: Learn "higher-level" speech models that encode semantic content

- The contextual modality informs the model what the speech is *about*
- Learning often involves learning to predict matched vs. mismatched multimodal pairs

Example data: Images with spoken captions

Dataset	Lang.	Images	Captions	Speakers	Duration	Speech type
Flickr Audio Captions	en	8000	40000	183	46	Read aloud
Synthetically Spoken COCO	en	123287	616767	1	601	Synthetic
Synthetically Spoken STAIR	ja	123287	616767	1	793	Synthetic
Speech COCO	en	123287	616767	8	601	Synthetic
Places Audio Captions (English)	en	400000	400000	2683	936	Spontaneous
Places Audio Captions (Hindi)	hi	100000	100000	139	316	Spontaneous
SpokenCOCO	en	123287	605000	2352	742	Read aloud

Chrupała, "Visually grounded models of spoken language-A survey of datasets, architectures and evaluation techniques." *Journal of Artificial Intelligence Research,* 2021.

Learning with contextual modalities

Example data: Video description/instruction datasets

- Like spoken captions, the speech is assumed to be “about” the video
- But the modalities are much less well aligned than spoken captions

Dataset	Clips	Captions	Videos	Duration	Source	Year
Charades [48]	10k	16k	10,000	82h	Home	2016
MSR-VTT [58]	10k	200k	7,180	40h	Youtube	2016
YouCook2 [67]	14k	14k	2,000	176h	Youtube	2018
EPIC-KITCHENS [7]	40k	40k	432	55h	Home	2018
DiDeMo [15]	27k	41k	10,464	87h	Flickr	2017
M-VAD [52]	49k	56k	92	84h	Movies	2015
MPII-MD [43]	69k	68k	94	41h	Movies	2015
ANet Captions [26]	100k	100k	20,000	849h	Youtube	2017
TGIF [27]	102k	126k	102,068	103h	Tumblr	2016
LSMDC [44]	128k	128k	200	150h	Movies	2017
How2 [45]	185k	185k	13,168	298h	Youtube	2018
HowTo100M	**136M**	**136M**	**1.221M**	**134,472h**	Youtube	2019

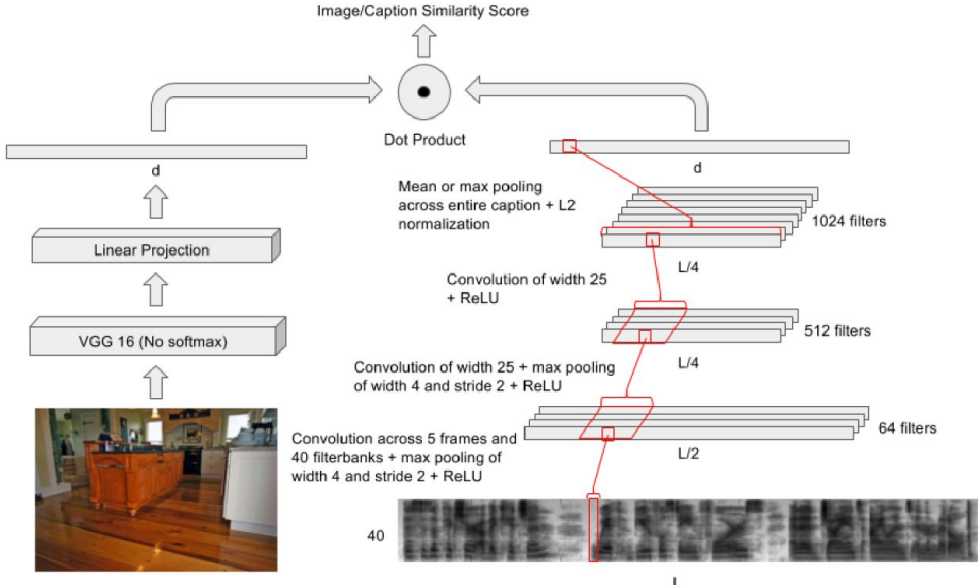
Miech et al., "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips.“ *ICCV* 2019.

Learning to relate images and spoken captions

Approach: Given images and spoken captions, learn an image model and a speech model that produce similar representations for matched image/ caption pairs and dissimilar ones otherwise

Evaluation: Cross-modal retrieval-based

$$\mathcal{L}(\theta) = \sum_{j=1}^{B} \max(0, S_j^c - S_j^p + 1) + \max(0, S_j^i - S_j^p + 1)$$



Synnaeve, Versteegh, & Dupoux, “Learning words from images and speech,” NeurIPS Workshop on Learning Semantics 2014.

Harwath & Glass. "Deep multimodal semantic embeddings for speech and images." ASRU 2015.

Merkx et al., "Language learning using speech to image retrieval," Interspeech 2019.

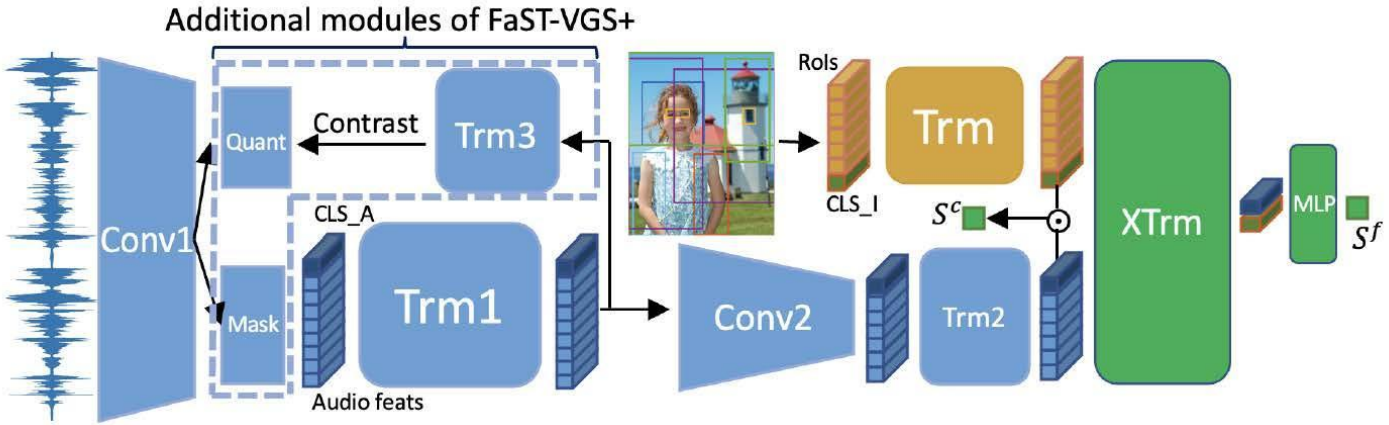
Sanabria et al., "Talk, don’t write: A study of direct speech-based image retrieval," Interspeech 2021.

Harwath, Torralba, & Glass, "Unsupervised learning of spoken language with visual context," *NeurIPS* 2016.

Visually grounded models as pre-trained representations

FaST-VGS and FaST-VGS+:

- Transformer-based models initialized with wav2vec2
- Trained with cross-modal retrieval (+ masked language modeling) loss
- Used as a pre-trained representation model, excels on syntactic & semantic tasks



Peng & Harwath, “Fast-slow transformer for visually grounding speech,” ICASSP 2022.

Peng & Harwath, “Self-supervised representation learning for speech using visual grounding and masked language modeling,” in AAAI SAS workshop, 2022.

References

[Sumby & Pollack 1954] W. H. Sumby & I. Pollack. "Visual contribution to speech intelligibility in noise." *JASA* 26.2 (1954): 212-215.

[Reisbert et al. 1987] D. Reisberg et al.. “Easy to hear but hard to understand: A speechreading advantage with intact auditory stimuli.” In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology lip-reading*. (pp. 97±113). London: Erlbaum, 1987.

[McGurk & McDonald] H. McGurk &J. MacDonald. "Hearing lips and seeing voices." *Nature* 264.5588 (1976): 746-748.

[Rosenbaum 1997] L. D. Rosenblum et al., "The McGurk effect in infants." *Perception & psychophysics* 59.3 (1997): 347-357.

[Mills 1987] A. E. Mills, “The development of phonology in the blind child.” *Hearing by eye: The psychology of lip reading*. Dodd, B. Campbell, R. Hove, UK: Lawrence Erlbaum Associates 145–162.

[Legerstee 1990] M. Legerstee, "Infants use multimodal information to imitate speech sounds." *Infant behavior and development* 13.3 (1990): 343-354.

[Saenko et al. 2009] K. Saenko et al., “Multistream Articulatory Feature-Based Models for Visual Speech Recognition ,” *TPAMI* 2009.

[Zhu et al. 2007] B. Zhu, T. J. Hazen, and J. Glass. “Multimodal speech recognition with ultrasonic sensors” Interspeech 2007.

[Scultz & Wand 2010] Schultz & Wand, “Modeling Coarticulation in EMG-based Continuous Speech Recognition,” In *Speech Communication Journal*, volume 52, 2010.

[Ngiam et al. 2011] Ngiam et al., "Multimodal deep learning," ICML 2011.

[Srivastava & Salakhutdinov 2014] Srivastava & Salakhutdinov, "Multimodal Learning with Deep Boltzmann Machines," *JMLR* 2014.

References

[Hotelling 1936] H. Hotelling. “Relations between two sets of variates.” *Biometrika*, 28:321– 377, 1936.

[Wang et al. 2016] Wang et al., “Deep variational canonical correlation analysis,” arXiv:1610.03454, 2016.

[Andrew et al. 2013] Andrew et al., “Deep canonical correlation analysis,” ICML 2013.

[Hermann & Blunsom 2014] Hermann & Blunsom, “Multilingual models for compositional distributed semantics,” ACL 2014.

[Shi et al. 2022] Shi et al., “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction,” ICLR 2022

[Chrupala 2021] G. Chrupala, "Visually grounded models of spoken language-A survey of datasets, architectures and evaluation techniques." *Journal of Artificial Intelligence Research,* 2021.

[Miech et al. 2019] Miech et al., "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips.“ ICCV 2019.

[Harwath et al. 2016] Harwath, Torralba, & Glass, "Unsupervised learning of spoken language with visual context," NeurIPS 2016.

[Synnaeve et al. 2014] Synnaeve, Versteegh, & Dupoux, “Learning words from images and speech,” NeurIPS Workshop on Learning Semantics 2014.

[Harwath & Glass 2015] Harwath & Glass. "Deep multimodal semantic embeddings for speech and images." ASRU 2015.

[Merkx et al. 2019] Merkx et al., "Language learning using speech to image retrieval,” Interspeech 2019.

[Sanabria et al. 2021] Sanabria et al., "Talk, don’t write: A study of direct speech-based image retrieval," Interspeech 2021.

[Peng & Harwath 2022a] Peng & Harwath, “Fast-slow transformer for visually grounding speech,” ICASSP 2022.

[Peng & Harwath 2022b] Peng & Harwath, “Self-supervised representation learning for speech using visual grounding and masked language modeling,” in AAAI SAS workshop, 2022.

Benchmarking SSL techniques

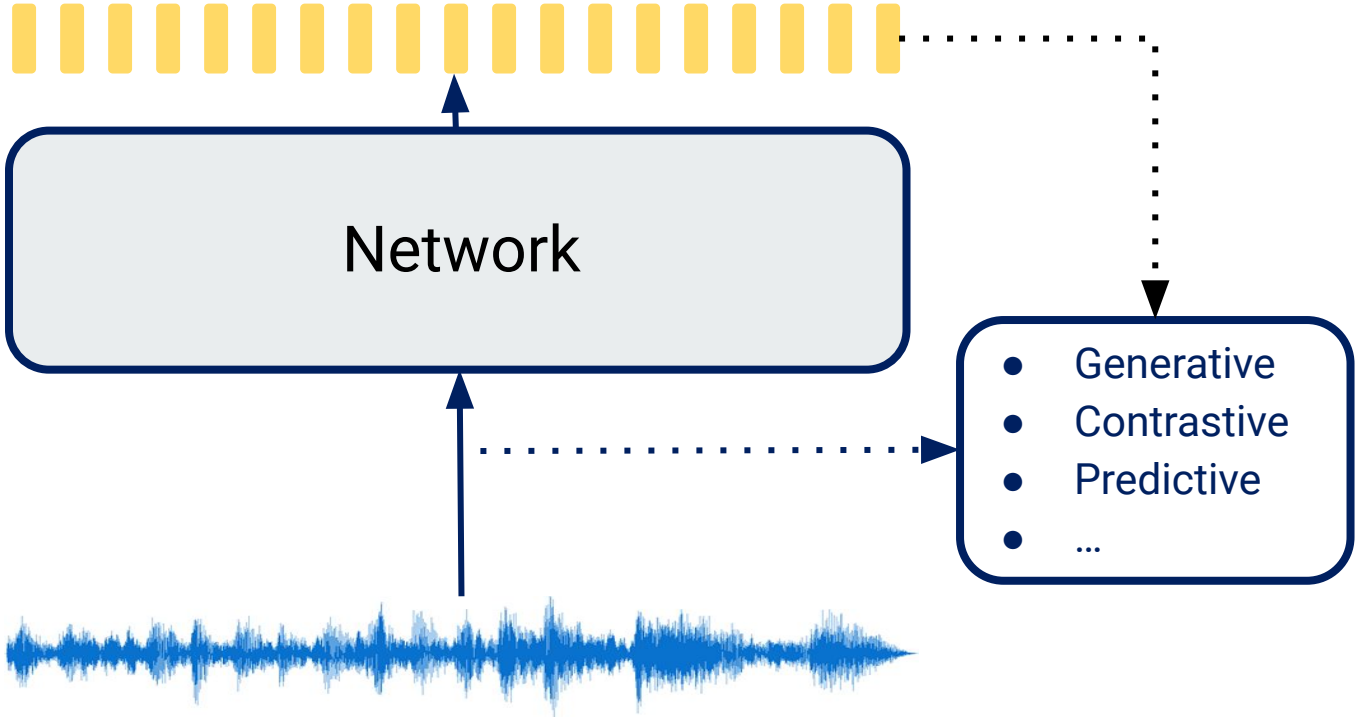
Shang-Wen Li

Benchmarking SSL techniques

- SSL pre-training
- Fine-tuning & evaluation on downstream tasks

SSL pre-training

- Pre-train networks with pretext tasks (described previously, the core of SSL algorithms)



Pre-training datasets

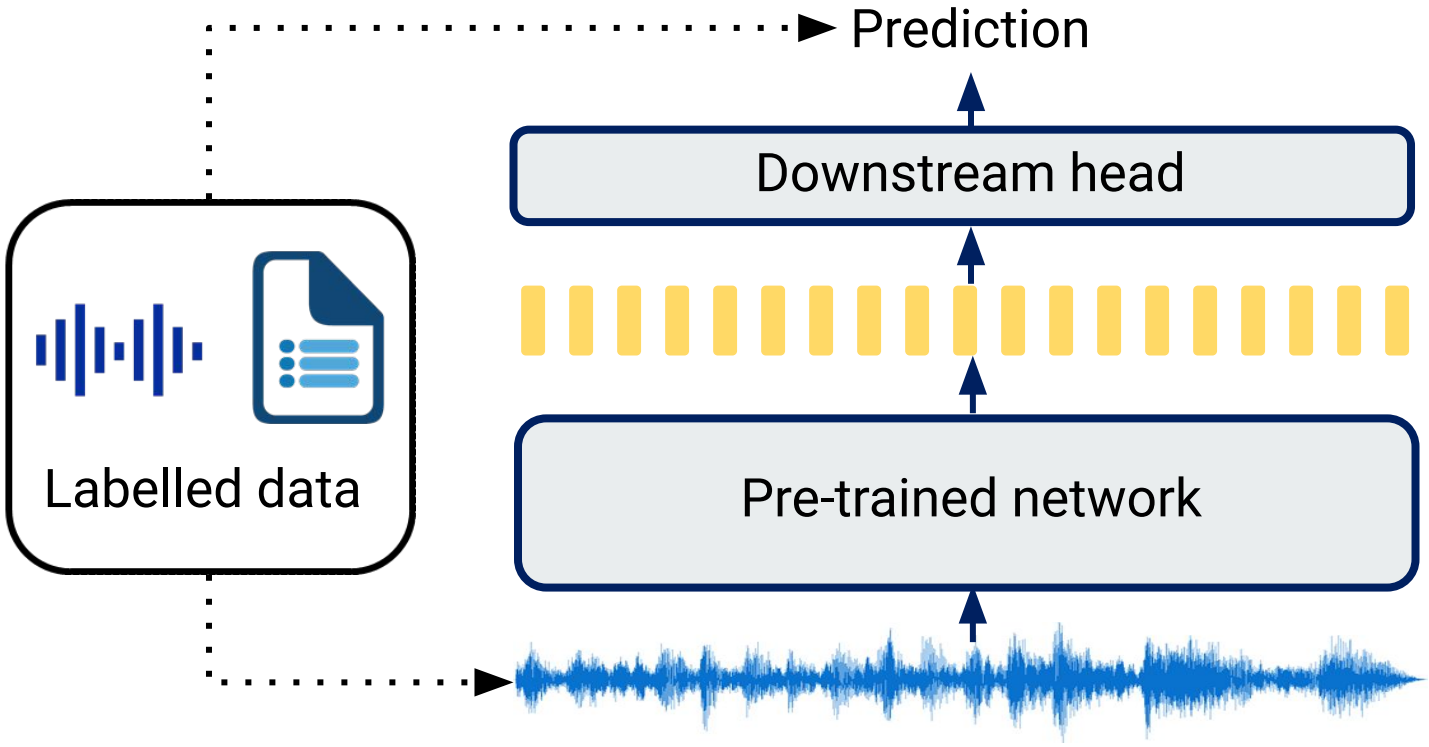
- Sizes
  - LS 50[1]/100[2] hrs
  - LS 960 hrs[3]
  - LL 60k hrs[4]
  - VoxPopuli 400k hrs[5]
  - Alexa (internal) 10k hrs[6]
- w/ or w/o labels
- Sometimes multilingual
- Partial (*) or combined corpora

1. PASE
2. CPC
3. HuBERT, wav2vec 2.0, DeCoAR, TERA, ConvDMM, ...
4. HuBERT, wav2vec 2.0, UniSpeech-SAT, WavLM
5. XLS-R
6. wav2vec-c

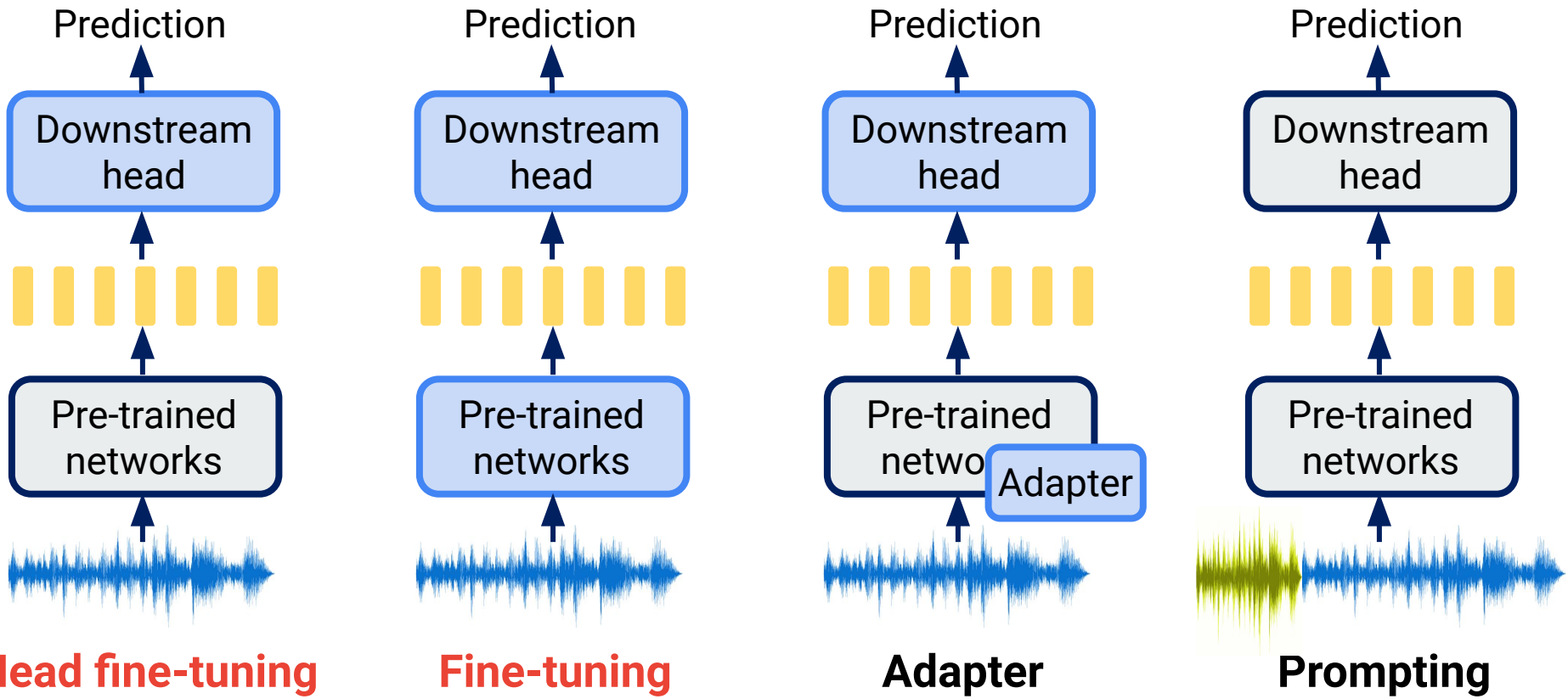
	Dataset	Languages	Size (hrs)	License
Unlabeled	LibriLight (LL)	EN	60k	MIT License
	AudioSet	Multi	2.5k	CC BY 4.0
	AVSpeech	Multi	3.1k	CC BY 4.0
	Fisher	EN	2k	Linguistic Data Consortium (LDC)
	Alexa-10k	EN	10k	Not released
	Didi Callcenter	ZH	10k	Not released
	Didi Dictation	ZH	10k	Not released
Labeled	LibriSpeech (LS)	EN	960	CC BY 4.0
	Wall Street Journal (WSJ)	EN	81	Linguistic Data Consortium (LDC)
	Common Voice (CV-dataset)	Multi	11k*	CC0
	Multilingual LS (MLS)	Multi	50k hrs	CC BY 4.0
	VoxPopuli (VP)	Multi	400k*	CC0
	BABEL (BBL)	Multi	1k	IARPA Babel Agreement
	GigaSpeech	EN	40k*	Apache-2.0 License
	TED-LIUM 3 (TED3)	EN	450	CC BY-NC-ND 3.0
	TED-LIUM 2 (TED2)	EN	118	CC BY-NC-ND 3.0
	Switchboard (SWB)	EN	260	Linguistic Data Consortium (LDC)
	TIMIT	EN	4	Linguistic Data Consortium (LDC)
	VoxLingua107	Multi	6.6k	CC BY 4.0
	Open Mandarin	ZH	1.5k	CC BY-NC-ND 4.0, Apache License v.2.0, Linguistic Data Consortium (LDC)
	HKUST	ZH	168/200	Linguistic Data Consortium (LDC)
	AISHELL-1	ZH	178	Apache License v.2.0

Fine-tuning & evaluation on downstream tasks

- Various fine-tuning techniques, downstream tasks, and scenarios
- Fine-tuning



Fine-tuning techniques



Fine-tuning & evaluation on downstream tasks

- Downstream tasks
  - ASR, PR, SID, Speaker diarization, Speech enhancement, Spoken translation, Emotion recognition, TTS, Event detection …
- Scenarios
  - Full training
  - Few-shot
  - Transfer learning (pre-training <> downstream mismatched)
  - Multilinguality

Datasets for downstream tasks

- Corpora w/ labels
- Popular benchmarks, popular speech processing tasks
- Fine-tune: official or sampled (e.g., few-shot) train/dev set
- Evaluation: official test set*

ASR: automatic speech recognition
PR: phoneme recognition
PC: phoneme classification
SID: speaker identification
ASV: automatic speaker verification
Sentiment: sentiment analysis
ST: speech translation
QbE: query by example or spoken term detection
IC: intent classification
AED: audio event detection
LID: language identification

Dataset	Language	Task	License
LibriSpeech (LS)	EN	ASR/PR/PC/SID	CC BY 4.0
Wall Street Journal (WSJ)	EN	ASR/PR/PC/SID	Linguistic Data Consortium (LDC)
Common Voice (CV-dataset)	Multi	ASR/PR/PC	CC0
Multilingual LS (MLS)	Multi	ASR	CC BY 4.0
VoxPopuli (VP)	Multi	ASR	CC0
BABEL (BBL)	Multi	ASR	IARPA Babel Agreement
GigaSpeech	EN	ASR	Apache-2.0 License
TED-LIUM 3 (TED3)	EN	ASR	CC BY-NC-ND 3.0
TED-LIUM 2 (TED2)	EN	ASR	CC BY-NC-ND 3.0
Switchboard (SWB)	EN	ASR	Linguistic Data Consortium (LDC)
TIMIT	EN	ASR/PR/PC	Linguistic Data Consortium (LDC)
VoxLingua107	Multi	LID	CC BY 4.0
Open Mandarin	ZH	ASR	CC BY-NC-ND 4.0, Apache License v.2.0, Linguistic Data Consortium (LDC)
HKUST	ZH	ASR	Linguistic Data Consortium (LDC)
AISHELL-1	ZH	ASR	Apache License v.2.0
Hub5'00	EN	ASR	Linguistic Data Consortium (LDC)
DIRHA	EN	ASR	https://dirha.fbk.eu/node/107
CHiME-5	EN	ASR	https://chimechallenge.github.io/chime6/download.html
Alexa-eval	EN	ASR	Not released
INTERFACE	Multi	Sentiment	No information
MOSEI	EN	Sentiment	https://github.com/A2Zadeh/CMU-MultimodalSDK/blob/master/LICENSE.txt
VCTK	EN	SID/ASV	CC BY 4.0
VoxCeleb1	Multi	SID/ASV	CC BY 4.0
Fluent Speech Commands (FSC)	EN	IC	CC BY-NC-ND 4.0
QUESST 2014 (QUESST)	Multi	QbE	?
LS En-Fr	En-Fr	ST	CC BY 4.0
CoVoST-2	Multi	ST	CC0
ALFFA	Multi	ASR-multi	MIT License
OpenSLR-multi	Multi	ASR-multi	CC BY-SA 3.0 US, CC BY-SA 4.0, CC BY 4.0, CC BY-NC-ND 4.0, Apache License v.2.0
AED datasets	-	AED	CC BY 4.0 (MUSAN, Speech Commands, NSynth, Bird Audio Detection), CC0 (Spoken Language Identification), Non-Commercial (TUT)

Experiment settings

- ASR (LS, WSJ) and SID are popular
- Also see diversity in downstream tasks and corpora
- Other diversity
  - Pre-training corpora
  - Same/different pre-train, fine-tune corpora
  - Limited fine-tuning data

Work	Pre-training corpus	Task	Dataset		Transfer	Used fine-tuning labels
			Training (fine-tuning)	Test		
CPC	LS 100 hrs	PC	LS 100 hrs	LS 100 hrs	-	80 hrs
		SID	LS 100 hrs	LS 100 hrs	-	80 hrs
PASE	LS 50 hrs	SID	VCTK	VCTK	✓	44 hrs
		Sentiment	INTERFACE	INTERFACE	✓	3 hrs
		PR	TIMIT	TIMIT	✓	4 hrs
		ASR	DIRHA	DIRHA	✓	11 hrs
Audio2Vec	AudioSet	AED	6 AED datasets	6 AED datasets	✓	Check original paper for details
APC	LS 360 hrs	ASR	WSJ si284	WSJ dev93	✓	72 hrs
		ST	LS En-Fr	LS En-Fr	-	236 hrs
		SID	WSJ si284	WSJ si284	✓	65 hrs
wav2vec	LS 80/960 hrs, LS 960 hrs + WSJ si284	ASR	WSJ si284	WSJ eval92	✓	81 hrs
		PR	TIMIT	TIMIT	✓	4 hr
PhasePredict	AudioSet	AED	7 AED datasets	7 AED datasets	✓	Check original paper for details
Bidir-CPC	LS 960 hrs, CPC-8k	ASR	WSJ si284, LS 960 hrs, TED3	WSJ eval92, LS test-clean, LS test-other, TED3, SWB	✓	81/960/450 hrs
		ASR-multi	ALFFA	ALFFA	✓	4 languages, 5.2-18.3 hrs
		ASR-multi	OpenSLR-multi	OpenSLR-multi	✓	21 languages, 4.4-265.9 hrs
MockingJay	LS 360 hrs	PC	LS 360 hrs	LS test-clean	-	0.36/1.8/3.6/18/45/360 hrs
		SID	LS 100 hrs	LS 100 hrs	-	90 hrs
		Sentiment	MOSEI	MOSEI	✓	65 hrs
CPC modified	LS 100 hrs	PC	LS 100 hrs	LS 100 hrs	-	80 hrs
	LS 100 hrs, LS 960 hrs, LL 60k hrs	PC	CV-dataset	CV-dataset	✓	1 hrs
vq-wav2vec	LS 960 hrs	ASR	WSJ si284	WSJ eval92	✓	81 hrs
		PR	TIMIT	TIMIT	✓	4 hrs
DeCoAR	LS 100/360/460/960 hrs, WSJ si284	ASR	WSJ si284	WSJ eval92	-	25/40/81 hrs
		ASR	LS 100/360/460/960 hrs	LS test-clean, LS test-other	-	100/360/460/960 hrs
MT-APC	LS 360 hrs	ASR	WSJ si284	WSJ dev93	✓	72 hrs
		ST	LS En-Fr	LS En-Fr	-	236 hrs
PASE+	LS 50 hrs	PR	TIMIT	TIMIT	✓	4 hrs
		ASR	DIRHA	DIRHA	✓	11 hrs
		ASR	CHiME-5	CHiME-5	✓	50 hrs
AALBERT	LS 360 hrs	PC	LS 100 hrs	LS 100 hrs	-	80 hrs
		SID	LS 360 hrs	LS 360 hrs	-	288 hrs

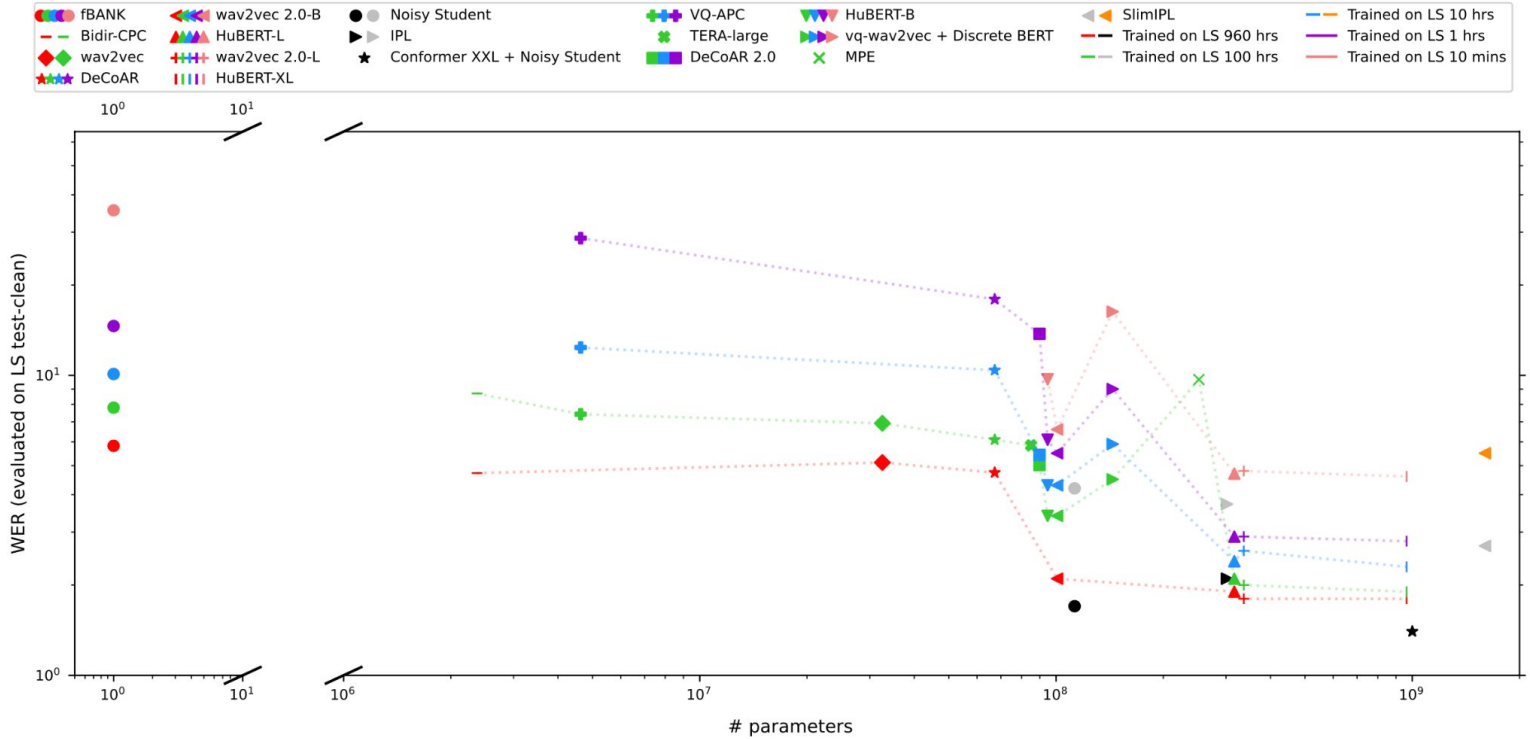
Experiment settings (cont'd)

- ASR (LS, WSJ) and SID are popular
- Also see diversity in downstream tasks and corpora
- Other diversity
  - Pre-training corpora
  - Same/different pre-train, fine-tune corpora
  - Limited fine-tuning data

Work	Pre-training corpus	Task	Dataset		Transfer	Used fine-tuning labels
			Training (fine-tuning)	Test		
BMR	WSJ si284, LS 960 hrs	ASR	WSJ si284	WSJ eval92	-	81 hrs
		PR	WSJ si84/si284	WSJ dev93	-	15/81 hrs
vq-APC	LS 360 hrs	PC	WSJ si284	WSJ dev93	✓	81 hrs
		SID	WSJ si284	WSJ si284	✓	65 hrs
vq-wav2vec + DiscreteBERT	LS 960 hrs	ASR	LS 100 hrs	LS test-clean, LS test-other	-	10 mins, 1/10/100 hrs
speech-XLNet	LS 960 hrs + WSJ si284 + TED2	PR	TIMIT	TIMIT	✓	4 hrs
		ASR	WSJ si284	WSJ eval92	-	7/14/30/81 hrs
MPC	Didi Callcenter, Didi Dictation, Open Mandarin	ASR-zh	HKUST	HKUST	✓	168 hrs
		ASR-zh	AISHELL-1	AISHELL-1	✓	178 hrs
	SWB, Fisher 1k, LS 960 hrs	ASR	SWB	Hub5'00	-	260 hrs
MPE	WSJ si284, LS 960 hrs	ASR	WSJ si284	WSJ eval92	-	25/40/81 hrs
		ASR	LS 100/360/960 hrs	LS test-clean	-	100/360/960 hrs
ConvDMM	LS 50/360/ 960 hrs	PC/PR	WSJ si284	WSJ eval92	✓	5/50/100 mins, 4/8/40 hrs
wav2vec 2.0	LS 960 hrs, LL 60k hrs	ASR	LS 960 hrs	LS test-clean, LS test-other	-	10 mins, 1/10/100/960 hrs
		PR	TIMIT	TIMIT	✓	4 hrs
NPC	LS 360 hrs	PC	WSJ si284	WSJ dev93	✓	81 hrs
		SID	WSJ si284	WSJ si284	✓	65 hrs
DeCoAR 2.0	LS 960 hrs	ASR	LS 100 hrs	LS test-clean, LS test-other	-	1/10/100 hrs
TERA	LS 100/360/ 960 hrs	PC	LS 100 hrs	LS 100 hrs	-	80 hrs
		SID	LS 100 hrs	LS 100 hrs	-	80 hrs
		PR	TIMIT	TIMIT	✓	4 hrs
		ASR	LS 100 hrs	LS test-clean	-	100 hrs
HuBERT	LS 960 hrs, LL 60k hrs	ASR	LS 960 hrs	LS test-clean, LS test-other	-	10 mins, 1/10/100/960 hrs
wav2vec-c	Alexa-10k	ASR	Alexa-eval	Alexa-eval	✓	1k hrs
UniSpeech-SAT	LL 60k hrs + GigaSpeech-10k + VP-24k	Multi	SUPERB	SUPERB	✓	Check SUPERB paper for details
WavLM	LL 60k hrs + GigaSpeech-10k + VP-24k	Multi	SUPERB	SUPERB	✓	Check SUPERB paper for details
XLS-R	VP-400k + MLS + CV-dataset-7k + VL + BBL	ASR	VP, MLS, CV-dataset, BBL, LS	VP, MLS, CV-dataset, BBL, LS	-	Check original paper for detai
		SID	VoxCeleb1	VoxCeleb1	✓	
		ST	CoVoST-2	CoVoST-2	✓	
		LID	VL	VL	-	

Achievement - ASR

- LS (test-clean), SOTA (partial SSL), good perf. @ limited fine-tuning data
- More efficient than pseudo labeling, #param ↑ perf. ↑



Achievement - more tasks

- SSL also yields SOTA in more tasks/corpora beyond ASR
- Expected more as the area gaining attention

Tasks	Datasets	non-SSL SOTA	SSL SOTA
ASR (WER ↓)	LS test-clean/other	2.1/4.0 [1]	1.4/2.6 [2]
IC (Acc ↑)	FSC	98.8 [3]	99.3 [4]
SID (Acc ↑)	VoxCeleb1	94.8 [5]	95.5 [6]
ASV (EER ↓)	VoxCeleb1	3.1 [7]	2.4 [8]
QbE (MTWV ↑)	QUESST (EN)	10.6 [9]	11.2 [4]

The need of consolidated benchmarks

- Too many variables in experiment settings, hard to compare results
- Efforts to establish consistent benchmarks
- SUPERB, LeBenchmark, ZeroSpeech, HEAR, NOSS, and HARES

References

[1] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, “Iterative Pseudo-Labeling for Speech Recognition,” in Proceedings of the Annual Conference of the International Speech Communication Association, 2020.

[2] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, “Pushing the limits of semi-supervised learning for automatic speech recognition, in Workshop on Self-Supervised Learning for Speech and Audio Processing, NeurIPS, 2020.

[3] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, “Speech Model Pre-Training for End-to-End Spoken Language Understanding,” in Proceedings of the Annual Conference of the International Speech Communication Association, 2019.

[4] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, “Unispeech-sat: Universal speech representation learning with speaker aware pre-training,” 2021

[5] M. Hajibabaei and D. Dai, “Unified hypersphere embedding for speaker recognition,” arXiv preprint arXiv:1807.08312, 2018.

[6] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” 2021.

[7] A. Hajavi and A. Etemad, “Siamese capsule network for end-to-end speaker recognition in the wild,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2021.

[8] Y. Wang, A. Boumadane, and A. Heba, “A fine-tuned wav2vec 2.0/hu- bert benchmark for speech emotion recognition, speaker verification and spoken language understanding,” arXiv preprint arXiv:2111.02735, 2021.

[9] L. J. Rodr´ıguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, “Gtts-ehu systems for quesst at mediaeval 2014.” in MediaEval, 2014

Summary of part 1

- Historical Context of Representation Learning
- Speech Representation Learning Paradigms
  - Generative
  - Contrastive
  - Predictive
- Benchmarks for Self-Supervised Learning Approaches
  - SOTA in ASR, IC, SID, ASV, QbE
  - Diverse in experiment settings

Topics in Part 2

- Analysis SSL techniques
- Applying SSL to zero resource understanding
- Better utilization of SSL
- Toolkits for Self-Supervised Speech Representation Learning and demo
- Future directions

Analysis of self-supervised representations

Katrin Kirchhoff

Analysis of Self-Supervised Speech Models

- Impressive performance of SSSMs on variety of downstream tasks
- Can we gain deeper insights into how and why they work?
  - What information is encoded at different layers?
  - How robust are they to distributional shifts in training/test data?
  - How does their performance depend on data, model size, network architecture, or training criteria?
- Do they generalize across languages?
- Do they generalize to related (non-speech) tasks?

Information content at different layers

- What do networks learn at different representational layers?
- Analysis methods:
  - How similar are embeddings extracted from different layers with each other/acoustic features...
    - Canonical correlation analysis
      - find linear projections v,w that maximize correlation between continuously-valued representations *X* and *Y*

        $$v*, w* = argmax_{v,w} corr(v^T X, w^T Y)$$

      - project randomly sampled vectors from X and Y and compute aggregate correlation
    - Cosine similarity: $sim(x, y) = \frac{xy}{||x||||y||}$

  ... or with discrete classes:

    - Mutual information $I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) log \frac{P(x,y)}{P(x)P(y)}$

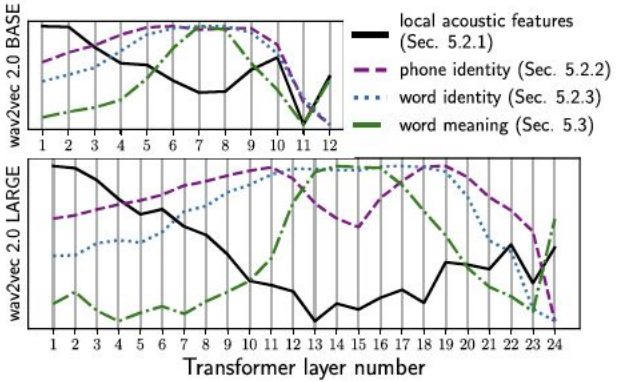
      where x = clustered vector, y = class label

Information content at different layers

- Probing tasks: train downstream classifiers on representations from different layers
  - Phone/character classification, word similarity prediction, speaker identification, …
- Gradient analysis
  - Contributions of weights, hidden states to overall gradient of downstream objective function

Information content at different layers

- Analysis of embeddings from different layers in wav2vec2.0: [Pasad et al., 2021]



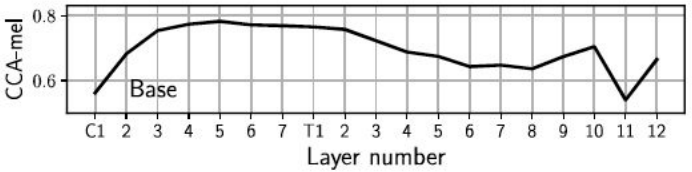


Fig. 4. *CCA similarity between layer representations and fbank; Ci: CNN layer i, Tj: transformer layer j.*

- Without fine-tuning, model show autoencoder-style behavior: early and late layers are more similar to acoustic input features; intermediate layers diverge more strongly

Information content at different layers

- Intermediate layers provide more information about higher-level classes (phone/word information)

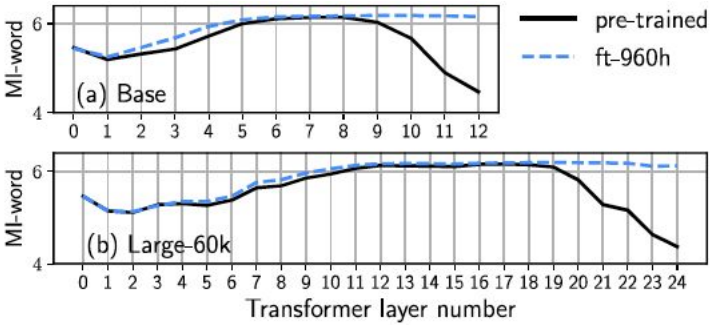


Fig. 6. *MI with word labels (max: 6.2).*

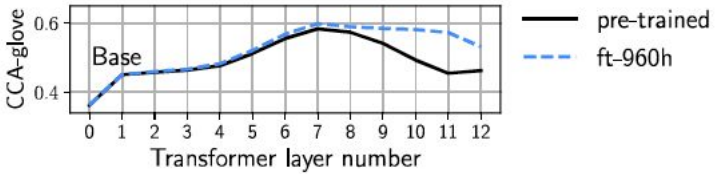


Fig. 8. *CCA similarity with GloVe embeddings.*

From: [Pasad et al. 2021]

- Behavior of top layers changes after fine-tuning: higher-level class information is retained

Information content at different layers

- Information about other categories:
  - Lower layers provide more information about speaker identity [Chen et al., 2021, Chen et al., 2022]
  - Language/accent information provided by higher layers [Ling et al., 2020]

Information content at different layers

- Parallel insights in other domains (language, vision):
  - Transformer-style pretrained language models:
    - Layers retrace ‘steps of traditional NLP pipeline’: POS tagging -> parsing-> named entity recognition -> semantic role labeling -> coreference], showing progression from lower-level to higher-level classes [Tenney et al., 2019]
  - Self-supervised vision transformers:
    - Intermediate layers provide most information about higher-level classes [Grigg et al., 2021]

Effect of training criterion vs. architecture

- [Chung et al., 2021]: Analysis of representations from different pretraining approaches distinguished by:
  - Training criterion (autoregressive predictive coding, contrastive predictive coding, masked predictive coding)
  - Model architecture (CNN/RNN/Transformer)
  - Directionality of input (uni vs. bidirectional)
- Similarity of representations from different models
- Correlation with downstream speaker/phone classification accuracy
- => Training criterion matters most, followed by directionality of input

Effect of training data size

- No surprises here: larger data sets lead to better performance
- Example: 960h LibriSpeech vs. 8k hrs multi-source data set

	LibriSpeech							
	dev-clean		dev-other		test-clean		test-other	
	10%	100%	10%	100%	10%	100%	10%	100%
LibriSpeech								
LogFilterbank (OpenSeq2Seq)	-	6.67	-	18.67	-	6.58	-	19.61
LogFilterbank (ours)	19.83	6.63	38.97	18.77	19.65	6.43	41.26	20.16
CPC-LibriSpeech	15.07	6.70	33.55	19.77	14.96	6.91	36.05	21.60
CPC-8k	**13.92**	**6.20**	**30.85**	**17.93**	**13.69**	**6.25**	**32.81**	**19.10**
+ LM decoding								
LogFilterbank (OpenSeq2Seq)	-	4.75	-	13.87	-	4.94	-	15.06
LogFilterbank (ours)	12.49	4.87	28.71	14.14	12.29	5.04	31.03	15.25
CPC-LibriSpeech	9.66	4.87	24.72	14.34	9.41	5.05	26.77	16.06
CPC-8k	**8.86**	**4.35**	**22.10**	**12.96**	**8.70**	**4.72**	**24.15**	**14.47**

From: [Kawakami et al., 2020]

- Note: larger data sets also imply higher data *diversity*
- Limits of large pretrained speech models not yet tested

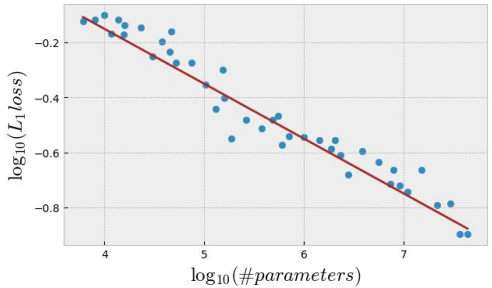
Effect of model size

- How does performance (test loss) depend on model size (number of parameters)? [Pu et al., 2021]
  - Scaling law for autoregressive language models [Kaplan et al., 2020]

$$L(N) = (N_c/N)^{\alpha_N}$$

L: test loss
N: number of parameters
N_c, α_N: constants

- Study for self-supervised speech model (Mockingjay) [Pu et al., 2021]:
  - Varied number of parameters across 5 orders of magnitude (10^3 to 10^7)
  - *L*: reconstruction error for predicted speech frames



$$L(N) = (1778.28/N)^{0.2}$$

Caveat: not the same relationship for downstream tasks; data size may be bottleneck

Robustness to distribution shift

- Comparison with traditional speech features
  - Training on Librispeech (clean read speech); testing on Switchboard/TED Talks (difference in lexical domains, acoustics, speaking styles, accents) [Kawakami et al., 2020]
  - CPC model vs. log filterbank features:

	WSJ		**LibriSpeech**		**Tedlium**		**Switchboard**
	test92	test93	test-clean	test-other	dev	test	eval2000
WSJ							
LogFilterbank	16.78	23.26	46.27	73.27	58.61	62.55	96.44
CPC-LibriSpeech	11.89	15.66	31.05	56.31	45.42	47.79	83.08
CPC-8k	**10.77**	**14.99**	**29.18**	**51.29**	**38.46**	**39.54**	**69.13**
LibriSpeech							
LogFilterbank	14.42	21.08	6.43	20.16	26.9	25.94	61.56
CPC-LibriSpeech	14.28	20.74	6.91	21.6	26.53	27.14	63.69
CPC-8k	**13.31**	**18.88**	**6.25**	**19.10**	**21.56**	**21.77**	**53.02**
Tedlium							
LogFilterbank	20.35	27.23	24.05	47.27	18.75	19.31	74.55
CPC-LibriSpeech	15.01	19.52	17.77	36.7	15.28	15.87	61.94
CPC-8k	**13.17**	**17.75**	**16.03**	**32.35**	**13.67**	**13.88**	**47.69**

Robustness to distribution shift

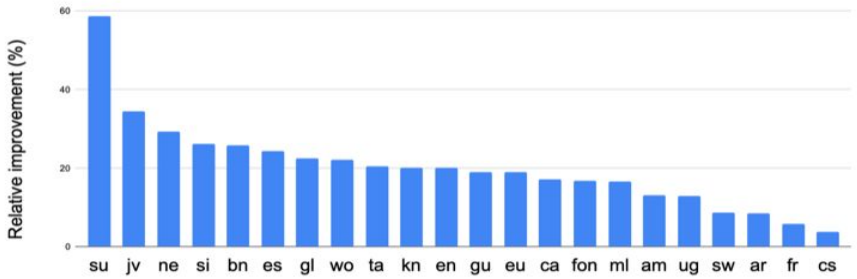
- Analysis of wav2vec 2.0 when trained/tested on 6 different domains [Hsu et al., 2021]
- pretraining on multiple domains is preferable:

PT on X	*WS-dev WER*			*CV-dev WER*			*VP-dev WER*			*Avg WER (over 6 devs)*		
	TD-10h	LS-10h	SB-10h	TD-10h	LS-10h	SB-10h	TD-10h	LS-10h	SB-10h	TD-10h	LS-10h	SB-10h
TD	11.32	9.87	11.10	36.93	34.53	46.04	18.59	16.67	22.41	22.65	20.34	22.85
LS	9.62	9.18	10.10	31.80	31.49	43.63	20.47	18.50	27.64	19.98	19.85	21.76
SF	14.65	12.79	99.25	44.14	43.08	94.53	22.88	24.73	99.97	23.60	22.51	83.09
TD+LS	9.08	8.00	8.95	28.54	27.34	37.59	14.77	14.43	17.77	17.25	16.21	17.63
TD+SF	10.64	9.83	10.01	32.98	32.02	36.09	16.19	16.69	17.25	17.69	16.90	17.90
LS+SF	9.76	8.72	9.32	28.67	28.29	34.12	15.49	16.18	19.60	15.86	15.06	16.97
TD+LS+SF	9.13	8.44	8.94	28.44	27.13	30.92	15.03	15.44	16.91	15.42	14.66	15.37

- pretraining on domain-matched data is best but even out-of-domain data helps close the gap with supervised models

Multilingual generalization

- Do self-supervised representations generalize across languages?
  - Representations from self-trained CPC model (8k mixed-speech data set, incl. 29 different languages) applied to ASR model on 22 different languages
  - Relative improvement over log mel filterbank features:



Improvement extends to languages not represented in the pretraining data!

From: [Kawakami et al., 2020]

Multilingual generalization

- Multilingual vs. monolingual pretraining
  - Multilingual training outperforms monolingual training even for small set of pretraining languages

Model	D	#pt	#ft	es	fr	it	ky	nl	ru	sv	tr	tt	zh	Avg
Number of pretraining hours per language				168h	353h	90h	17h	29h	55h	3h	11h	17h	50h	793h
Number of fine-tuning hours per language				1h	1h	1h	1h	1h	1h	1h	1h	1h	1h	10h
Baselines from previous work														
m-CPC† (Rivière et al., 2020)	LS_{100h}	10	1	38.7	49.3	42.1	40.7	44.4	45.2	48.8	49.7	44.0	55.5	45.8
m-CPC† (Rivière et al., 2020)	LS_{360h}	10	1	38.0	47.1	40.5	41.2	42.5	43.7	47.5	47.3	42.0	55.0	44.5
Fer et al.† (Fer et al., 2017)	BBL_{all}	10	1	36.6	48.3	39.0	38.7	47.9	45.2	52.6	43.4	42.5	54.3	44.9
Our monolingual models														
XLSR-English	CV_{en}	1	1	13.7	20.0	19.1	13.2	19.4	18.6	21.1	15.5	11.5	27.1	17.9
XLSR-Monolingual	CV_{mo}	1	1	6.8	10.4	10.9	29.6	37.4	11.6	63.6	44.0	21.4	31.4	26.7
Our multilingual models														
XLSR-10 (unbalanced)	CV_{all}	10	1	9.7	13.6	15.2	11.1	18.1	13.7	21.4	14.2	9.7	25.8	15.3
XLSR-10	CV_{all}	10	1	9.4	14.2	14.1	8.4	16.1	11.0	20.7	11.2	7.6	24.0	13.6
XLSR-10 (separate vocab)	CV_{all}	10	10	10.0	13.8	14.0	8.8	16.5	11.6	21.4	12.0	8.7	24.5	14.1
XLSR-10 (shared vocab)	CV_{all}	10	10	9.4	13.4	13.8	8.6	16.3	11.2	21.0	11.7	8.3	24.5	13.8
Our multilingual models (Large)														
XLSR-10	CV_{all}	10	1	7.9	12.6	11.7	7.0	14.0	9.3	20.6	9.7	7.2	22.8	12.3
XLSR-10 (separate vocab)	CV_{all}	10	10	8.1	12.1	11.9	7.1	13.9	9.8	21.0	10.4	7.6	22.3	12.4
XLSR-10 (shared vocab)	CV_{all}	10	10	7.7	12.2	11.6	7.0	13.8	9.3	20.8	10.1	7.3	22.3	12.2
Our Large XLSR-53 model pretrained on 56k hours														
XLSR-53	D_{53}	53	1	**2.9**	**5.0**	**5.7**	**6.1**	**5.8**	**8.1**	**12.2**	**7.1**	**5.1**	**18.3**	**7.6**

From: [Conneau et al., 2020]

From speech to sounds

- Can self-supervised representations generalize to non-speech sounds/other acoustic time series?
- Hear 2021 Neurips challenge: https://neuralaudio.ai/:
  - 19 downstream tasks (5 open, 14 blind) for classification of speech and non-speech audio signals
    - Emotion recognition, speech recognition, pitch prediction, environmental sound classification, music genre classification, ….
  - Fixed downstream classifiers; same pretrained representation needed to be used for all tasks
  - 29 submitted representation learning models

From speech to sounds

- Results:
  - No single representation is robust across all tasks
  - Speech representations perform reasonably well on related tasks (language identification, emotion detection)
  - Some tasks (e.g., vocal imitations) see generally poor performance – no standard pretraining data
  - => as of yet, no self-supervised model for general-purpose audio processing

References

- [Chen et al., 2021] S. Chen et al., 2021, WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing, https://arxiv.org/abs/2110.13900
- [Chen et al., 2022] S. Chen et al., 2022, Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition?, https://arxiv.org/abs/2204.12765
- [Conneau et al., 2020] A. Conneau et al., Unsupervised cross-lingual representation learning for speech recognition, 2020, https://arxiv.org/abs/2006.13979
- [Chung et al., 2021] A. Chung, Y. Belinkov, J. Glass, Similarity Analysis of Self-Supervised Speech Representations, ICASSP 2021
- [Grigg et al., 2021] T. G. Grigg et al., Do Self-Supervised and Supervised Methods Learn Similar Visual Representations, 2nd Workshop on Self-Supervised Learning: Theory and Practice, NeurIPS 2021, https://arxiv.org/abs/2110.00528
- [Hsu et al., 2021] W.N. Hsu et al., Robust Wavevec 2.0: Analyzing domain shift in self-supervised pre-training, Interspeech 2021
- [Kaplan et al., 2020] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” arXiv preprint arXiv:2001.08361, 2020
- [Kawakami et al., 2020] K. Kawakami et al., Learning Robust and Multilingual Speech Representations, 2020, https://arxiv.org/abs/2001.11128
- [Ling et al., 2020] S. Ling, J. Salazar, Y. Liu, K. Kirchhoff, BERTphone: Phonetically-aware encoder representations for utterance-level speaker and language recognition, Odyssey Speaker and Language Recognition workshop, 2020
- [Pu et al., 2021] J. Pu et al., Scaling Effect of Self-Supervised Speech Models, Interspeech, 2021
- [Tenney et al., 2019] I. Tenney et al., BERT rediscovers the classical NLP pipeline, ACL 2019

From Representation Learning to Zero Resources

Shinji Watanabe

From Representation Learning to Zero Resources

Shinji Watanabe

1. Unsupervised Speech Recognition
2. ASR-TTS Technique
3. Zero Resource Speech Technologies and Challenges
4. Textless NLP

Overview of this section

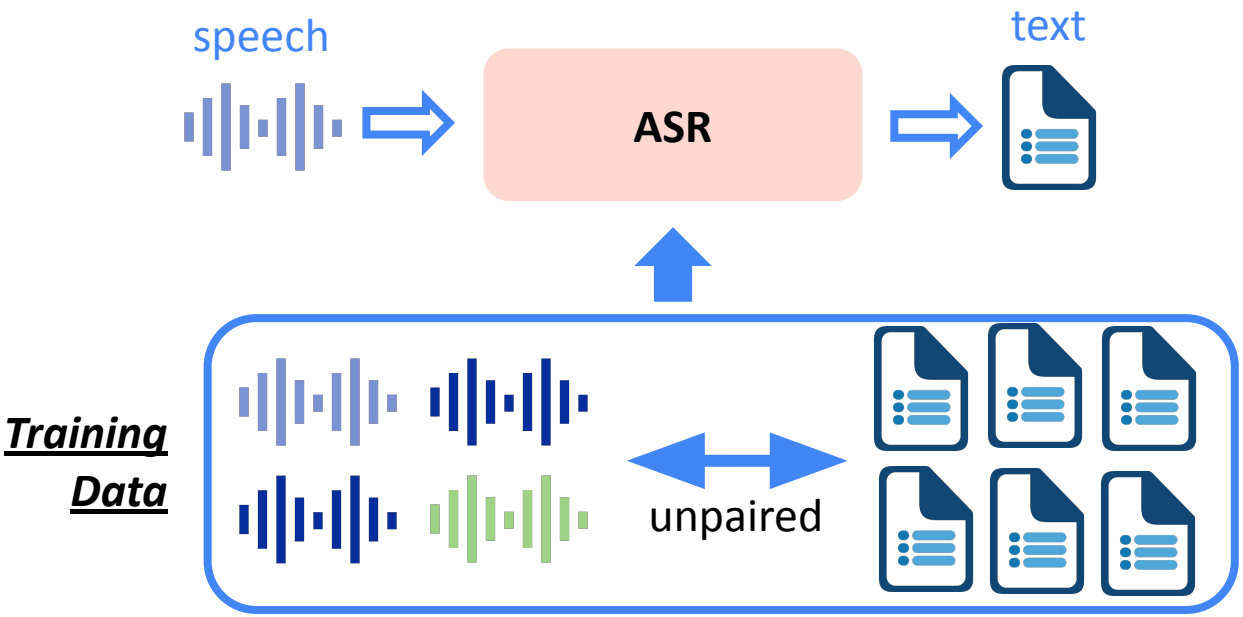
- Previous sections mainly discuss the effectiveness of SSLR for downstream tasks with the **paired** data ← This section relaxes the conditions!
- This section introduces several application of using the **unpaired** data
  - Unsupervised speech recognition
  - ASR-TTS technique
  - Zero resource speech technologies and challenges
  - Textless NLP

From Representation Learning to Zero Resources

Shinji Watanabe

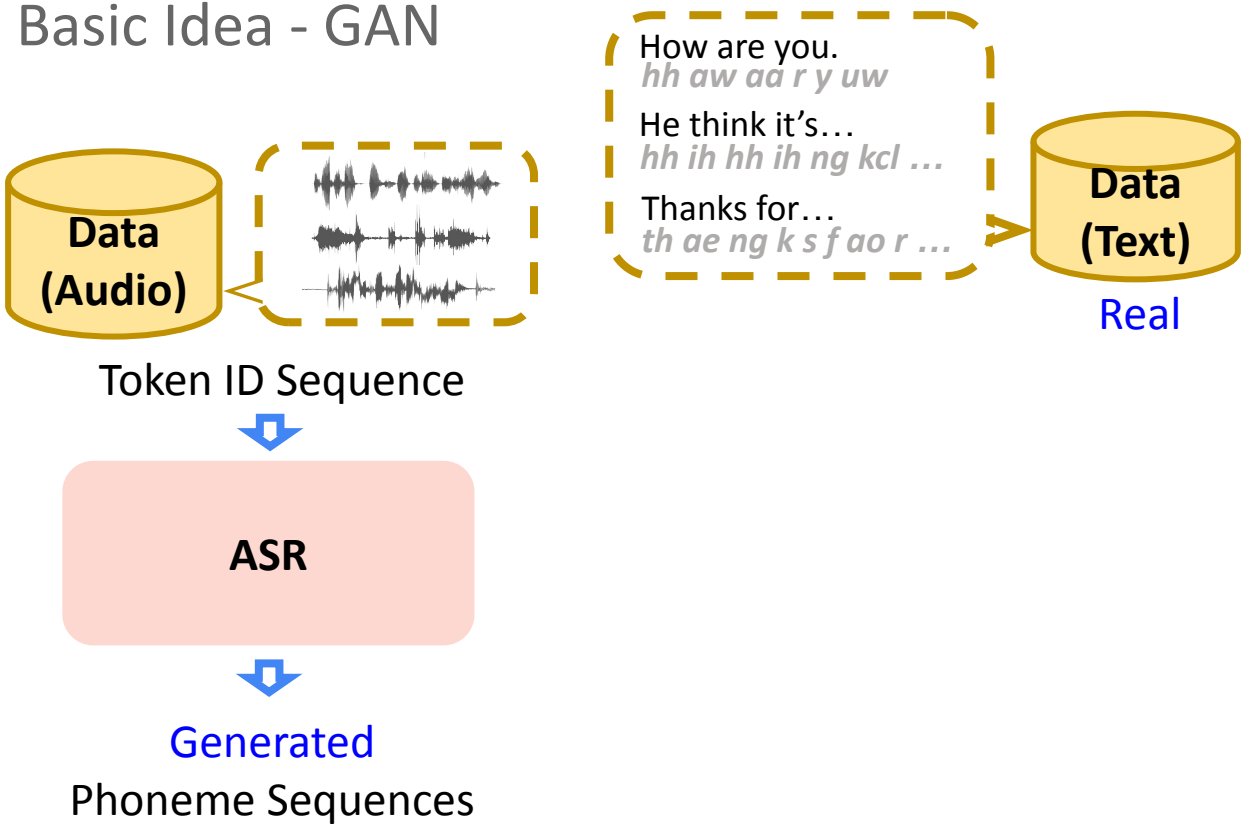
1. Unsupervised Speech Recognition
2. ASR-TTS Technique
3. Zero Resource Speech Technologies and Challenges
4. Textless NLP

Unsupervised Speech Recognition

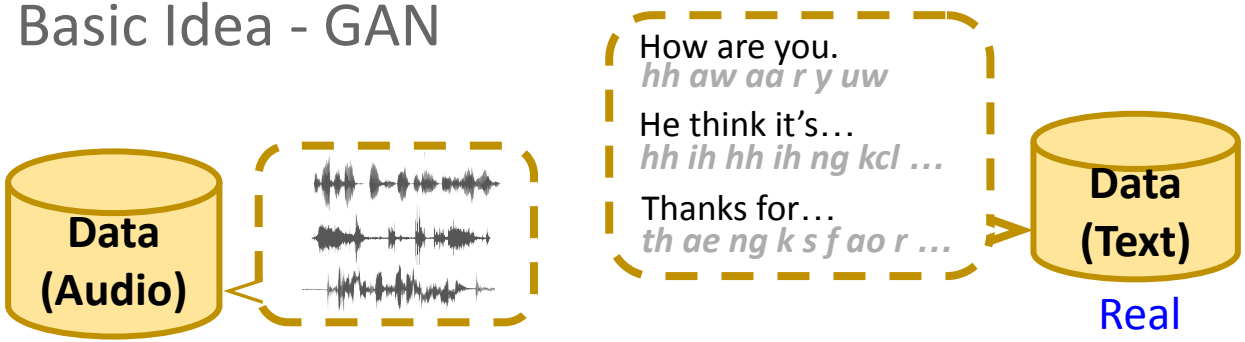


This can be achieved by Generative Adversarial Network (GAN).

Basic Idea - GAN

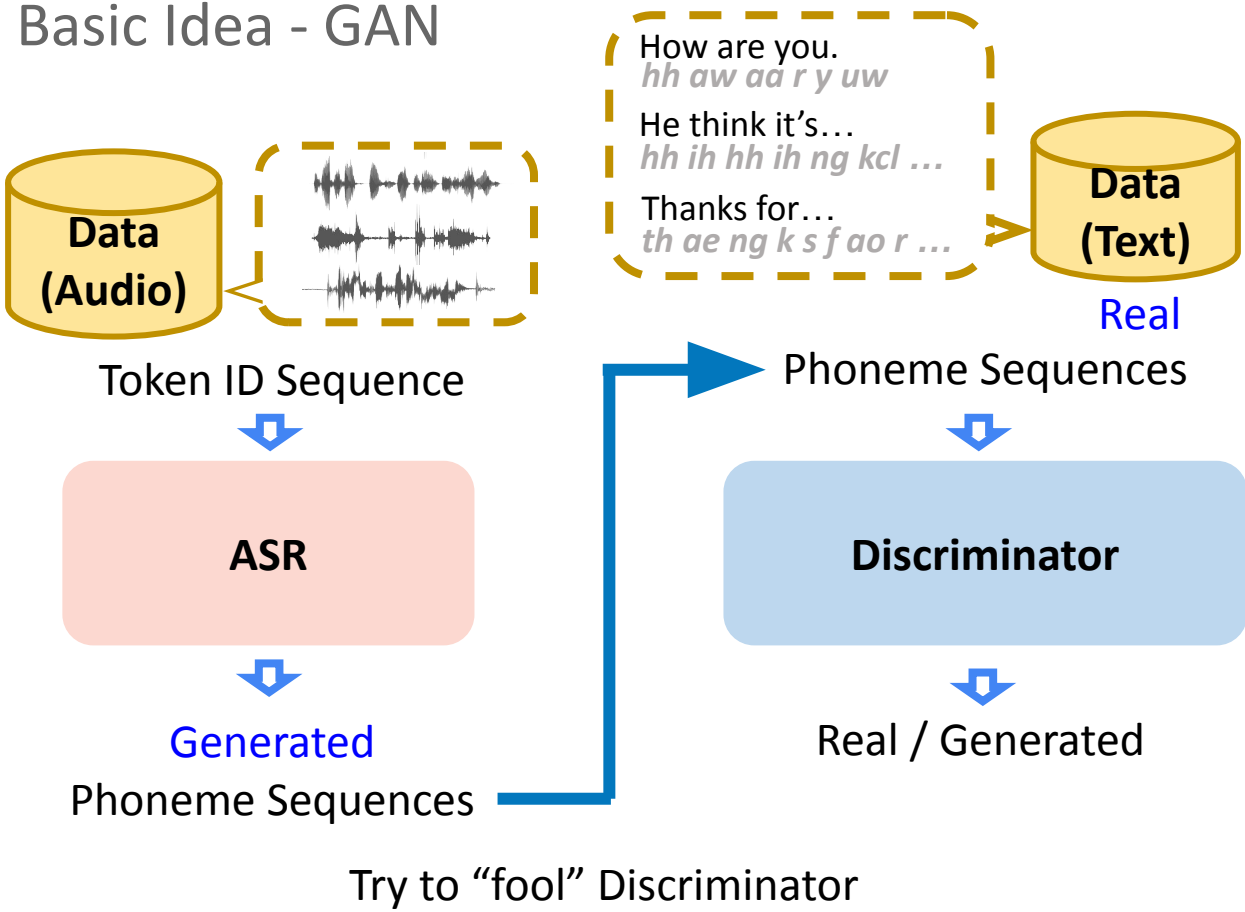


Basic Idea - GAN

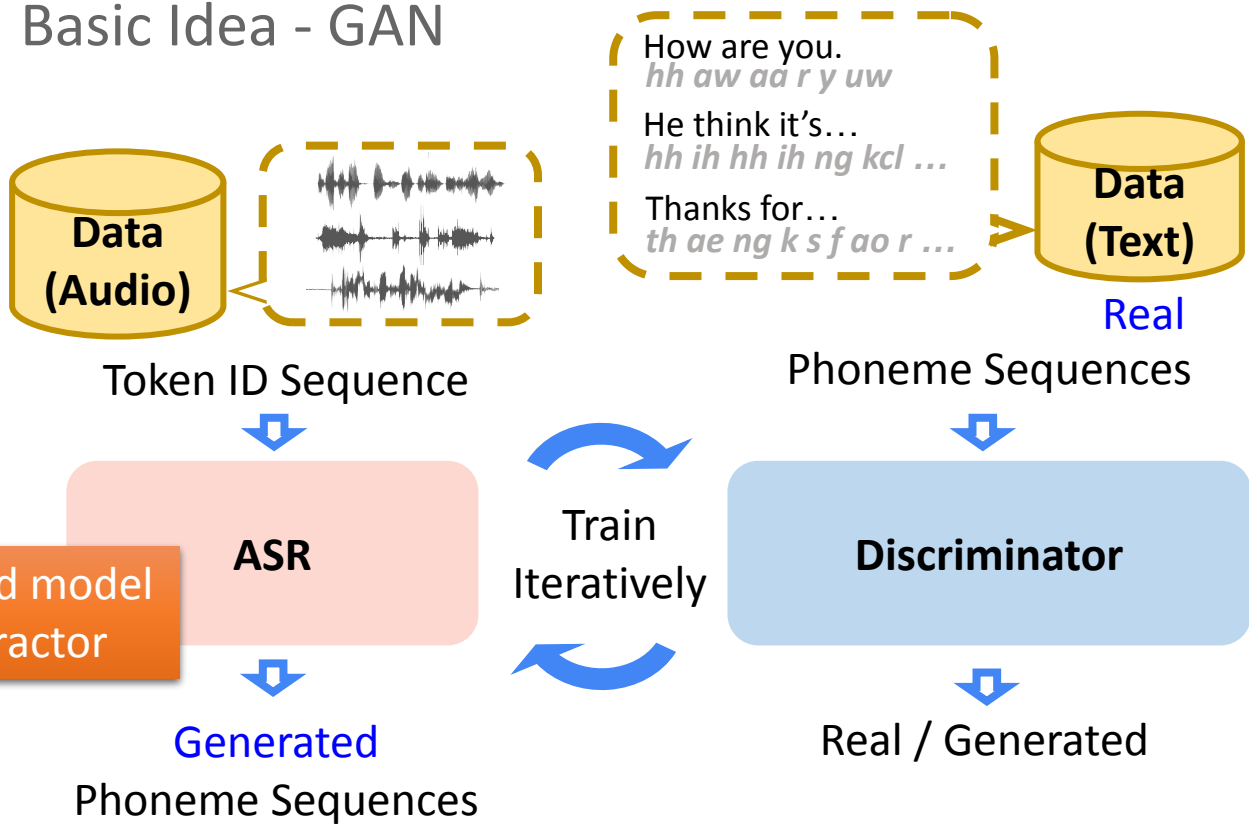


To distinguish *real* or *generated* phoneme sequence.

Basic Idea - GAN

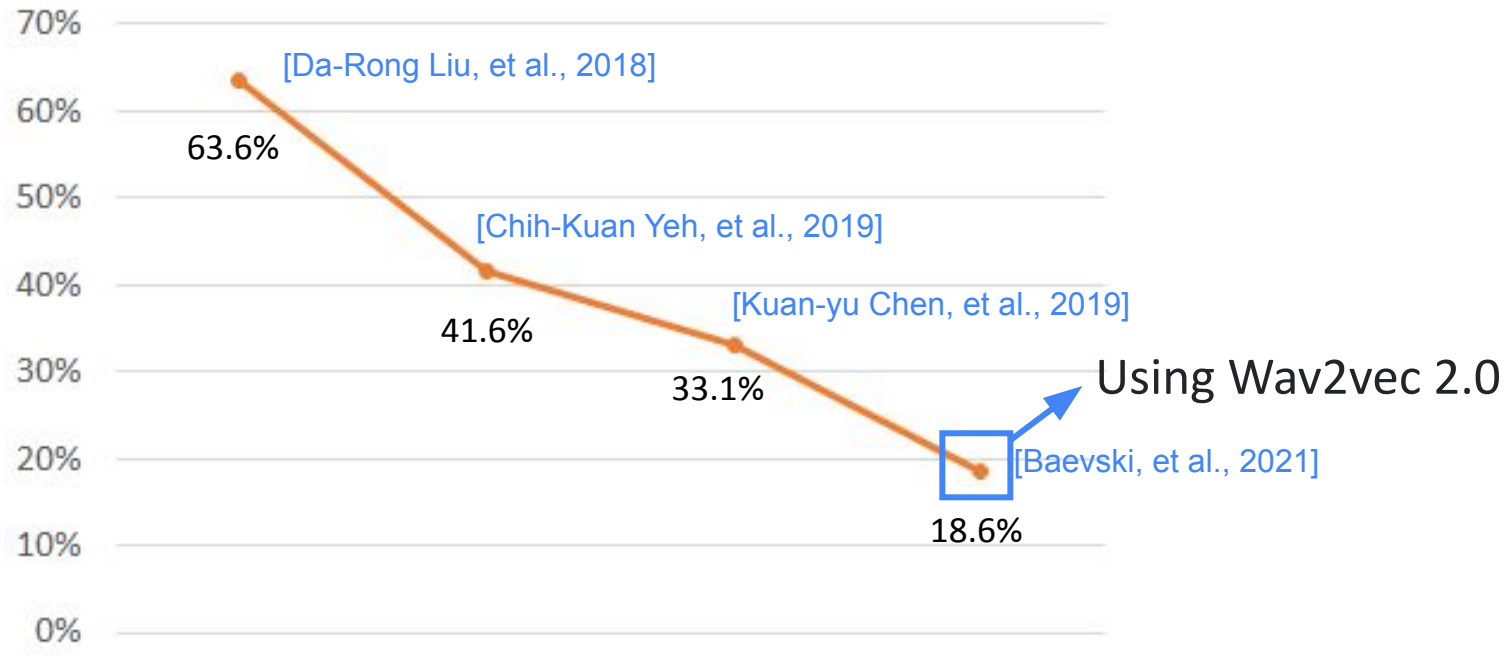


Basic Idea - GAN



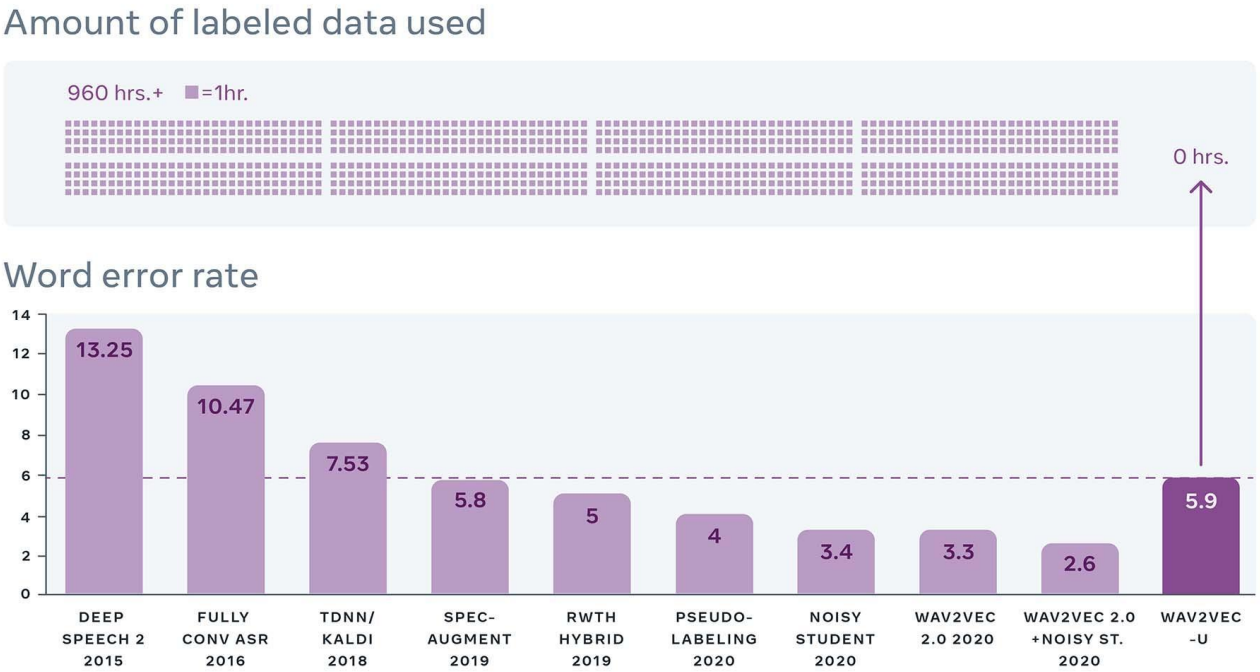
Unsupervised Speech Recognition - with SSL

TIMIT (Phoneme Error Rate)



Unsupervised Speech Recognition - with SSL

Librispeech



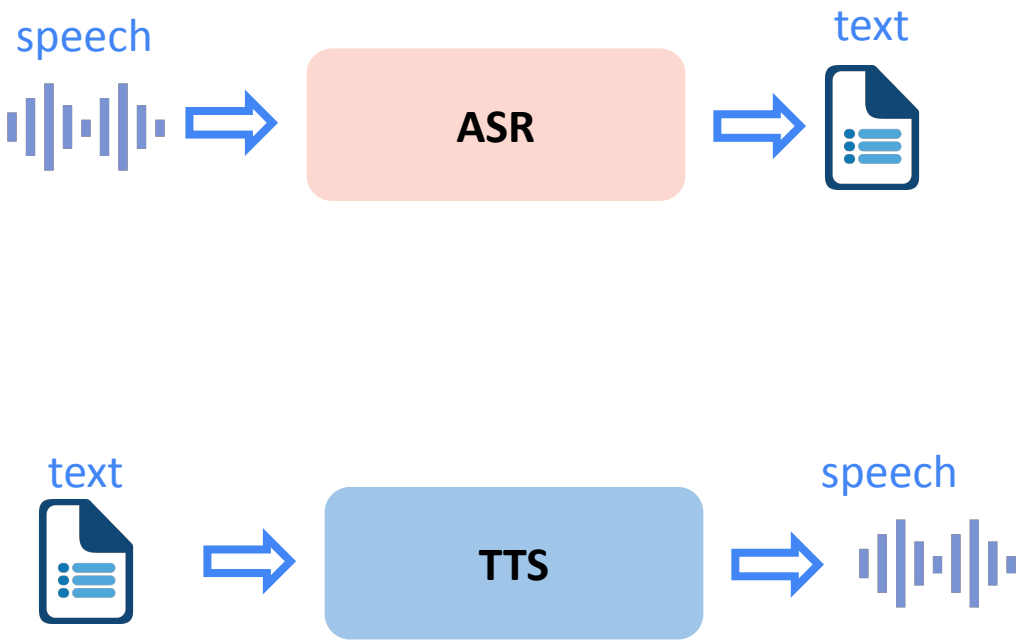
https://ai.facebook.com/blog/wav2vec-unsupervised-speech-recognition-without-supervision/

From Representation Learning to Zero Resources

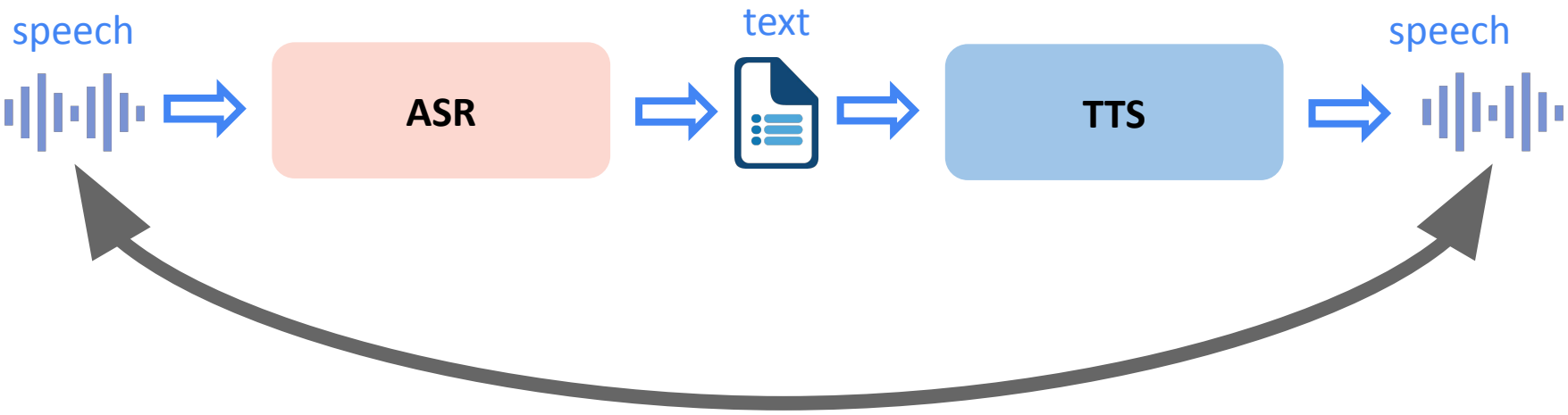
Shinji Watanabe

1. Unsupervised Speech Recognition
2. ASR-TTS Technique
3. Zero Resource Speech Technologies and Challenges
4. Textless NLP

ASR-TTS

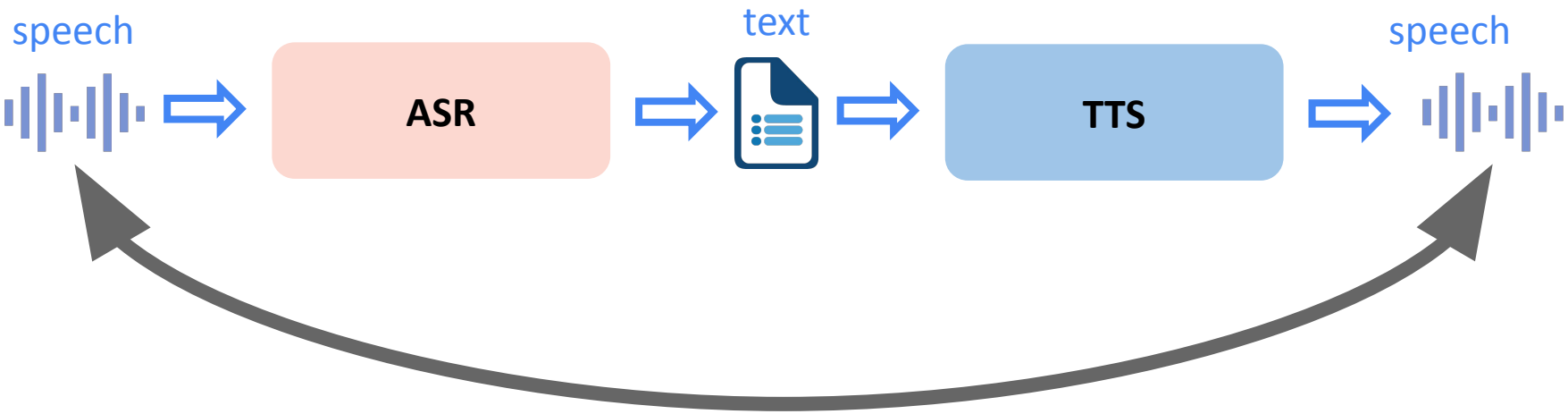


ASR-TTS



$$L(S, \hat{S}; \theta^{\text{asr}}, \theta^{\text{tts}})$$

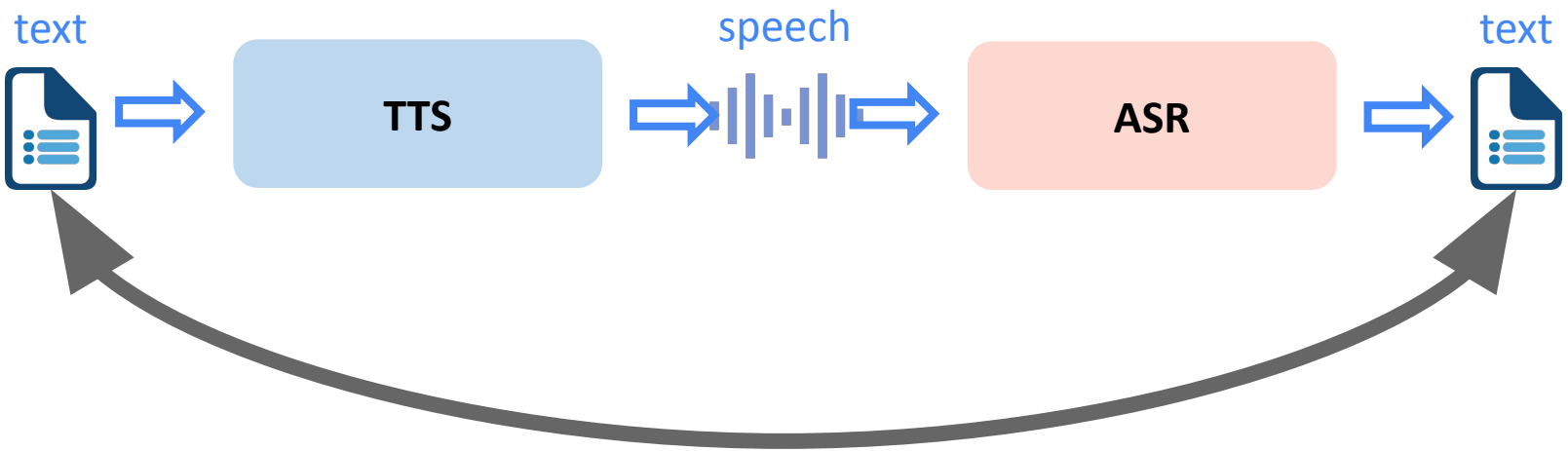
ASR-TTS



$$L(S, \hat{S}; \theta^{\text{asr}}, \theta^{\text{tts}})$$

- We can train both ASR and TTS only with the speech data
- The text data can be regarded as latent representation

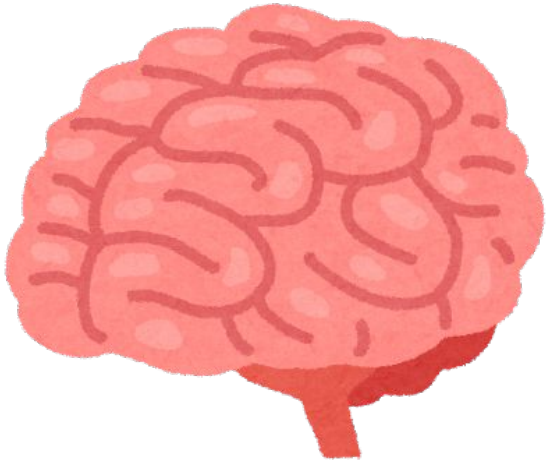
TTS-ASR



$$L(S, \hat{S}; \theta^{\text{asr}}, \theta^{\text{tts}})$$

- Do the same loop in the text only data

Speech recognition and synthesis



Joint modeling of speech recognition and synthesis is a very important concept in **neuroscience**

- **Phonological loop**
- **Speech chain**
- **Motor theory**

ASR-TTS framework

- Speech Chain [Tjandra+(2017)]
  - First success of joint training of ASR-TTS and TTS-ASR
  - Freezing TTS parameters while estimating ASR parameters and vice versa
- Cycle Consistency Training [Hori+(2019), Baskar+(2019)]
  - Jointly optimize both ASR and TTS parameters
  - Difficulty of obtaining gradients especially in after ASR due to the argmax operation

  → REINFORCE or Gamble softmax

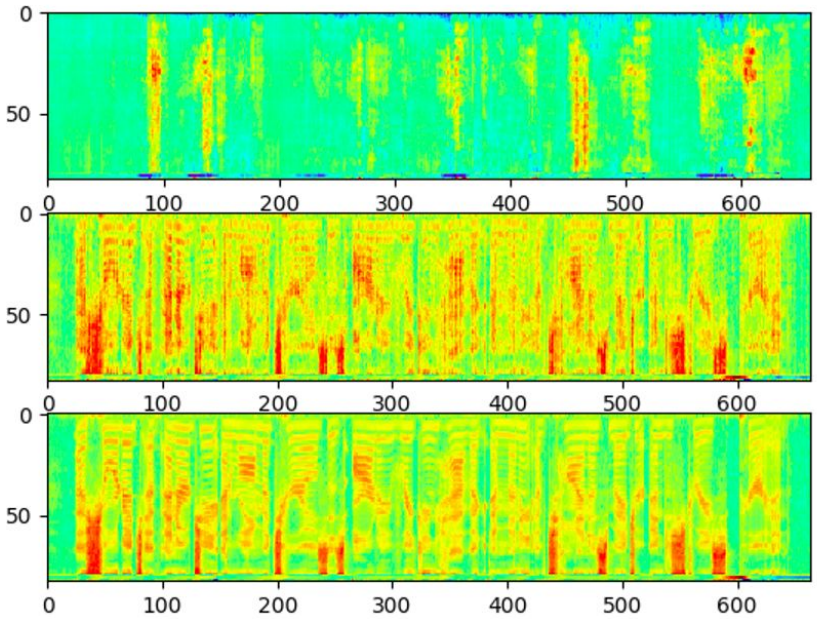
Note that this framework requires an initial ASR and TTS models trained with the paired data

Experimental results [Hori+(2019), Baskar+(2019)]

- English Librispeech corpus
  - Paired data: 100h to train ASR and TTS [Shen+ (2018)] models first
  - Unpaired data: 360h (only audio and/or text only): cycle consistency training

Model	Eval-clean WER [%]
Baseline	20.7
+ audio-only cycle E2E	17.0
+ both audio-only/text-only cycle E2E	**16.6**

Improving TTS quality as well!

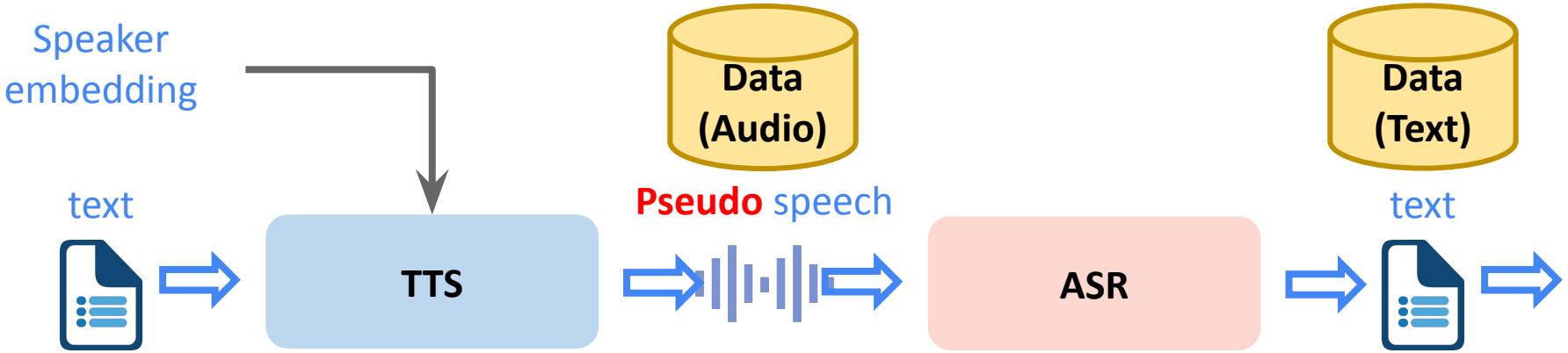


REFERENCE TEXT:

“has never been surpassed”

- Initial epoch
- Final epoch

Data augmentation views



By using various text only data and various speaker embeddings, we can generate huge amount of paired training data [Li+(2018), Ueno+(2019), Rosenberg+(2019), Laptev+(2019) etc.]

Experimental results [Ueno+(2019)]

- Corpus of Spontaneous Japanese (CSJ)
  - Paired data: Academic lecture 247h to train ASR and TTS models first
  - Unpaired data: Spontaneous lecture 281h (text only):

Model	Eval-Spontaneous WER [%]
Baseline	18.8
+ single speaker TTS augmentation	14.2
+ multi speaker TTS augmentation	**13.3**

- Feature level and phoneme level or data augmentations [Hayashi+(2018), Renduchintala+(2018)] have been studied.

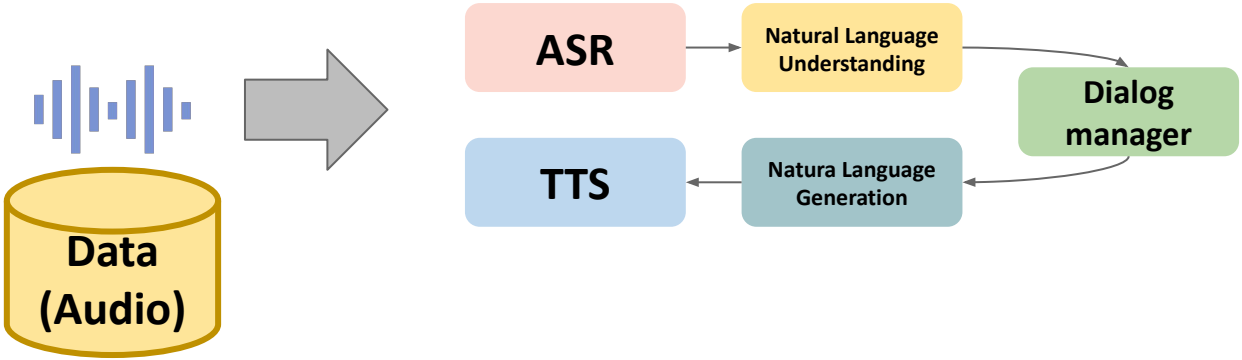
From Representation Learning to Zero Resources

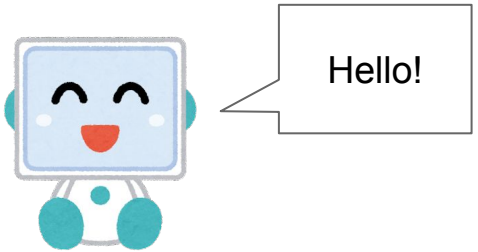
Shinji Watanabe

1. Unsupervised Speech Recognition
2. ASR-TTS Technique
3. Zero Resource Speech Technologies and Challenges
4. Textless NLP

Toward zero resource speech technologies

- Aim at discovering linguistic concepts from audio only (no text nor lexicon)
  - Phonemes, words, syntactic structure, semantics
- Inspired by the infant learning process
- Build spoken dialog systems in a zero resource setup for any language



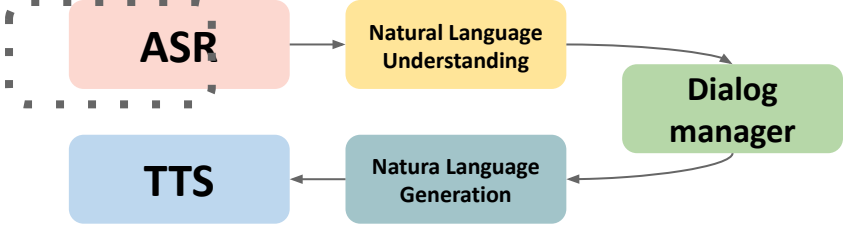


A lot of difficulties

- Technological difficulties
  - It is very difficult to build a dialog system only from the speech data
  - Need an appropriate milestones
- Common benchmarks
  - Compare the various technologies
- Evaluation metrics
  - It is difficult to map the obtained results with actual phonetic/linguistic representations

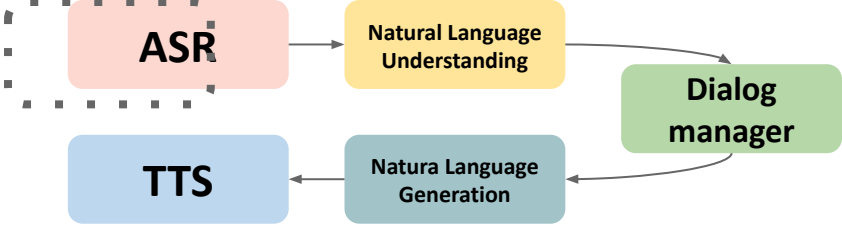
ZeroSpeech Challenge was launched to guide the community to tackle this challenging problems

ZeroSpeech Challenge 2015



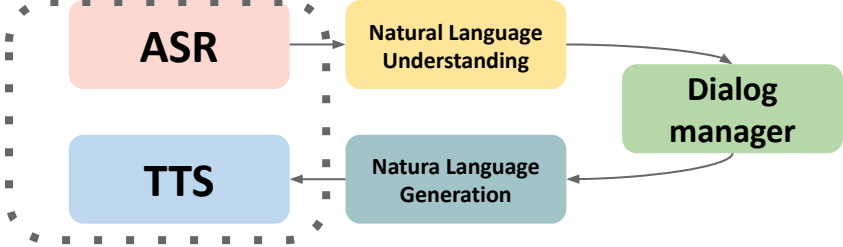
- First challenge (Special Session at Interspeech 2015)
- **Task target:** Build an **acoustic model** without using any linguistic annotations
- **Evaluation metric**:
  - **ABX score** within/across the speaker for the subword unit modeling track
    - Suppose speech features *a* and *x* for category *A*, and speech features *b* for category *B*
    - This should satisfy DTW(*a*, *x*) < DTW(*b*, *x*)
  - **F1-scores** (normalized edit distance and coverage scores) for the spoken term discovery track
- **Data**: Buckeye corpus of conversational **English** and **Xitsonga** section of the NCHLT corpus of South Africa's languages
- **Participants:** 7+ teams
- **Technical trends:** Top-down (spoken term discovery → Bottleneck feature), Dirichlet process Gaussian mixture model (DPGMM) with MFCC, Investigation of various clustering techniques with various features

ZeroSpeech Challenge 2017



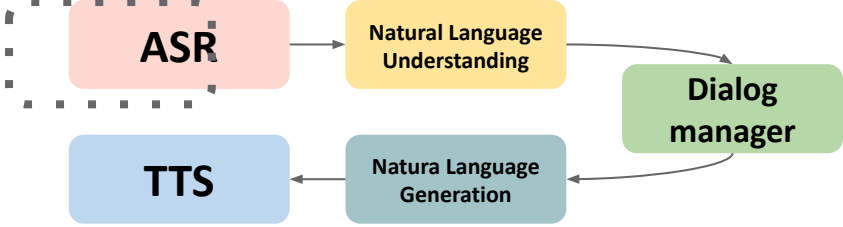
- **Task target:** Focus on an **acoustic model** with unseen language and speaker aspects
- **Evaluation metric**: ABX based on DTW (but the participants can customize the local metric), Spoken term detection metrics (F measure)
- **Data**: English and French from LibriVox, Mandarin (Thchs-30), and two surprising languages (German (Librivox) and Wolof (ALFFA))
- **Participants:** 9+ teams
- **Technical trends:** Autoencoder, top-down (DPGMM/GMM-HMM → Bottleneck feature, and its iterations)

ZeroSpeech Challenge 2019



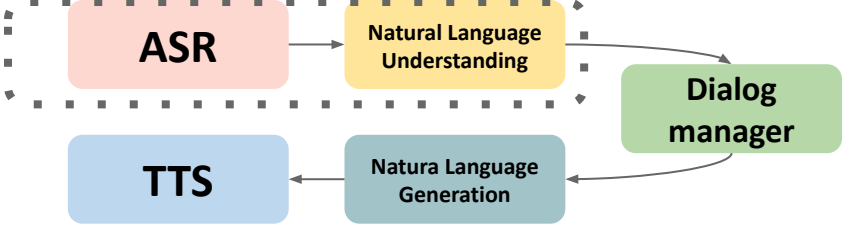
- **Task target:** Focus on an **ASR and TTS models without text**
- **Evaluation metric**: TTS quality measures (intelligibility, naturalness, speaker similarity), CER, ABX, and bitrate
- **Data**: English (development set), Indonesian (test set, A-STAR)
- **Participants:** 11 teams
- **Technical trends:** VAE, VQ-VAE, top-down → TTS, Auto-encoder

ZeroSpeech Challenge 2020



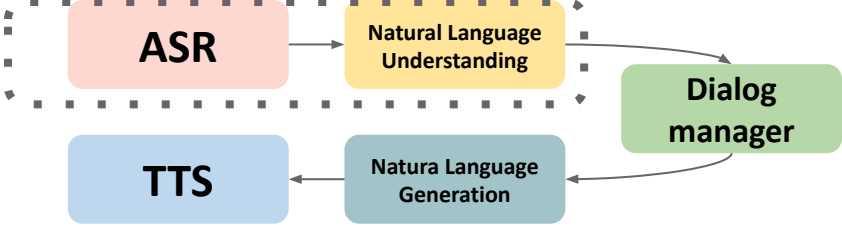
- Revisiting previous challenges (2017 and 2019) with different evaluation metrics.

ZeroSpeech Challenge 2021



- **Task target:** Focus on a **spoken language modeling** tasks
- **Evaluation metric**: phonetic, lexical, syntactic, and semantic metrics. Budget constraint
- **Data**: English (Librispeech and Librilight)
- **Participants:** 11 teams
- **Technical trends:** Self-supervised learning (CPC, APC, Hubert), pseudo-text LM, deep cluster,

ZeroSpeech Challenge 2021



- **Task target:** Focus on a **spoken language modeling** tasks
- **Evaluation metric**: phonetic, lexical, syntactic, and semantic metrics. Budget constraint
- **Data**: English (Librispeech and Librilight)
- **Participants:** 11 teams
- **Technical trends:** Self-supervised learning (CPC, APC, Hubert), pseudo-text LM, deep cluster,

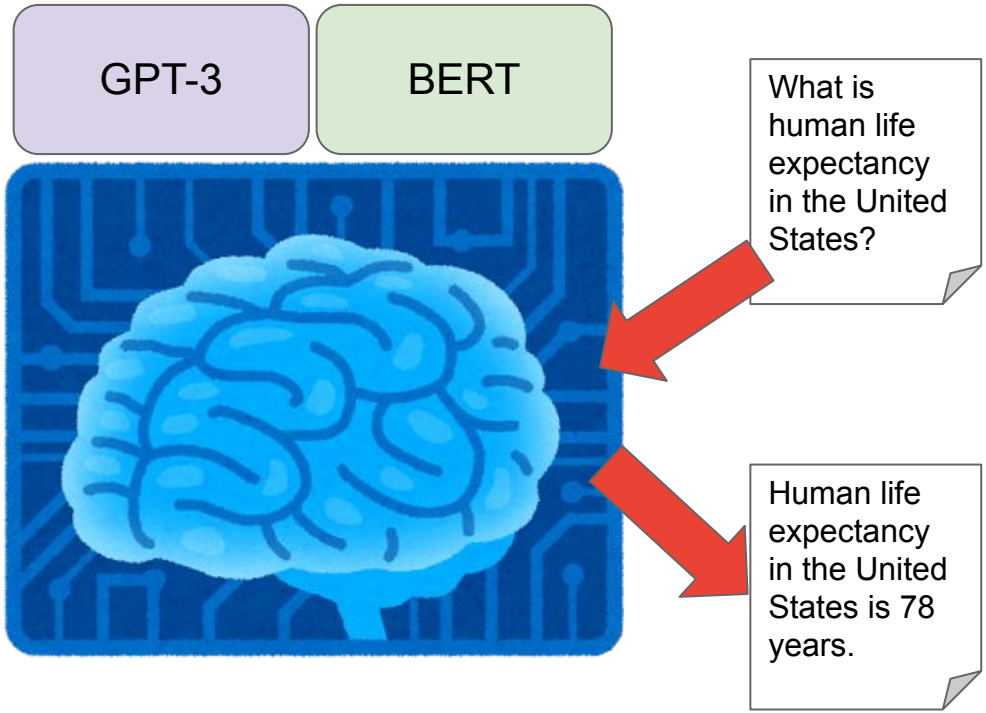
ZeroSpeech challenges has been significantly contributing to the progress of self-supervised learning representation

From Representation Learning to Zero Resources

Shinji Watanabe

1. Unsupervised Speech Recognition
2. ASR-TTS Technique
3. Zero Resource Speech Technologies and Challenges
4. Textless NLP

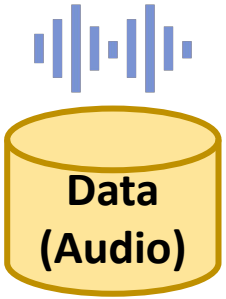
Pre-trained LM in NLP



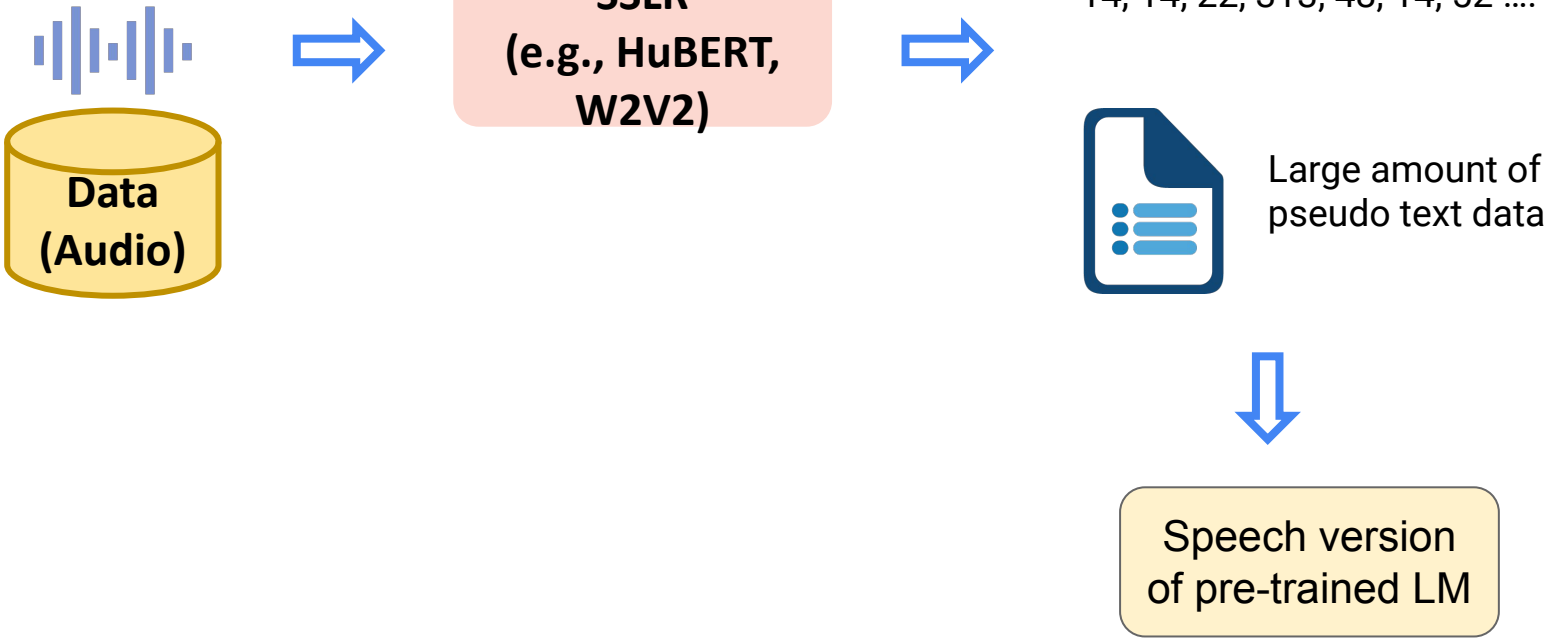
- Great success in various NLP tasks, including
  - Natural language understanding, Question-Answering, Summarization, Conversation generation
- Autoregressive models to ***predict*** future words

Example from https://beta.openai.com/examples/default-qa

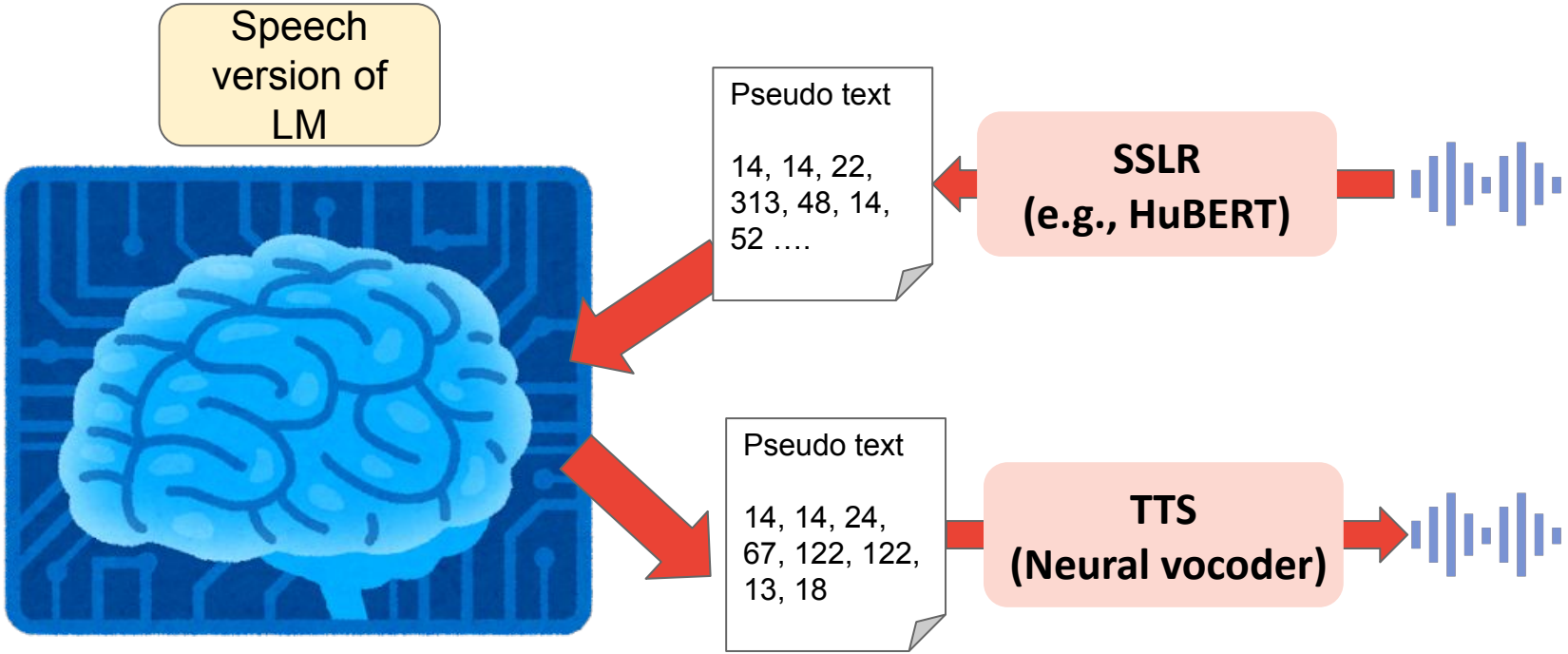
Perform NLP tasks without text



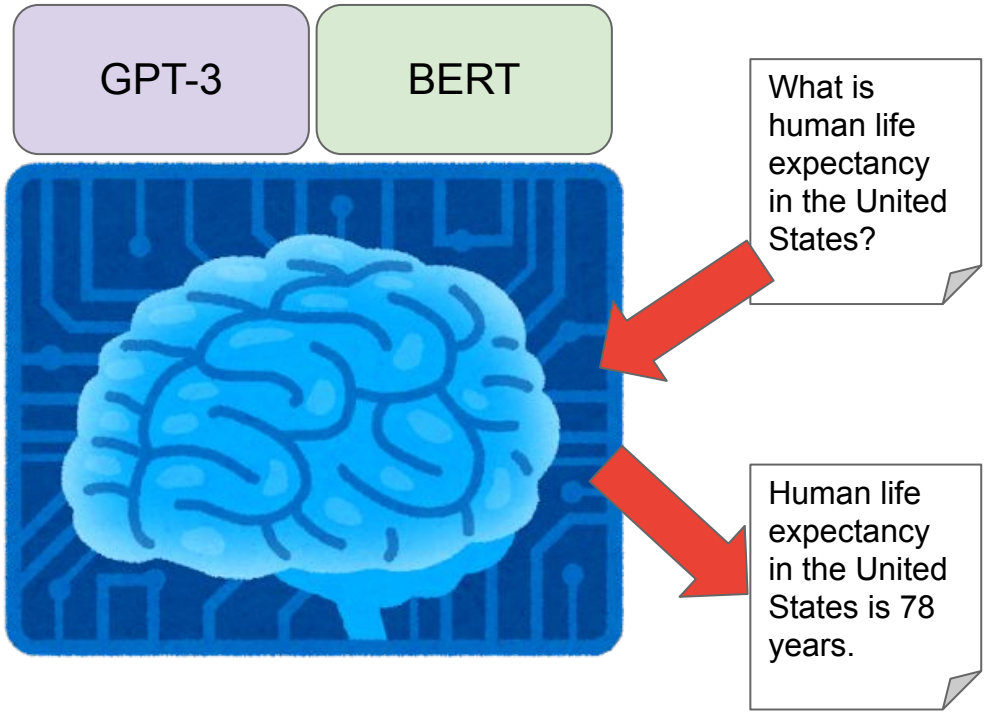
Perform NLP tasks without text



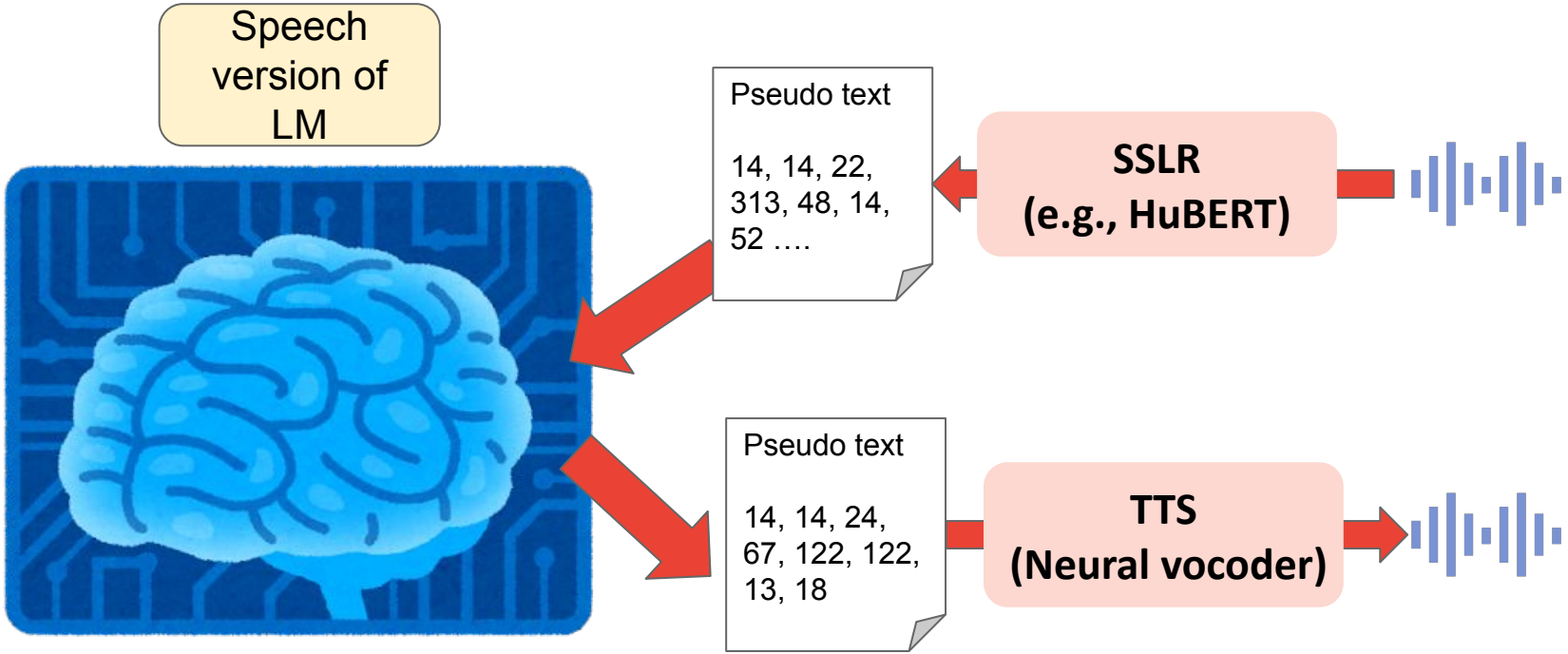
Generative Spoken Language Model (GSLM) [Lakhotia+(2021)]



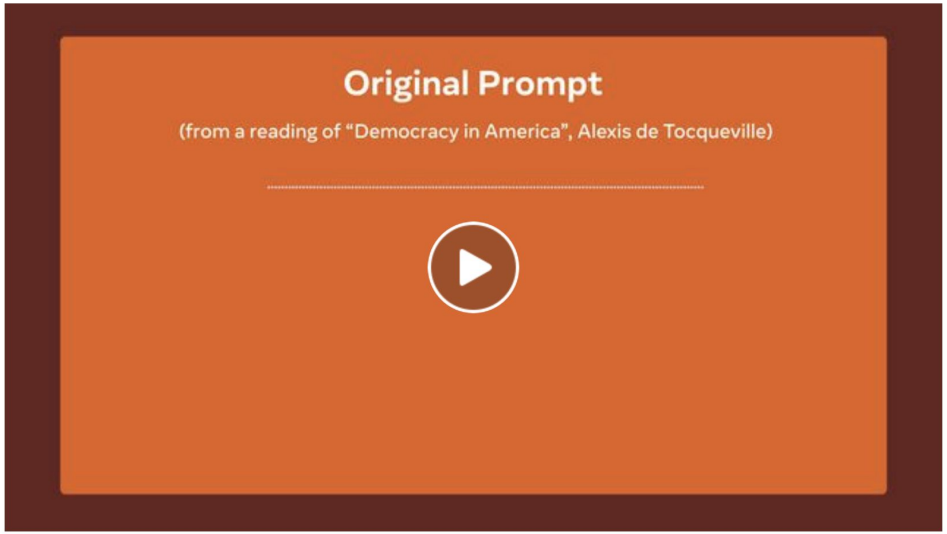
Pre-trained LM in NLP



Generative Spoken Language Model (GSLM) [Lakhotia+(2021)]



Examples



The model is extended to generate spoken dialogues [Nguyen+(2022)] with prosody and emotion controls [Kharitonov+(2021) etc.]

from https://ai.facebook.com/blog/textless-nlp-generating-expressive-speech-from-raw-audio/

Summary of this section

- This section reviews the unsupervised speech processing applications
  - Unsupervised speech recognition
  - ASR-TTS framework
  - Various zero resource speech technologies
  - Textless NLP
- The performance of these applications are boosted by SSLRs
  - Zerospeech challenges provide the milestone of SSLRs techniques

The progress of SSLRs and unsupervised/zero resource techniques are closely related

References

[Da-Rong Liu, et al., 2018] Da-Rong Liu, Kuan-Yu Chen, Hung-Yi Lee, Lin-shan Lee, Completely Unsupervised Phoneme Recognition by Adversarially Learning Mapping Relationships from Audio Embeddings, Interspeech, 2018

[Kuan-yu Chen, et al., 2019] Kuan-yu Chen, Che-ping Tsai, Da-Rong Liu, Hung-yi Lee and Lin-shan Lee, Completely Unsupervised Phoneme Recognition By A Generative Adversarial Network Harmonized With Iteratively Refined Hidden Markov Models, Interspeech, 2019

[Chih-Kuan Yeh, et al., 2019] Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, Dong Yu, Unsupervised Speech Recognition via Segmental Empirical Output Distribution Matching, ICLR, 2019

[Baevski, et al., 2021] Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, Michael Auli, Unsupervised Speech Recognition, NeurIPS, 2021

Topics beyond Accuracy

1. How to use SSL models
2. Security Issues
3. Data Bias
4. Compressing SSL Model

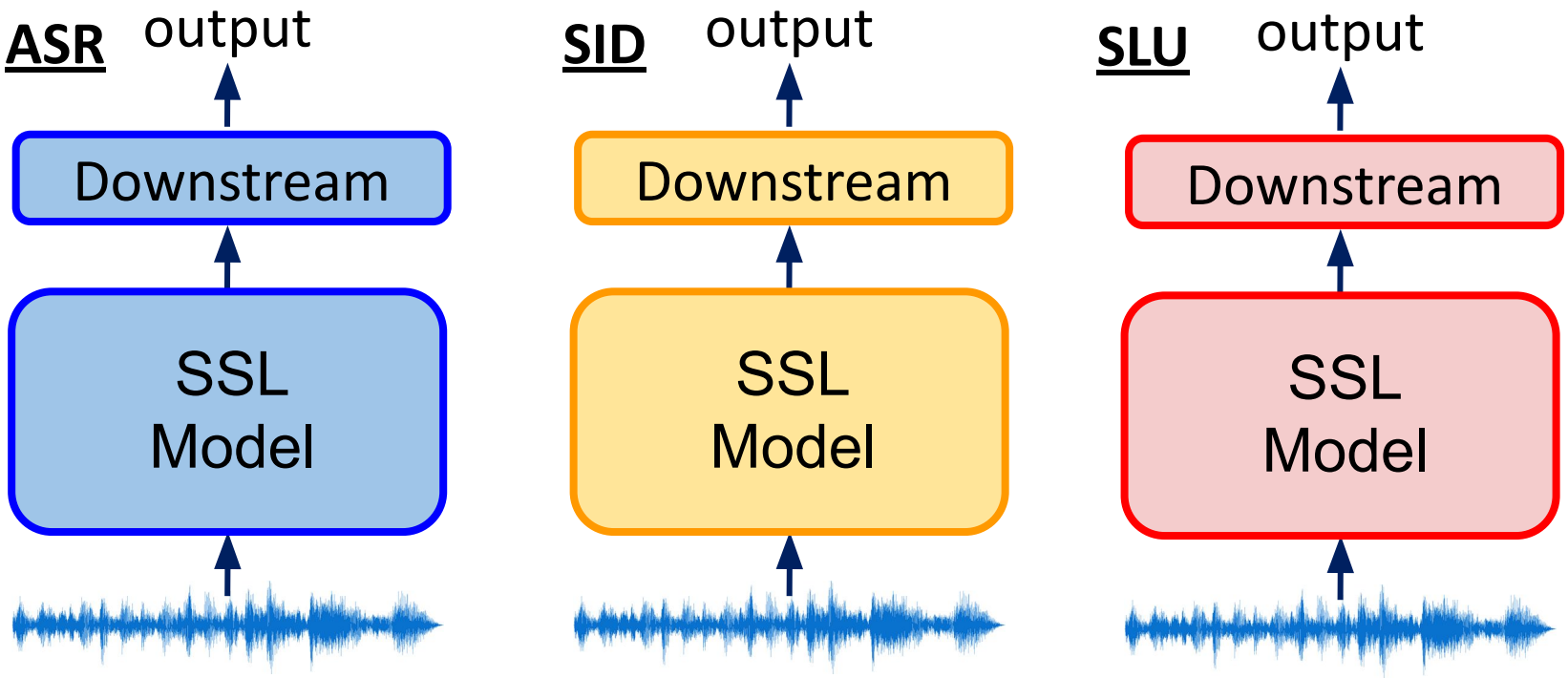
Hung-yi Lee

Topics beyond Accuracy

1. How to use SSL models
2. Security Issues
3. Data Bias
4. Compressing SSL Model

Hung-yi Lee

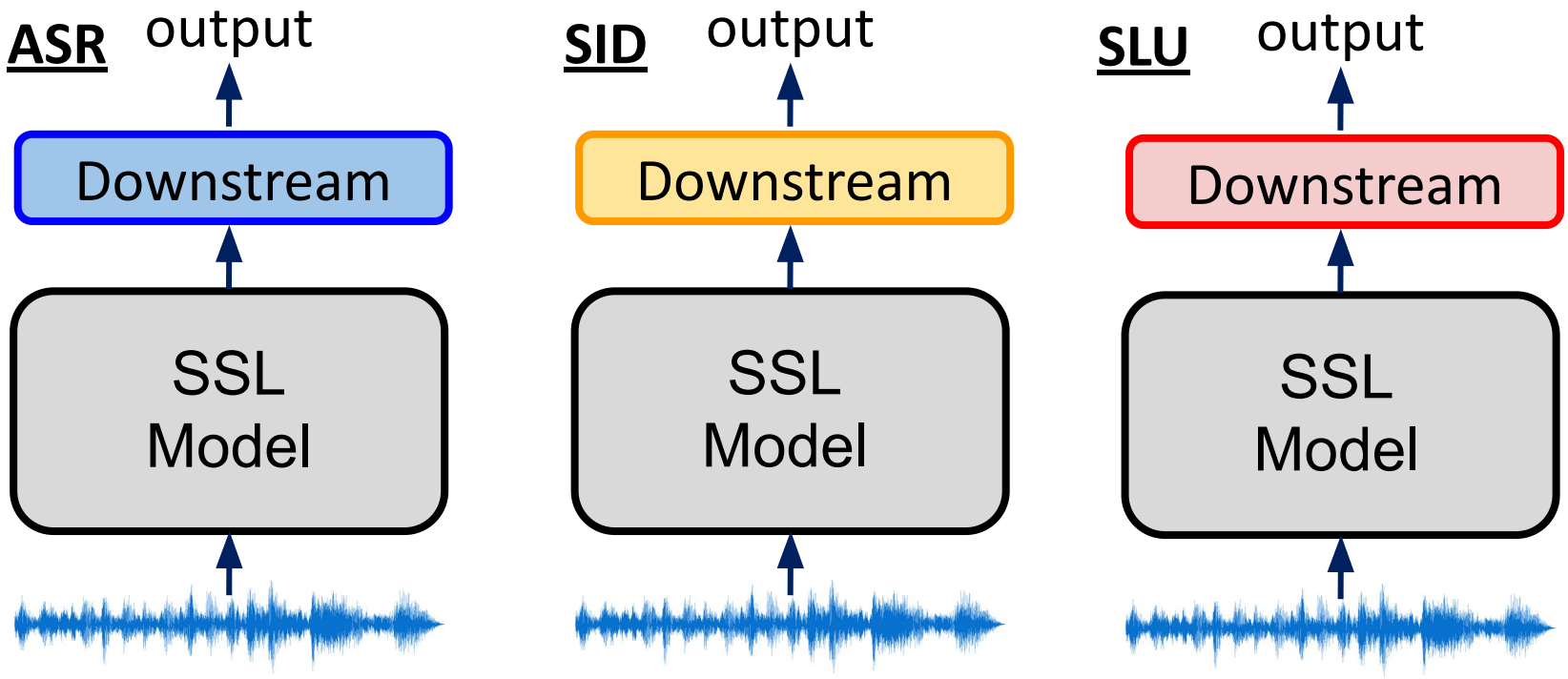
How to use SSL models - Fine-tuning



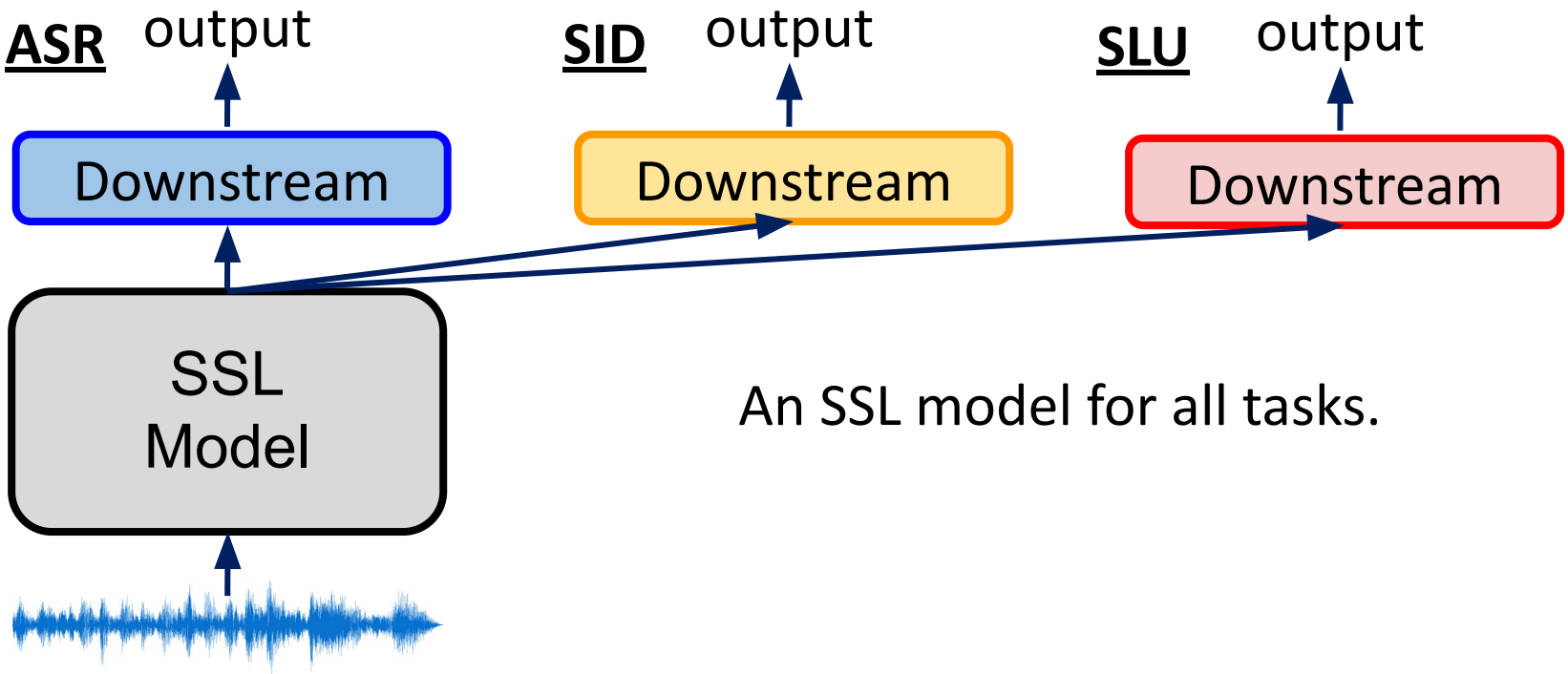
Weakness: Have to store a gigantic SSL model for each task.

(also called Head Fin-tuning)

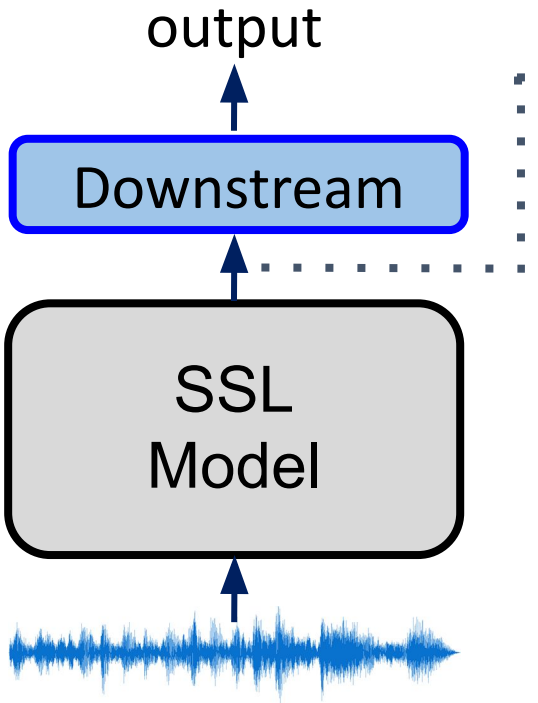
How to use SSL models - Feature Extractor



How to use SSL models - Feature Extractor



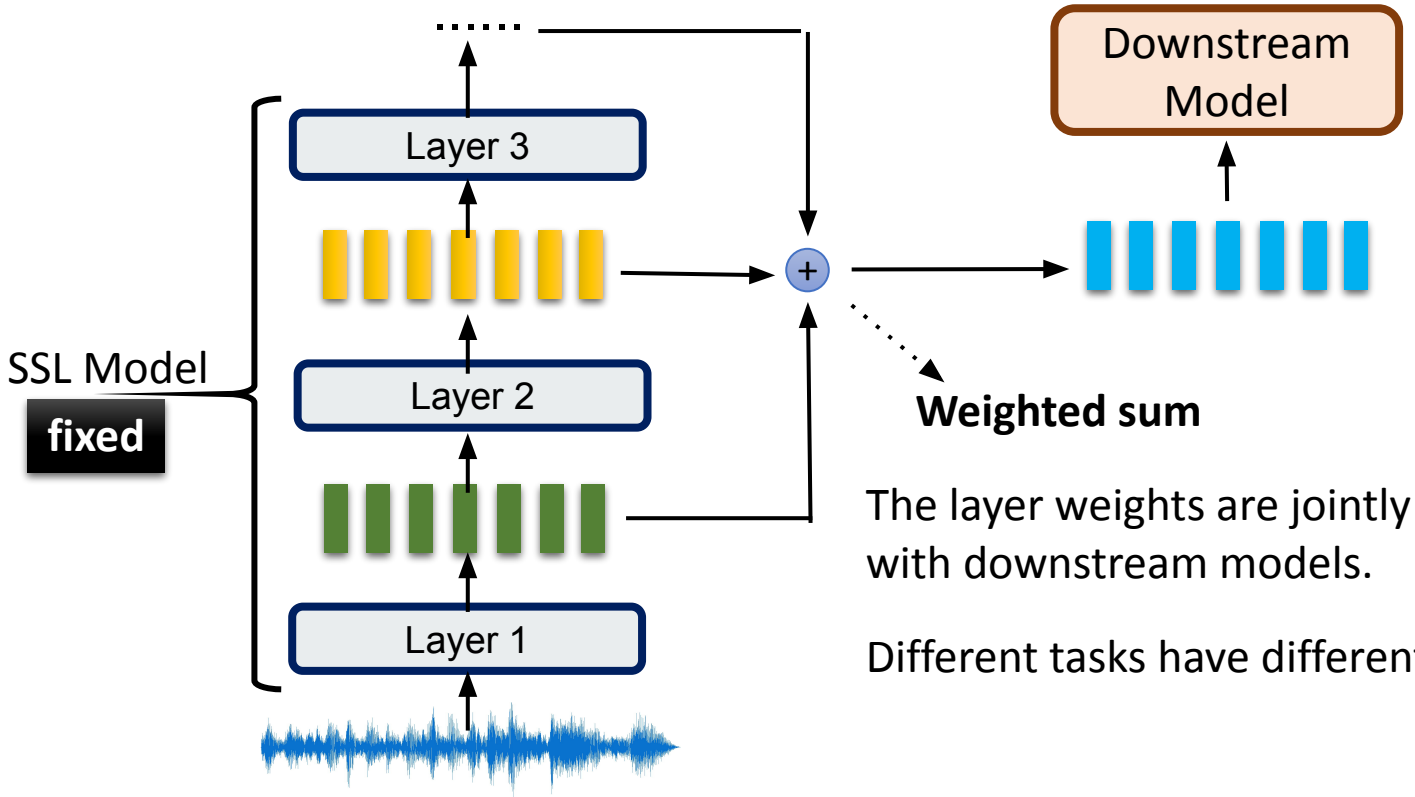
How to use SSL models - Feature Extractor



Last Layer of SSL model

- Do not lead to decent performance on the SUPERB benchmark
  - ➢ Large (>300M parameters) models perform poorly.
  - ➢ Most SSL models do not work well on speaker verification.
- Different layers contain different information

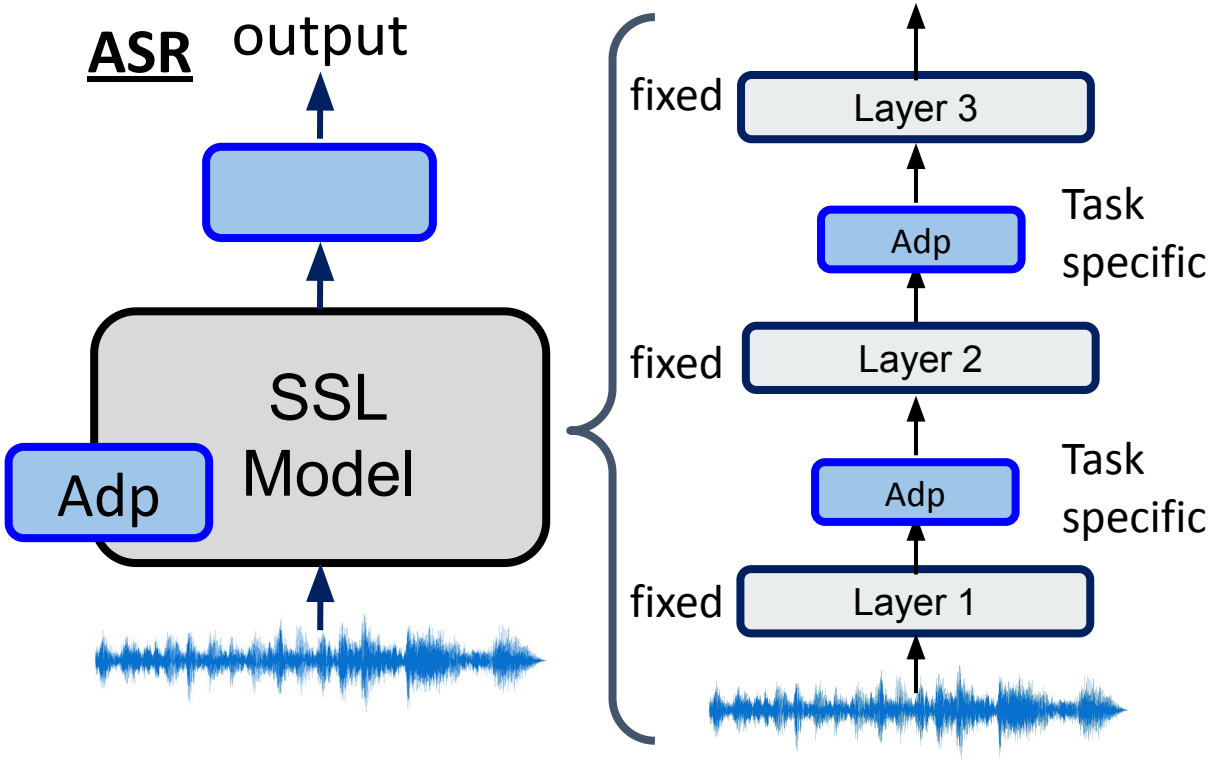
How to use SSL models - Feature Extractor



The layer weights are jointly learned with downstream models.

Different tasks have different weights.

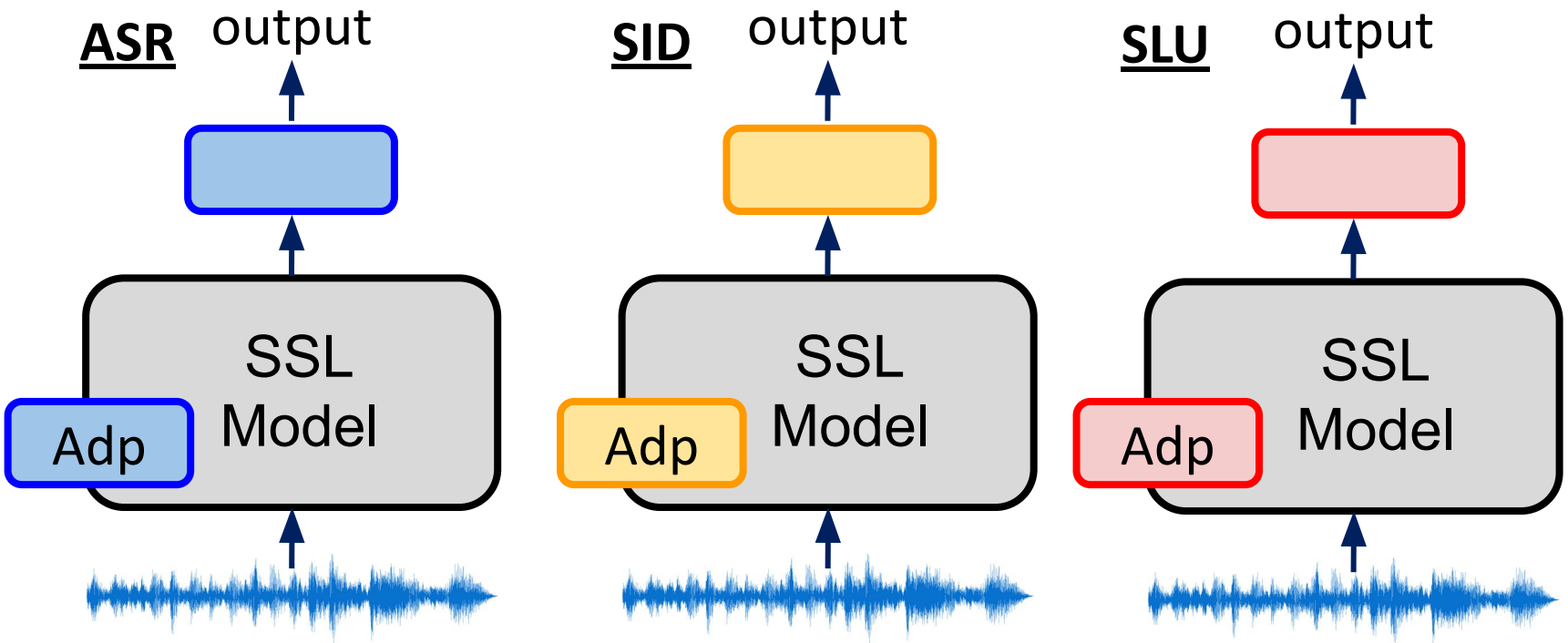
How to use SSL models - Adapter



Adapter has been widely studied in NLP.

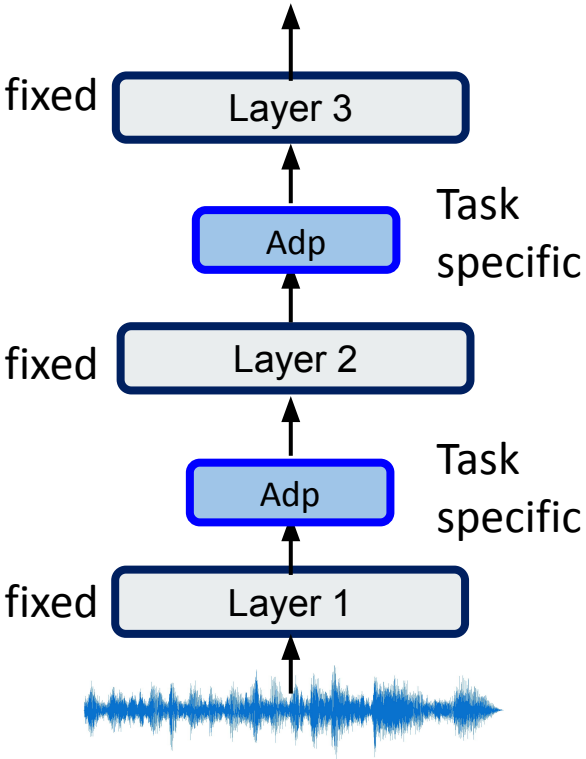
https://adapterhub.ml/

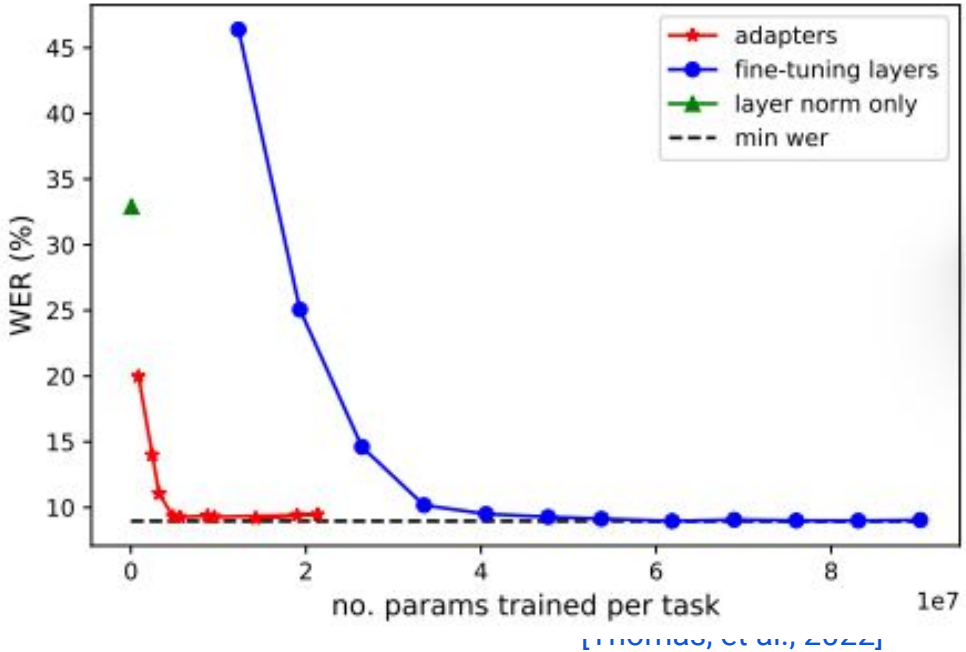
How to use SSL models - Adapter



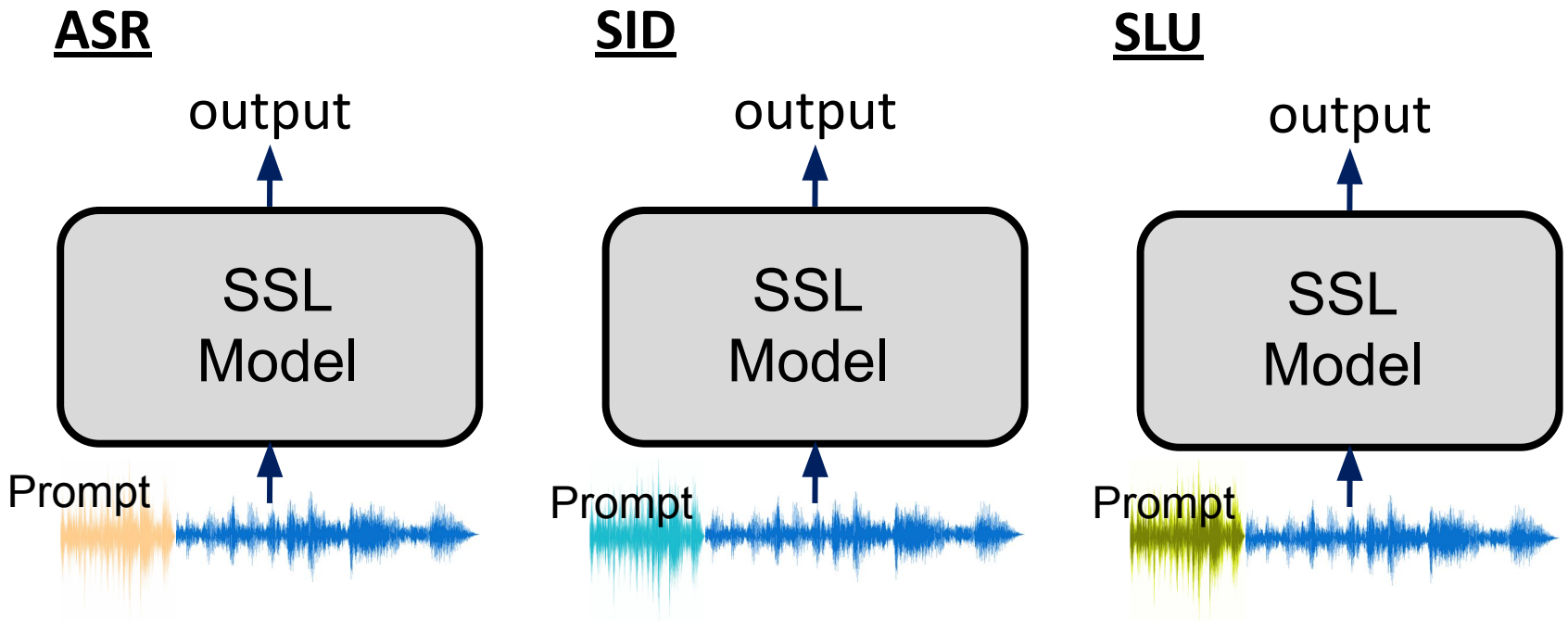
Instead the whole SSL model, only store an Adapter for each task

How to use SSL models - Adapter





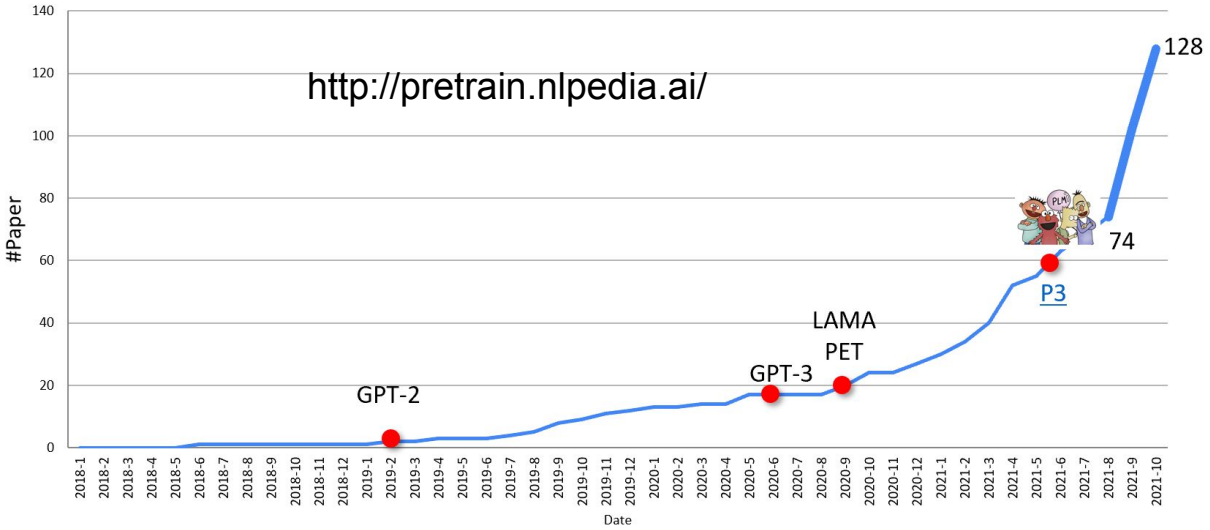
How to use SSL models - Prompt / Reprogramming



The SSL model will solve different tasks by adding extra input “Prompt”.

How to use SSL models - Prompt / Reprogramming

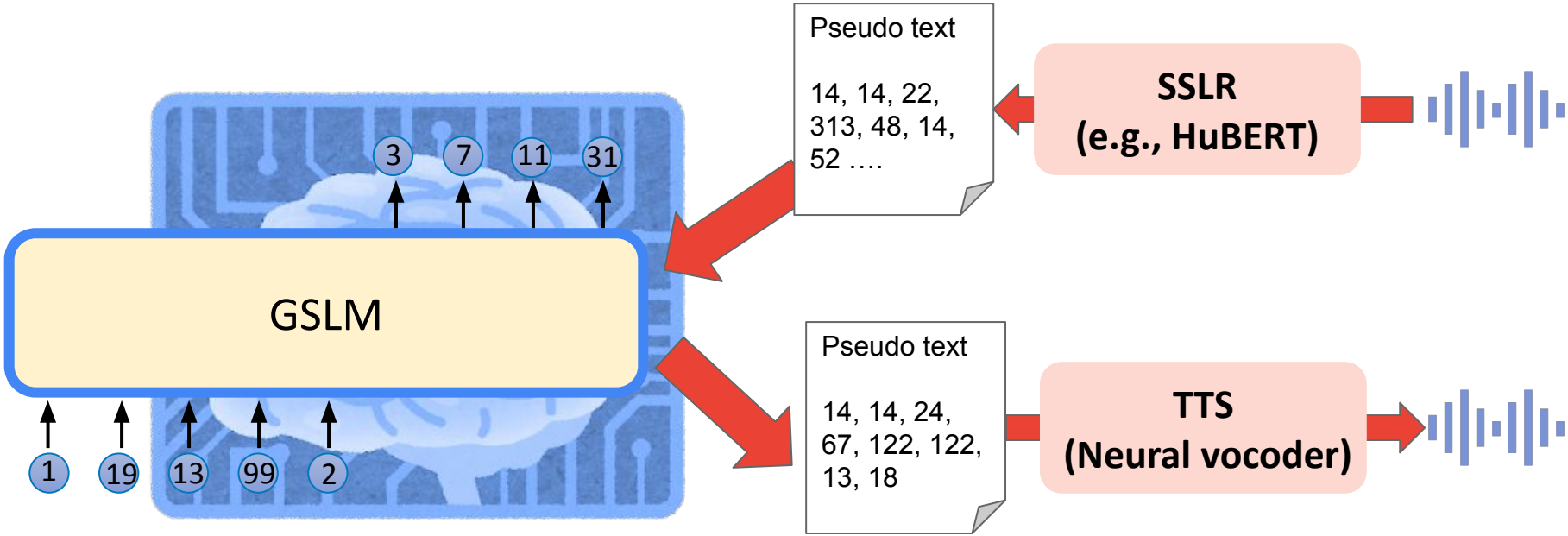
- NLP



- Speech
  - Reprogram voice comment models for time series classification [Chao-Han Huck Yang, et al., 2021]
  - Reprogram voice comment models for different languages [Hao Yen, et al., 2022]

Attempt to prompt GSLM [Kai-Wei Chang, et al., 2022]

Review: Generative Spoken Language Model (GSLM)

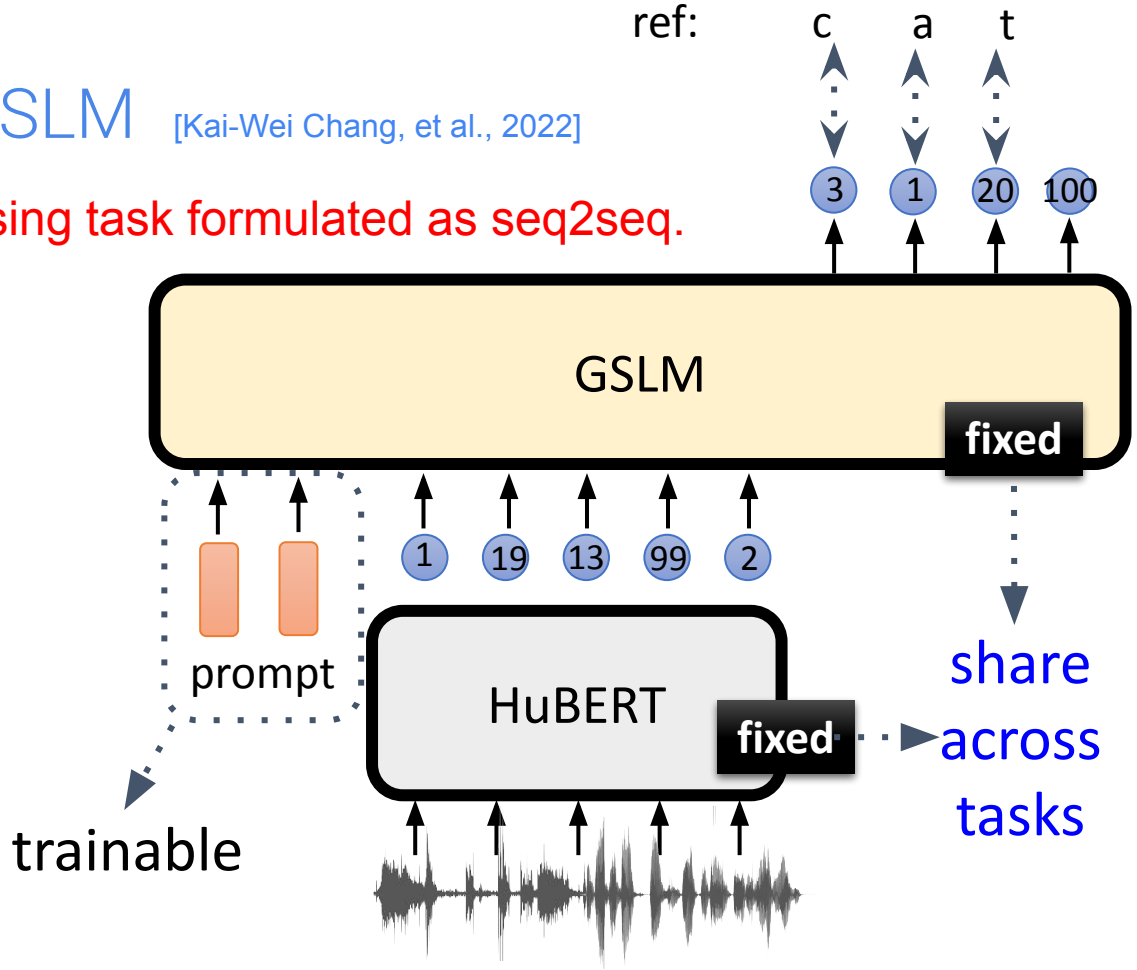


Attempt to prompt GSLM [Kai-Wei Chang, et al., 2022]

Apply on any speech processing task formulated as seq2seq.

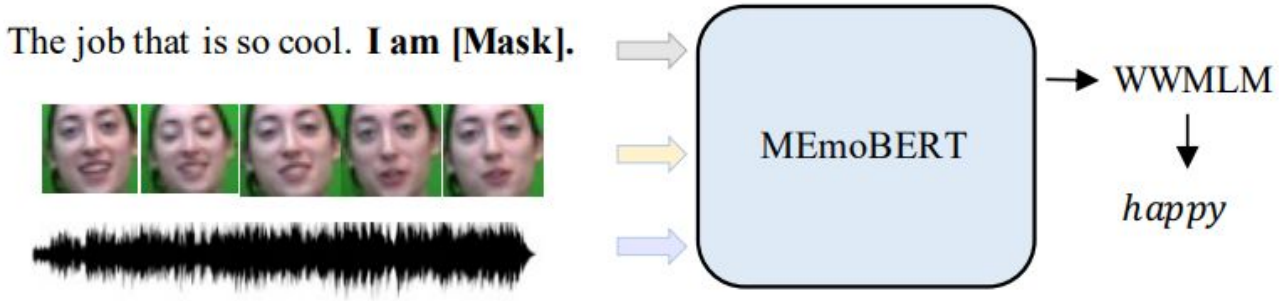
ASR

Char	Unit ID
a	1
b	2
c	3
d	4
…	…
<EOS>	100

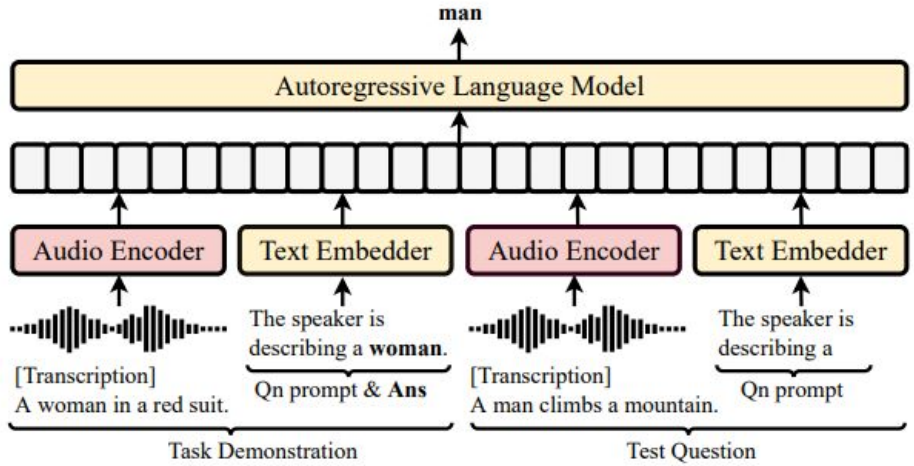


Prompt / Reprogramming SSL models for SLU

- MEmoBERT
  [Jinming Zhao,, et al., 2022]



- WAVPROMPT
  [Heting Gao, et al., 2022]



References

[Kai-Wei Chang, et al., 2022] Kai-Wei Chang, Wei-Cheng Tseng, Shang-Wen Li, Hung-yi Lee, An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks, arXiv, 2022

[Heting Gao, et al., 2022] Heting Gao, Junrui Ni, Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, WAVPROMPT: Towards Few-Shot Spoken Language Understanding with Frozen Language Models, arXiv, 2022

[Lakhotia, et al., 2021] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, Emmanuel Dupoux, On Generative Spoken Language Modeling from Raw Audio, TACL, 2021

[Thomas, et al., 2022] Bethan Thomas, Samuel Kessler, Salah Karout, Efficient Adapter Transfer of Self-Supervised Speech Models for Automatic Speech Recognition, ICASSP, 2022

[Chao-Han Huck Yang, et al., 2021] Chao-Han Huck Yang, Yun-Yun Tsai, Pin-Yu Chen, Voice2Series: Reprogramming Acoustic Models for Time Series Classification, ICML, 2021

[Hao Yen, et al., 2022] Hao Yen, Pin-Jui Ku, Chao-Han Huck Yang, Hu Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, Yu Tsao, A Study of Low-Resource Speech Commands Recognition based on Adversarial Reprogramming, Interspeech, 2022

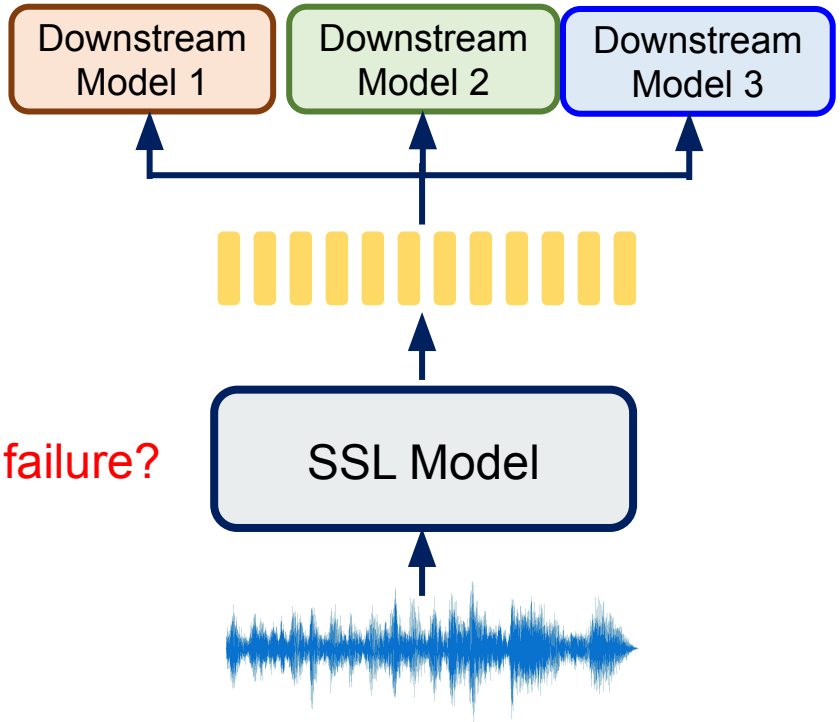
[Jinming Zhao,, et al., 2022] Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, Haizhou Li, MEmoBERT: Pre-training Model with Prompt-based Learning for Multimodal Emotion Recognition, ICASSP, 2022

Topics beyond Accuracy

1. How to use SSL models
2. Security Issues
3. Data Bias
4. Compressing SSL Model

Hung-yi Lee

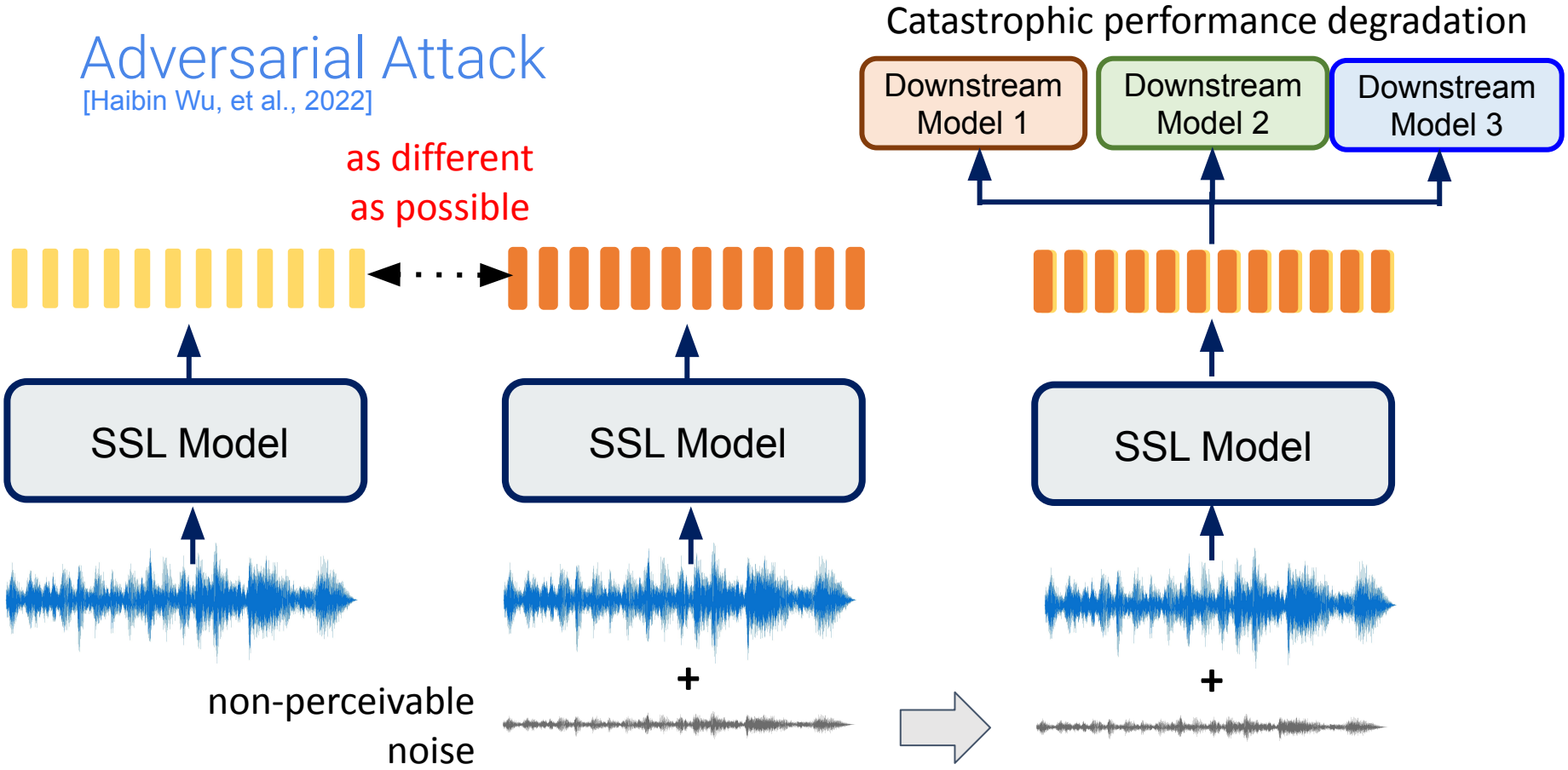
Single Point of Failure



single point of failure?

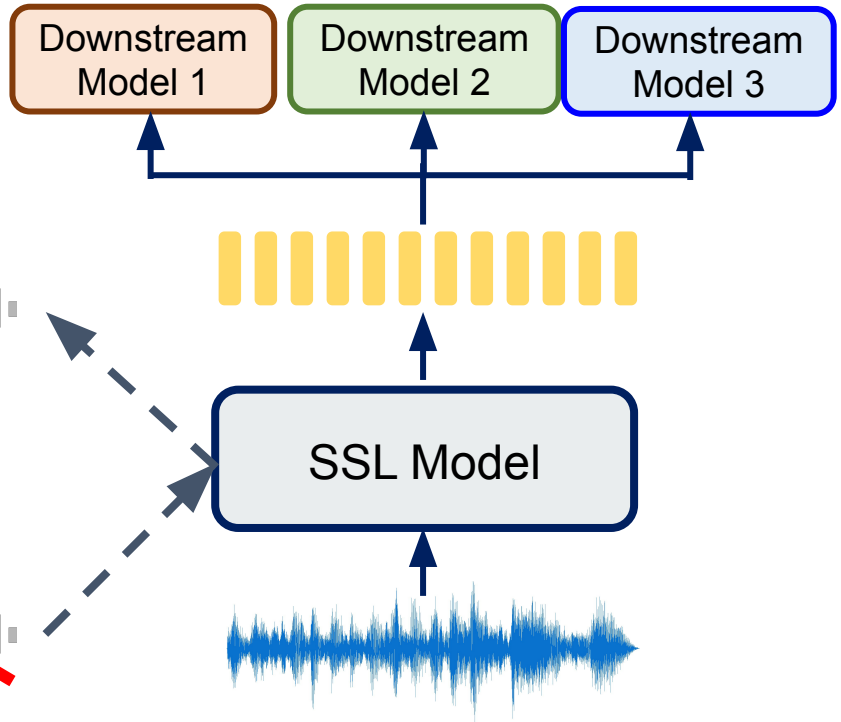
Adversarial Attack

[Haibin Wu, et al., 2022]



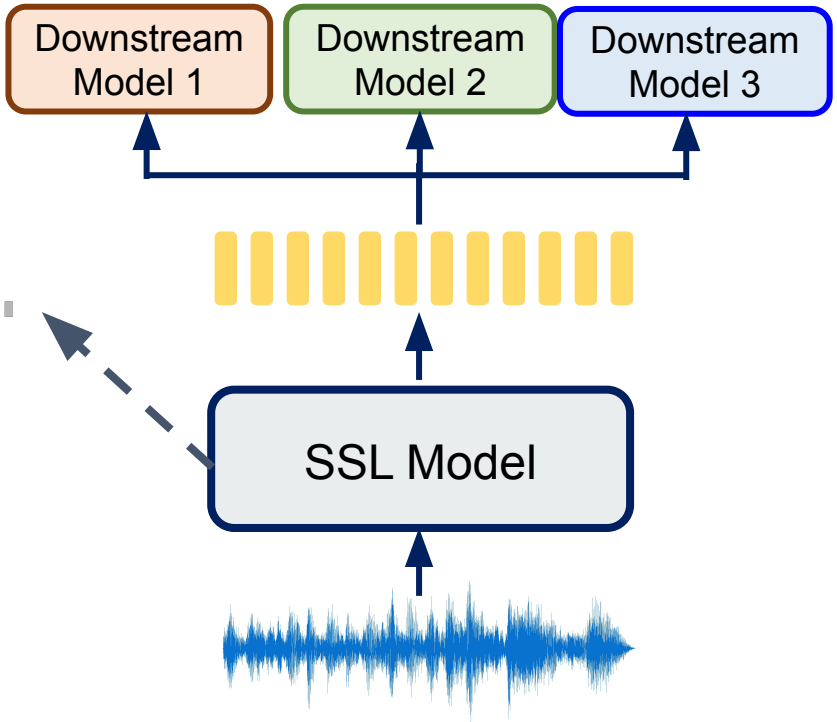
Privacy Attack

Recover? No successful results until now



Collected from real applications

Privacy Attack



Recover? No successful results until now

Membership Inference Attack

An utterance is in training set or not?

 Yes  No

High accuracy with naive approaches [Wei-Cheng Tseng, et al., 2022]

References

[Wei-Cheng Tseng, et al., 2022] Wei-Cheng Tseng, Wei-Tsung Kao, Hung-yi Lee, Membership Inference Attacks Against Self-supervised Speech Models, AAAI SAS workshop, 2022

[Haibin Wu, et al., 2022] Haibin Wu, Bo Zheng, Xu Li, Xixin Wu, Hung-yi Lee, Helen Meng, Characterizing the adversarial vulnerability of speech self-supervised learning, ICASSP, 2022

Topics beyond Accuracy

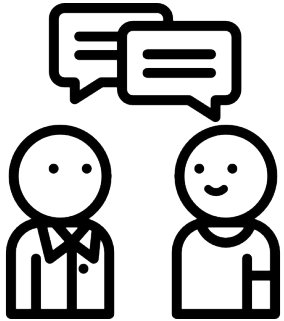
1. How to use SSL models
2. Security Issues
3. Data Bias
4. Compressing SSL Model

Hung-yi Lee

Data Bias

Demographic: gender, age, accent, …

Content: topic, word use, …



Prosody: speech rate, tone, …

Does data bias influence SSL models' performance on downstream tasks?

Data Bias

- **Speaking rate**
  - SSL models pre-trained on speech with faster speaking rate obtain worse performance. Lower speaking rate does not degrade performance. [Yen Meng, et al., 2022]
- **Gender**
  - Fine-tuning: gender-specific model leads to overall worse performance [Boito, et al., 2022]
  - The conclusion is mixed when using SSL model as feature extractor. [Yen Meng, et al., 2022][Boito, et al., 2022]
- **Language**
  - SSL models show a very small native language effect (except for CPC) [Millet, et al., 2022]
  - A monolingual wav2vec-2.0 is a good few-shot ASR learner in several languages. [Khurana, et al., 2022]

References

[Boito, et al., 2022] Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, Yannick Estève, A Study of Gender Impact in Self-supervised Models for Speech-to-Text Systems, arXiv, 2022

[Khurana, et al., 2022] Sameer Khurana, Antoine Laurent, James Glass, Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0, ICASSP, 2022

[Yen Meng, et al., 2022] Yen Meng, Yi-Hui Chou, Andy T. Liu, Hung-yi Lee, Don't speak too fast: The impact of data bias on self-supervised speech models, ICASSP, 2022

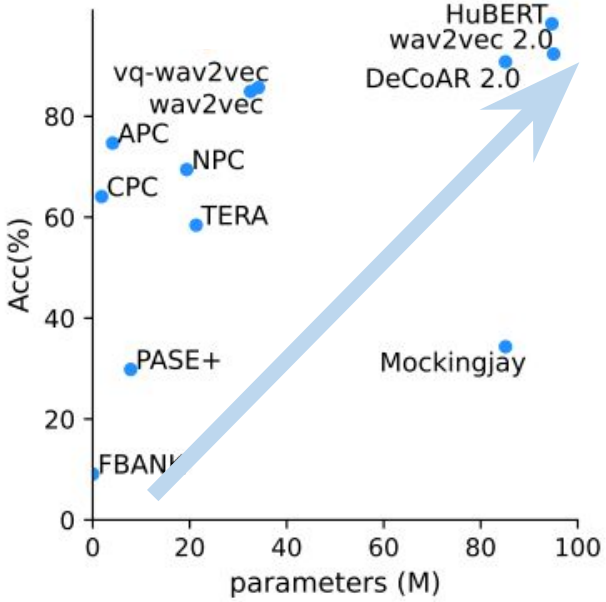
[Millet, et al., 2022] Juliette Millet, Ewan Dunbar, Do self-supervised speech models develop human-like perception biases?, AAAI SAS workshop, 2022

Topics beyond Accuracy

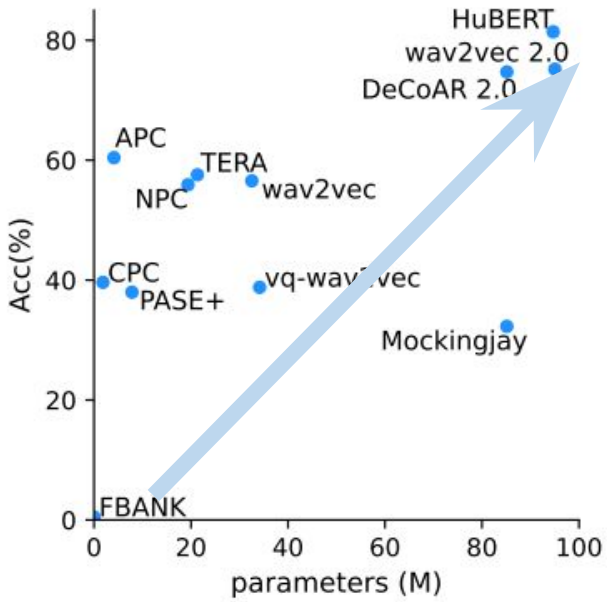
1. How to use SSL models
2. Security Issues
3. Data Bias
4. Compressing SSL Model

Hung-yi Lee

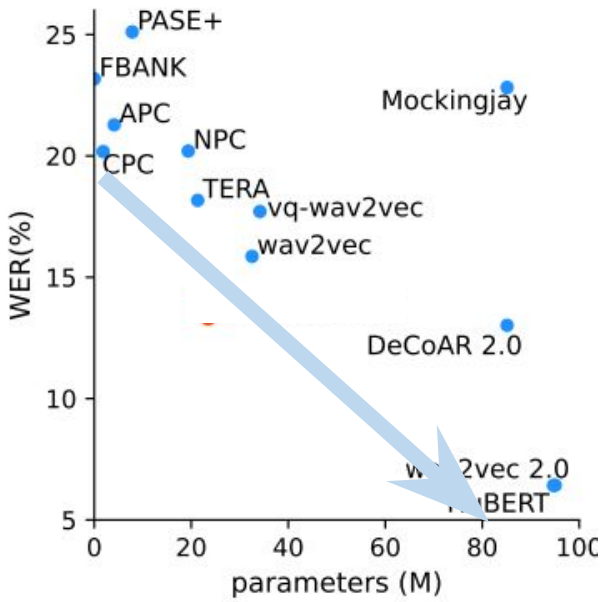
Larger Pre-trained Models lead to Better Results



Intent Classification



Speaker Identification



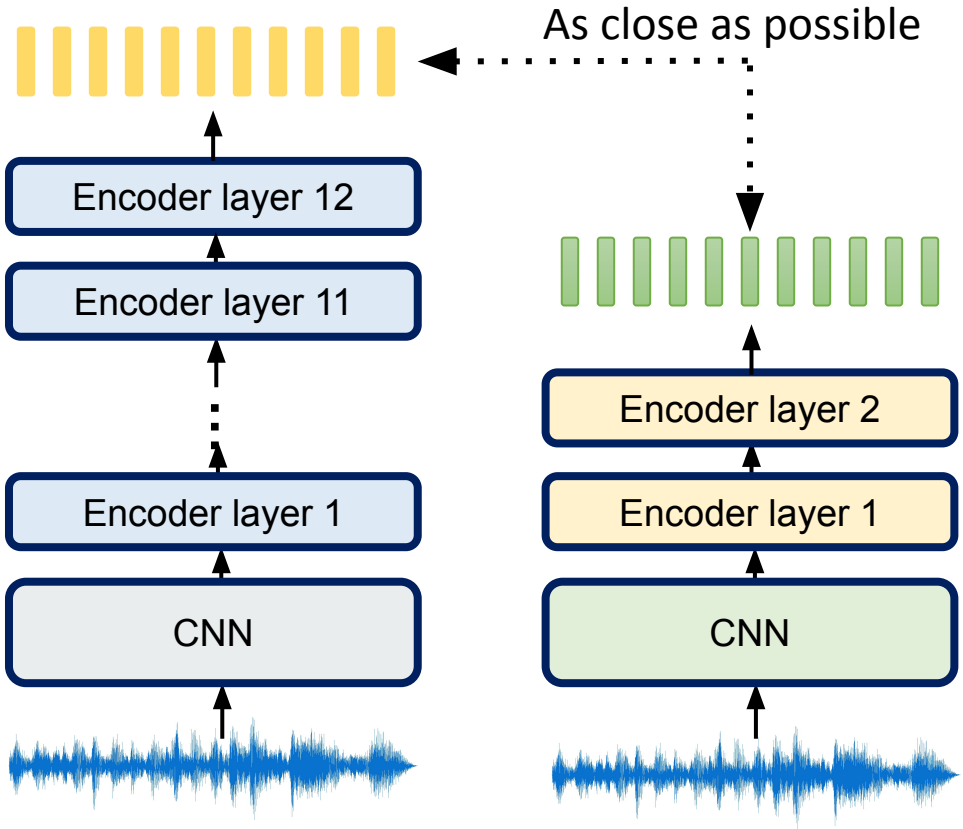
ASR (without LM)

From SUPERB benchmark

[Heng-Jui Chang, et al., 2022]

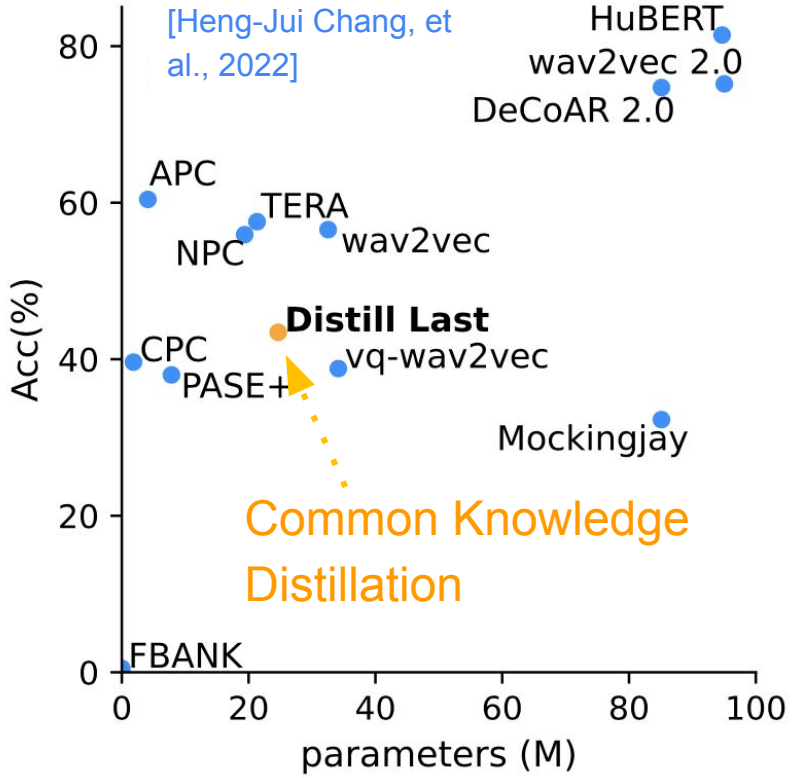
Knowledge Distillation

[Zilun Peng, et al., 2021]



Speaker Identification

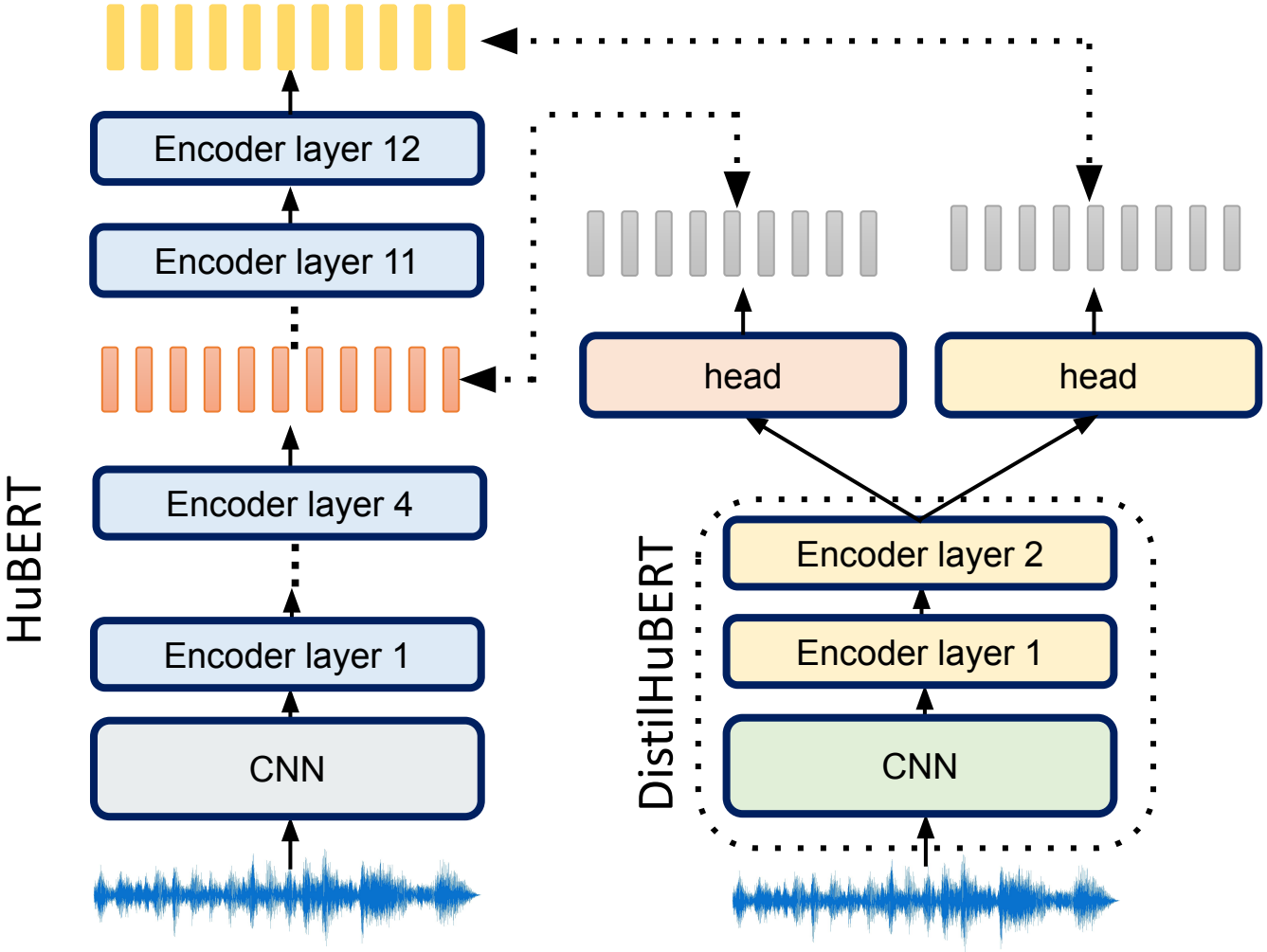
[Heng-Jui Chang, et al., 2022]



Each layer contains different information.

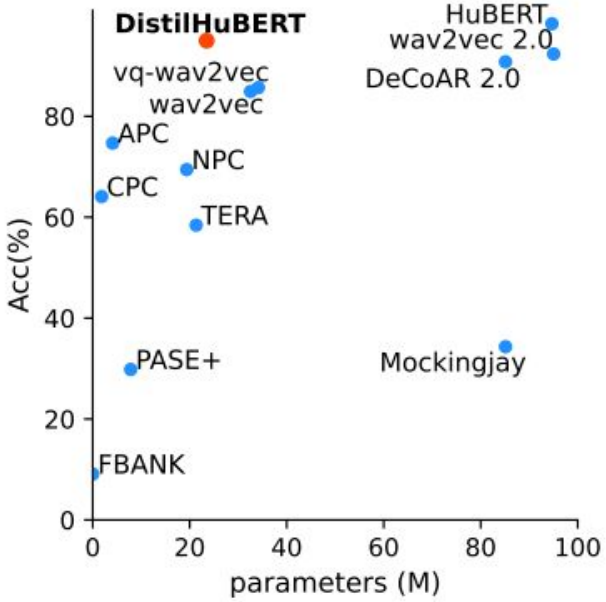
Learning from the last layer is not sufficient.

[Heng-Jui Chang, et al., 2022]

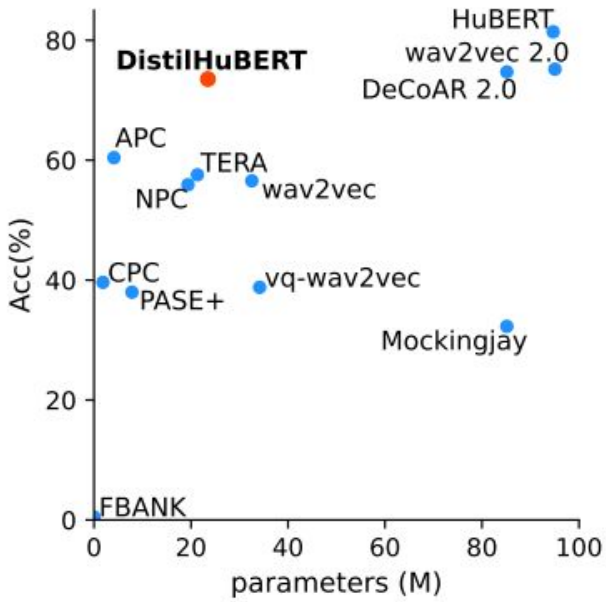


[Heng-Jui Chang, et al., 2022]

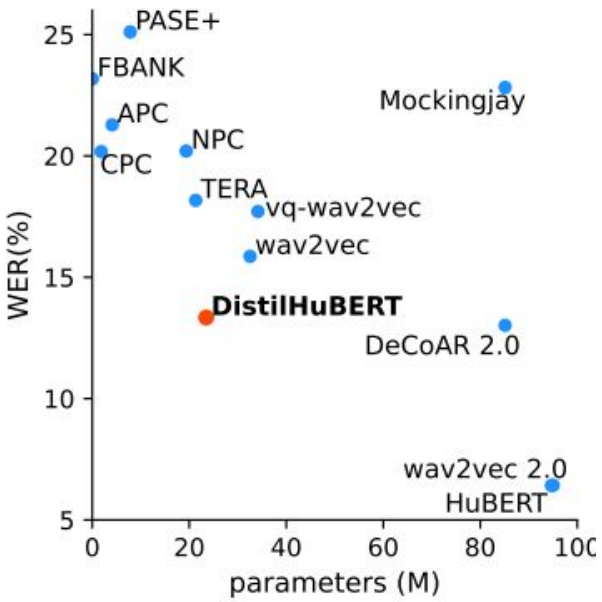
DistilHuBERT



Intent Classification



Speaker Identification

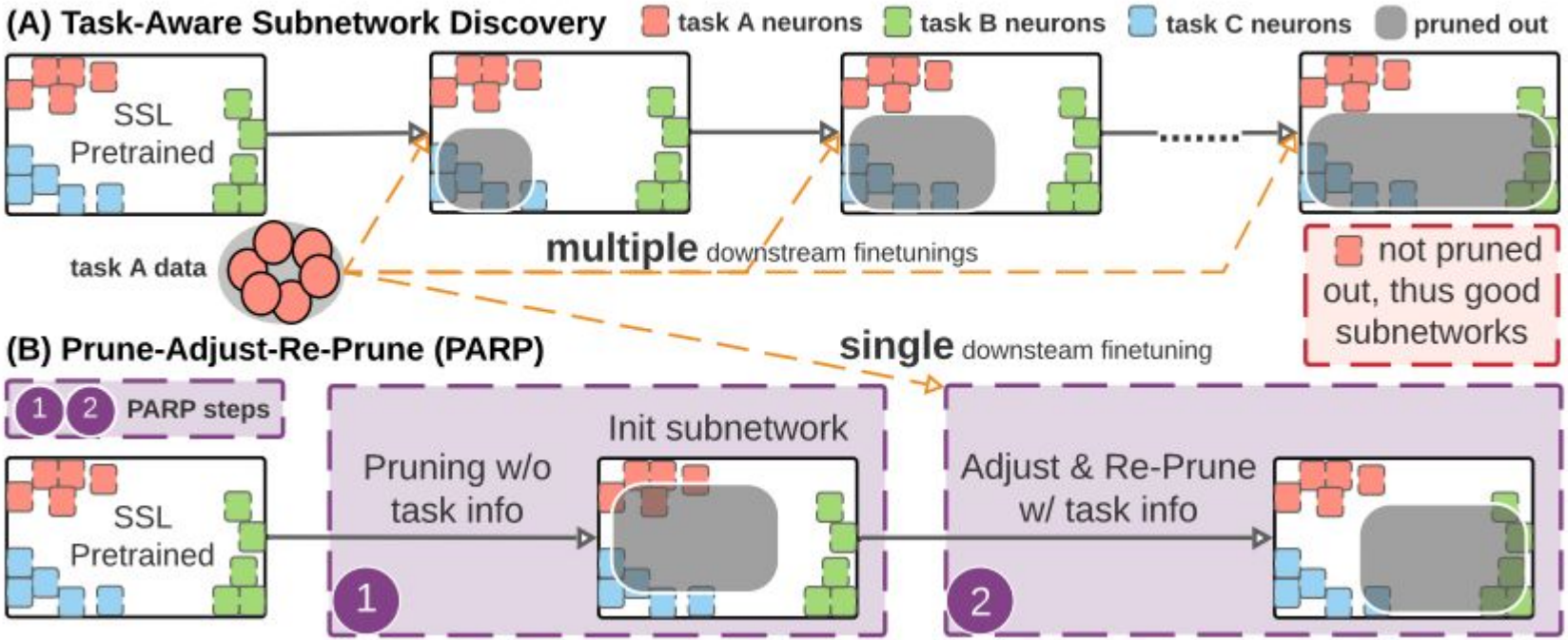


ASR (without LM)

Pruning

[Cheng-I Jeff Lai, et al., 2021]

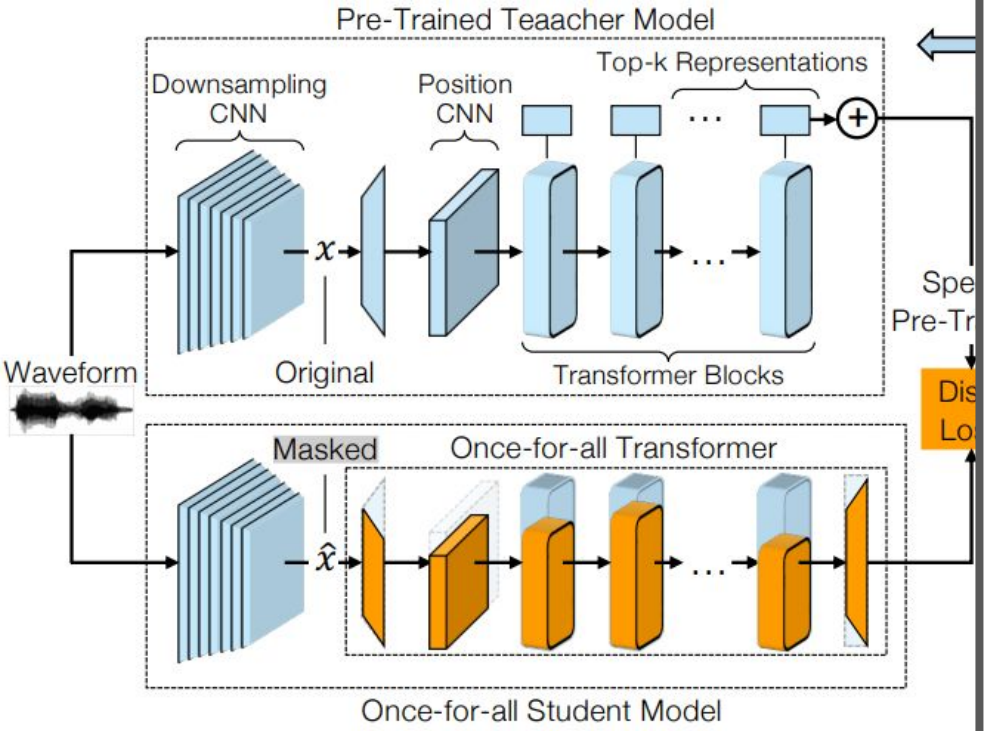
There are subnetworks from wav2vec 2.0 with an absolute 10.9%/12.6% WER decrease compared to the full model.

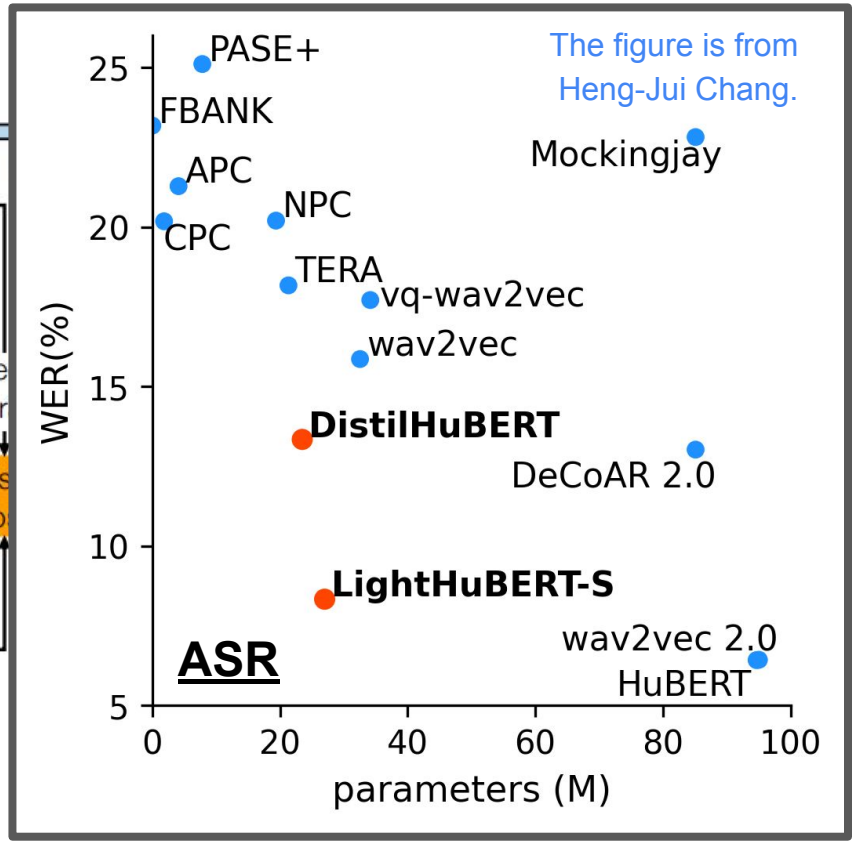


LightHuBERT

[Rui Wang, et al., 2022]

LightHuBERT obtains comparable performance to the HuBERT teacher in most tasks.





References

[Heng-Jui Chang, et al., 2022] Heng-Jui Chang, Shu-wen Yang, Hung-yi Lee, DistilHuBERT: Speech Representation Learning by Layer-wise Distillation of Hidden-unit BERT, ICASSP, 2022

[Cheng-I Jeff Lai, et al., 2021] Cheng-I Jeff Lai, Yang Zhang, Alexander H. Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David Cox, James Glass, PARP: Prune, Adjust and Re-Prune for Self-Supervised Speech Recognition, NeurIPS, 2021

[Zilun Peng, et al., 2021] Zilun Peng, Akshay Budhkar, Ilana Tuil, Jason Levy, Parinaz Sobhani, Raphael Cohen, Jumana Nassour. Shrinking Bigfoot: Reducing wav2vec 2.0 footprint, Interspeech, 2021

[Rui Wang, et al., 2022] Rui Wang, Qibing Bai, Junyi Ao, Long Zhou, Zhixiang Xiong, Zhihua Wei, Yu Zhang, Tom Ko, Haizhou Li, LightHuBERT: Lightweight and Configurable Speech Representation Learning with Once-for-All Hidden-Unit BERT, arXiv, 2022

Toolkits for Self-Supervised Speech Representation Learning

Shu-wen (Leo) Yang

Big picture: toolkits and their key support for SSL

- wav2vec, vq-wav2vec, wav2vec 2.0, data2vec
- HuBERT, wav2vec U, wav2vec U 2.0, GSLM, pGSLM…
  (most of the SOTA in speech SSL)

- The most comprehensive library for pre-trained models
- Standardized benchmark tools for speech SSL: SUPERB

- Pre-train HuBERT Task
- Utilize S3PRL pre-trained models as Frontend in many tasks

Big picture: toolkits and their key support for SSL

LeBenchmark

- Benchmark SSL on French with various tasks

SLUE

- Spoken Language Understanding Evaluation (SLUE) benchmark

textlesslib

- speech → discrete units → continued discrete units → speech

SpeechBrain

- Recipes of finetuning wav2vec2 for ASR
- A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding

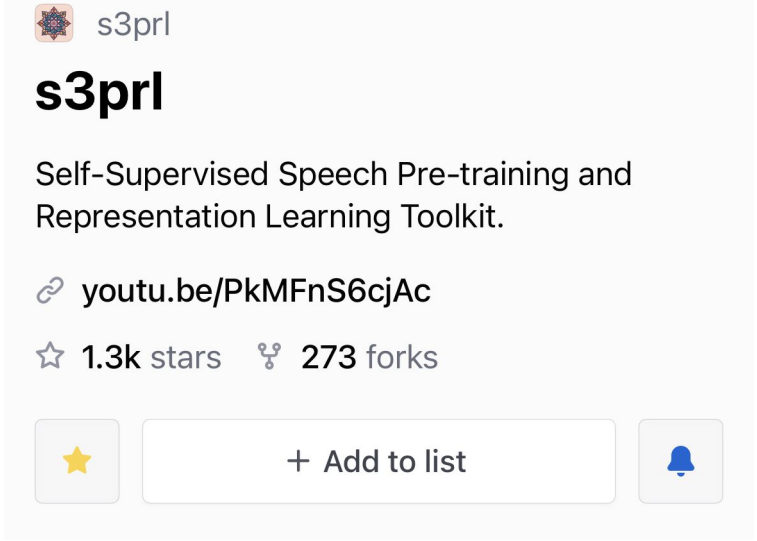
Big picture: toolkits and their key support for SSL

- Study/Propose pre-training methods → **Fairseq**
- Try all kinds of pretrained models for your own work → **S3PRL**
- Benchmark the quality of pretrained models:
  - SUPERB → **S3PRL**
  - SLUE → **slue-toolkit**
  - LeBenchmark → **le-benchmark**
  - A Fine-tuned Wav2vec2.0/HuBERT Benchmark → **SpeechBrain**
- Use SSL models for more & complicated tasks → **ESPnet & SpeechBrain**

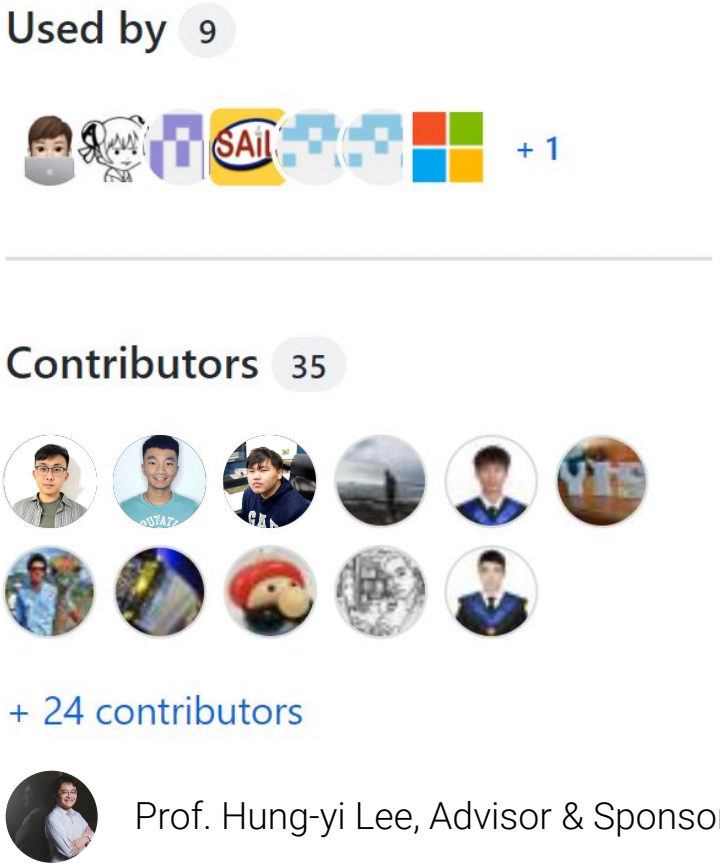
S3PRL

Self-**S**upervised **S**peech **P**re-training and **R**epresentation **L**earning

https://github.com/s3prl/s3prl/



https://github.com/s3prl/s3prl/



Prof. Hung-yi Lee, Advisor & Sponsor

Evolution and major usage of S3PRL

Pre-training

Pre-trained model collection

Downstream Benchmarking & SUPERB

2019


Andy T. Liu


Shu-wen Yang


Po-Han Chi

2020

Shu-wen Yang

Andy T. Liu

2021

Shu-wen Yang

Andy T. Liu

Po-Han Chi

Heng-Jui Chang

Xuankai Chang

Yung-Sung Chuang

Zili Huang

Wen-Chin Huang

Tzu-Hsien Huang

Kushal Lakhotia

Yist Lin Y.

Guan-Ting Lin

Jiatong Shi

Hsiang-Sheng Tsai

Wei-Cheng Tseng

Acknowledgement

Prof. Shinji Watanabe (CMU)

Abdelrahman Mohamed (Meta AI)

Shang-Wen (Daniel) Li (Meta AI)

Pre-training

Pre-training

Mockingjay

TERA

NPC

DistilHubert

AudioAlbert

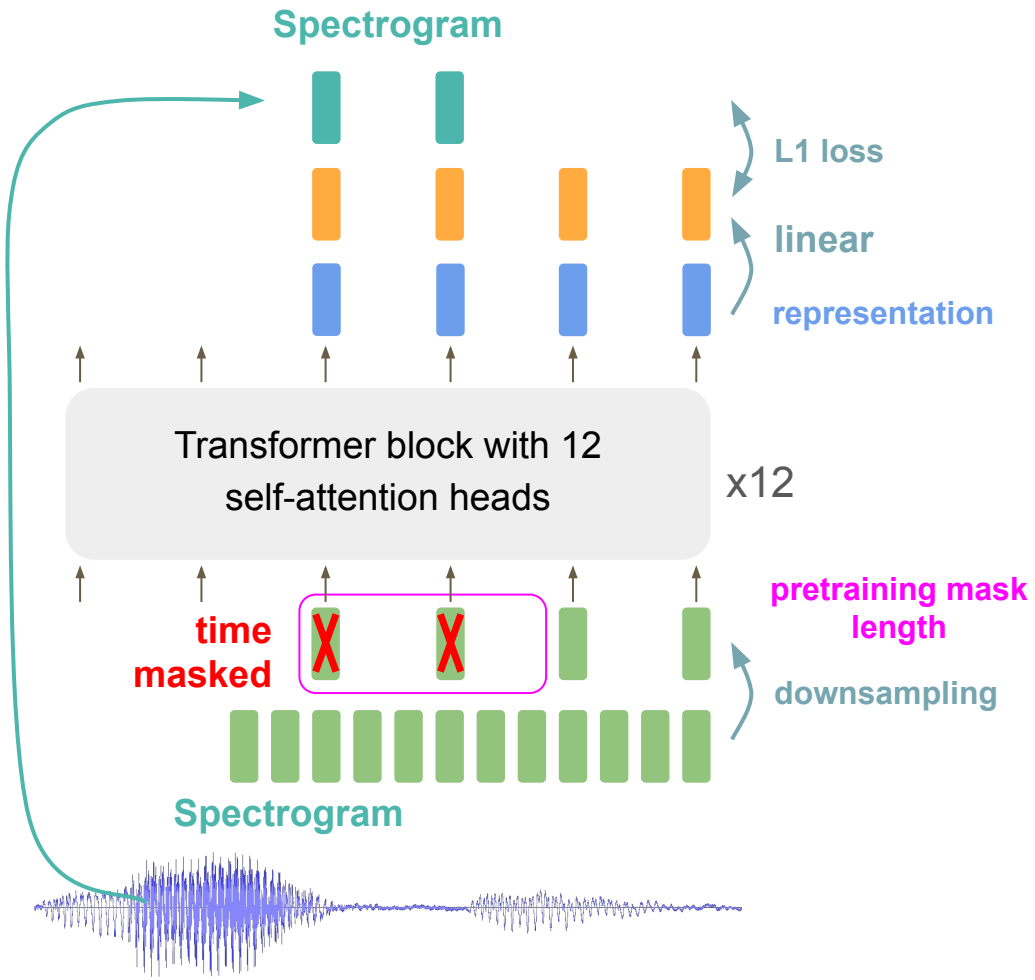
APC

VQ-APC

Pre-training

Pre-training

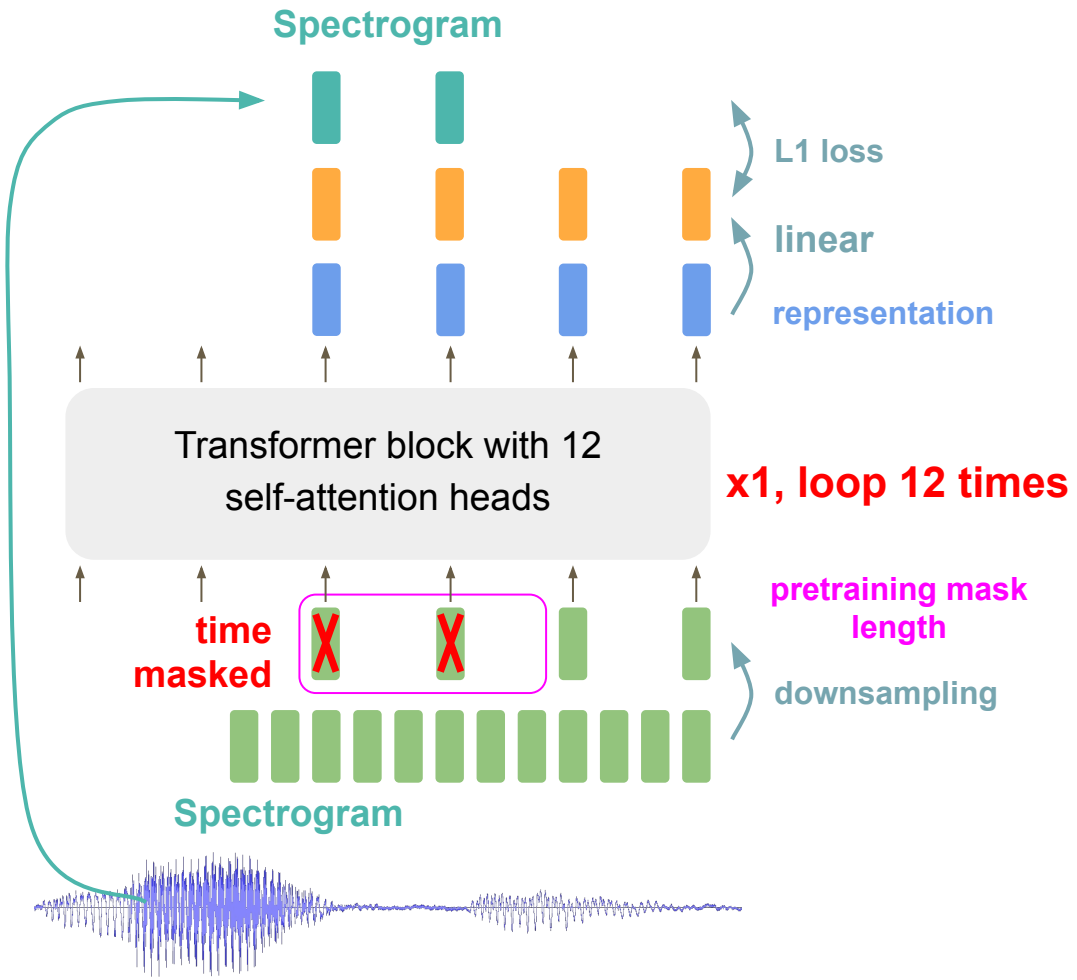
Mockingjay



Pre-training

Pre-training

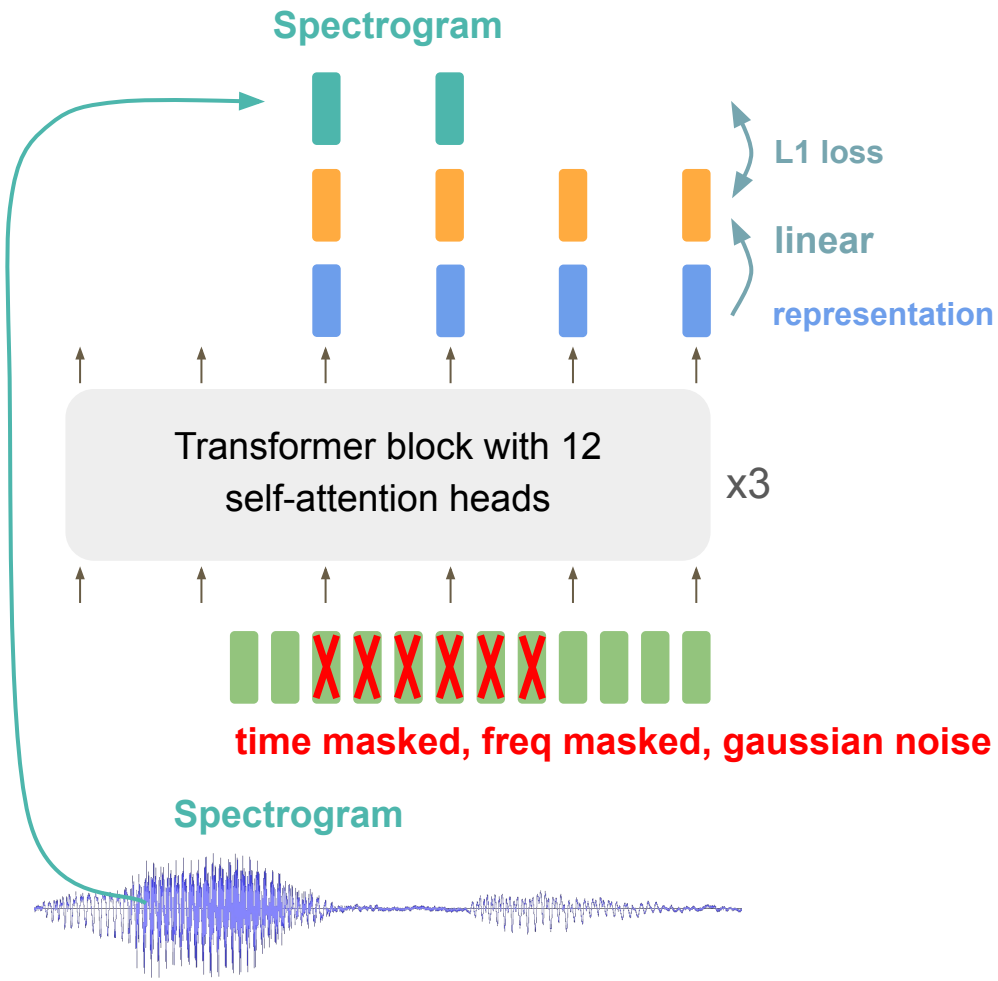
AudioAlbert



Pre-training

Pre-training

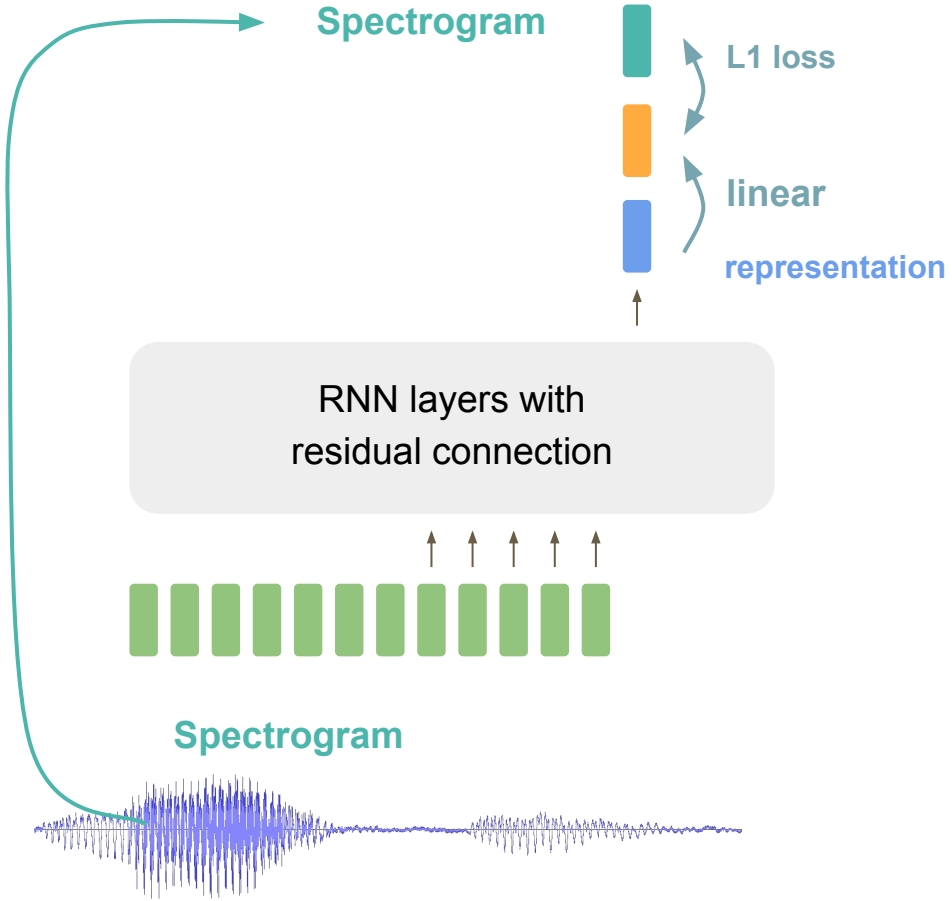
TERA



Pre-training

Pre-training

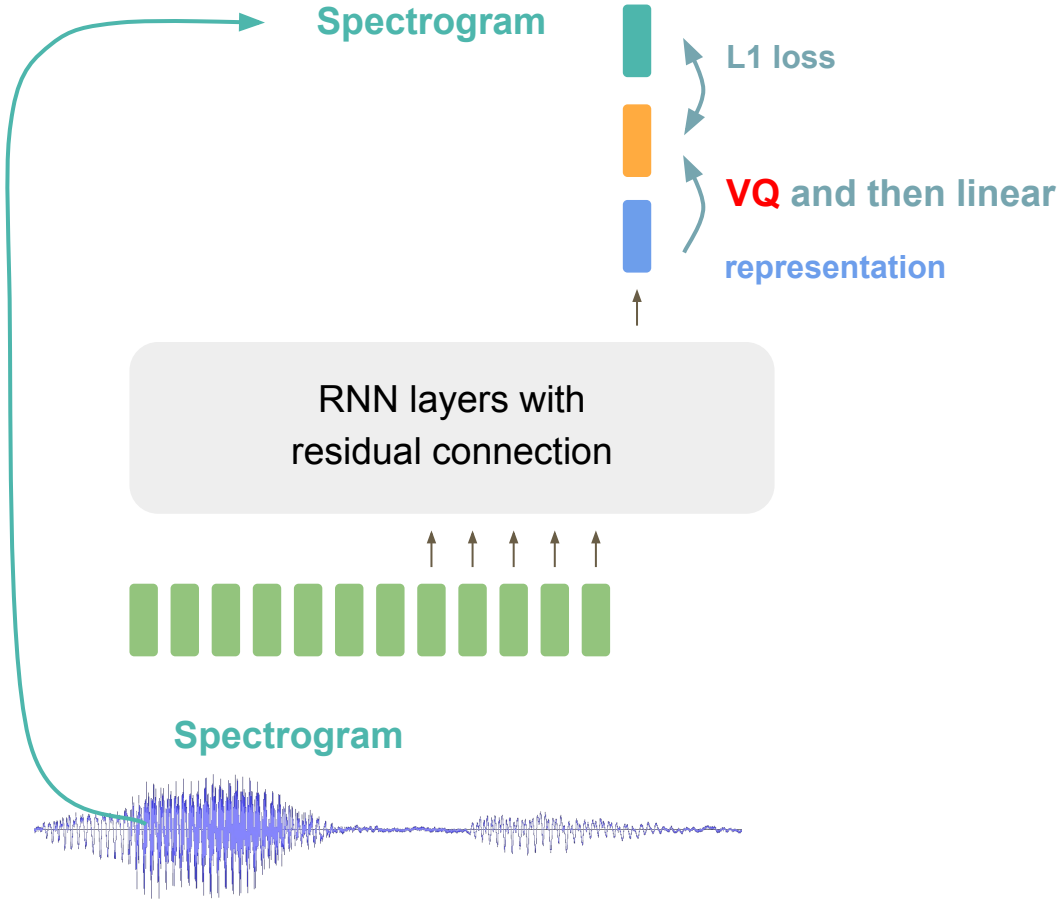
APC



Pre-training

Pre-training

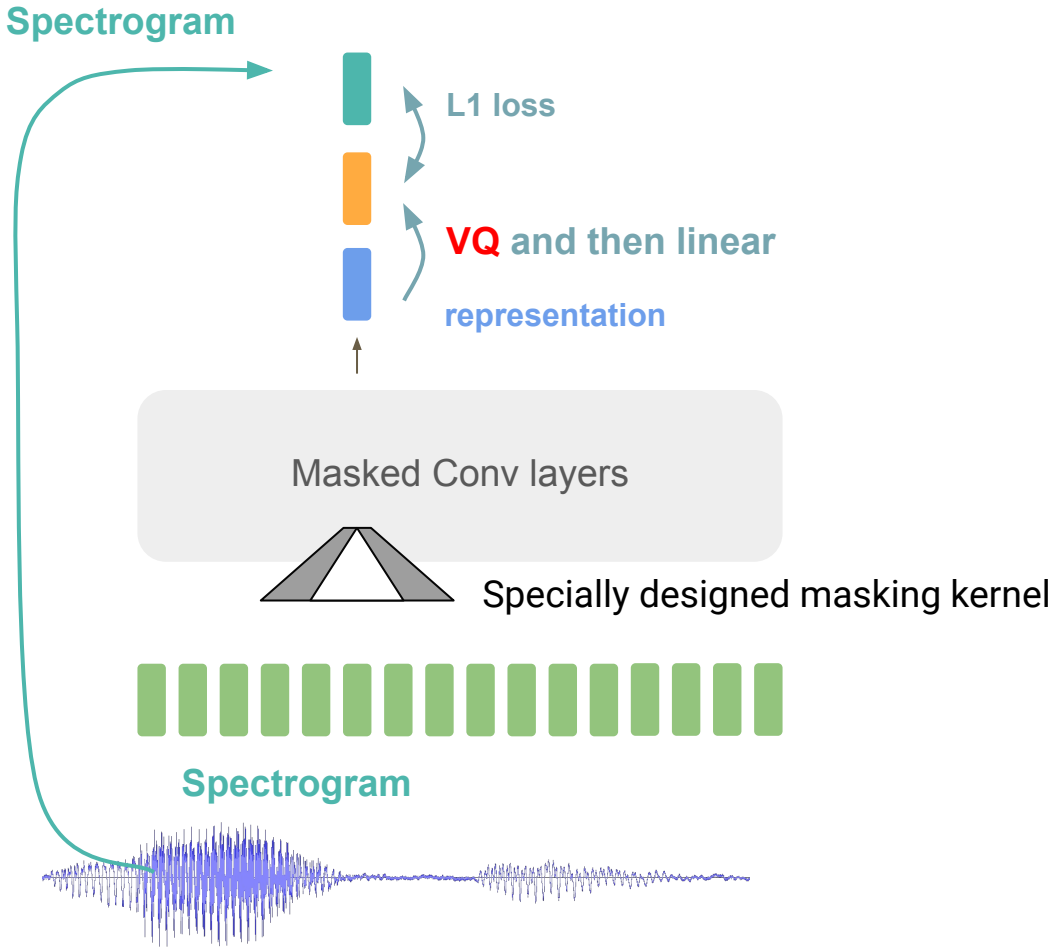
VQ-APC



Pre-training

Pre-training

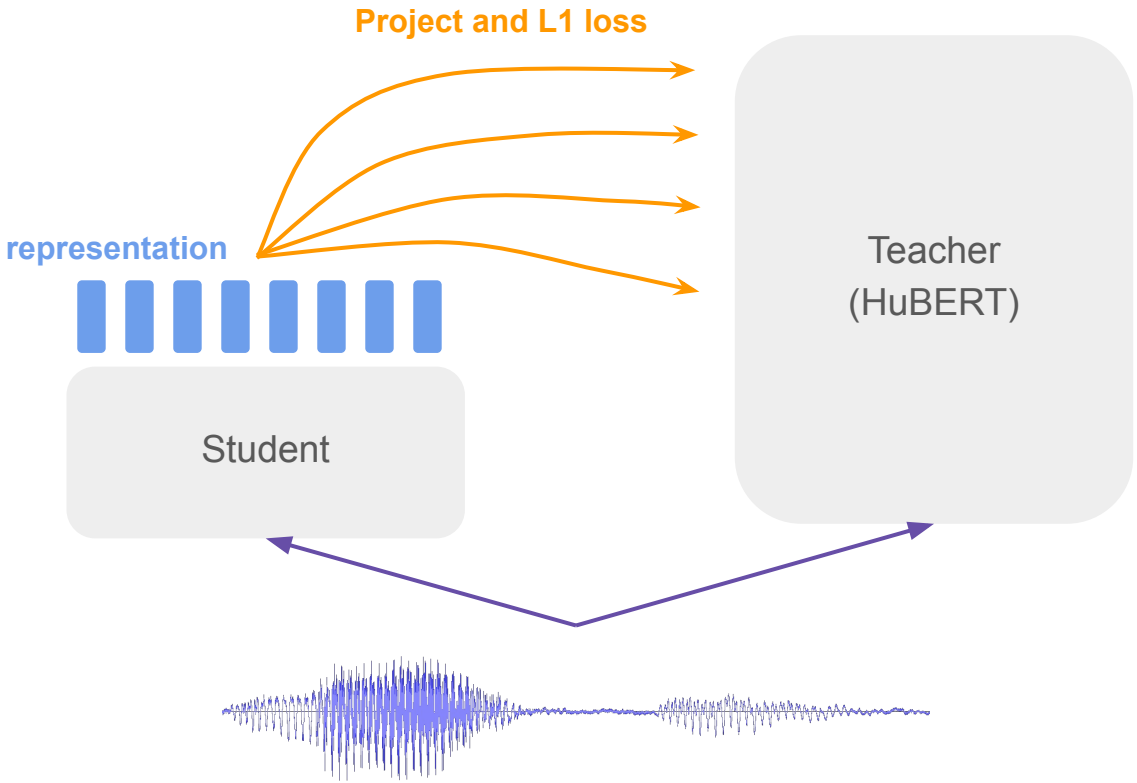
NPC



Pre-training

Pre-training

DistilHubert



Pre-training

- An unified trainer logic for all pretraining methods
- Support Distributed Data Parallel (DDP): Easily scale up!
- Where to start?
  - https://github.com/s3prl/s3prl/blob/master/s3prl/pretrain/README.md
  - Download LibriSpeech
  - Change the dataset path in the config file
  - Train! (`mockingjay` can be changed to `tera`/`audio_albert`/`distilhubert` … etc)

```
python run_pretrain.py -u mockingjay \
    -g pretrain/mockingjay/config_model.yaml -n ExpName
```

Pre-trained model collection

Pre-trained model collection

Isn't wav2vec 2.0 / HuBERT / WavLM / data2vec always the best?

No! Different task, different story, like VC

Generative

Mockingjay | TERA | AudioAlbert | APC

VQ-APC | NPC | DeCoAR | DeCoAR 2.0

Contrastive

Modified CPC | wav2vec | vq-wav2vec

discreteBERT | wav2vec 2.0

Predictive

HuBERT | Unispeech-SAT | WavLM | data2vec

Multi-task

PASE+

Distillation

DistilHuBERT | LightHuBERT

Pre-trained model collection

- Simplify model preparation
- Auto-downloading and caching

Installation

```
$ pip install s3prl torch torchaudio
```

Get model by name

```
model = s3prl.hub.wav2vec2().cuda()
# or
model = torch.hub.load("s3prl/s3prl", "wav2vec2").cuda()
# checkpoint will be downloaded and cached automatically
```

Pre-trained model collection

- One unified I/O interface for all models
- List of waveforms in → batched all-layer hidden states out
- Model-specific frontend preprocessing. E.g. Mel-spectrogram
  - done on-the-fly in forward
  - accelerated by GPU

Extract features

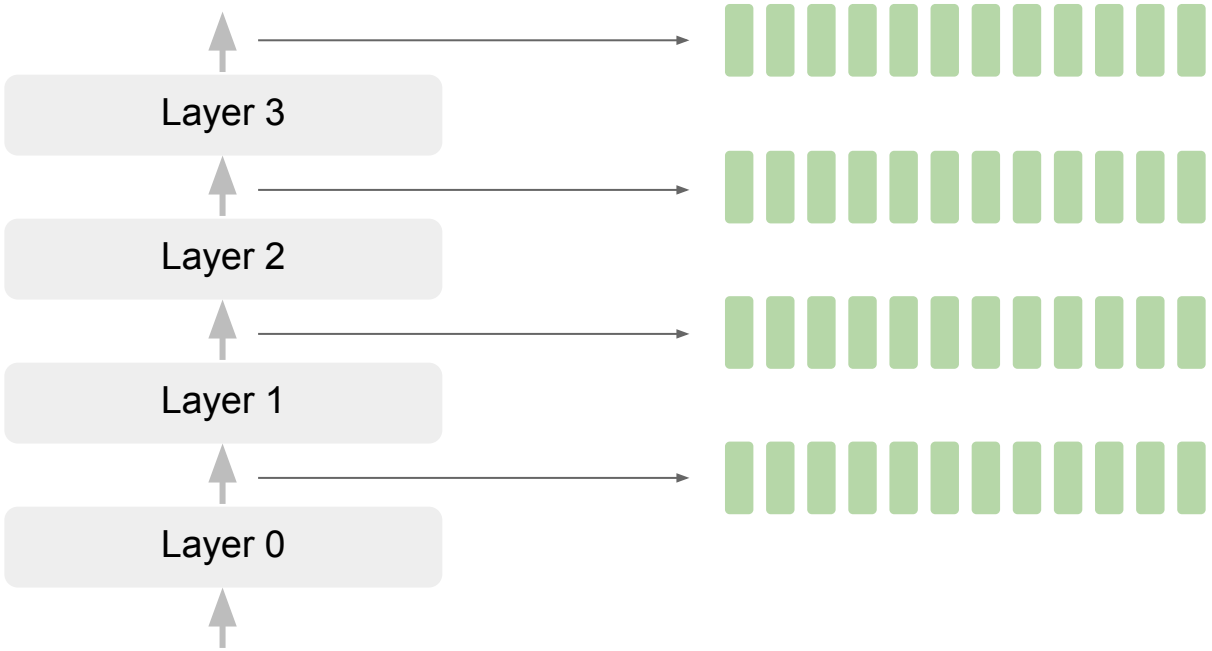
```
wav1 = torchaudio.load(“your audio path”).view(-1).cuda()
wav2 = ...
```

```
all_representations = model([wav1, wav2])
```

 # padding and masking is done automatically in the correct way

Pre-trained model collection

- Extract all the hidden states for all models



Pre-trained model collection

Playing with the representation

```
>>> type(all_representations)
<class 'dict'>
```

```
>>> len(all_representations[“hidden_states”])
13    # wav2vec2: conv’s output + 12 transformer layers’ outputs (in order)
```

```
>>> all_representations[“hidden_states”][0].dim()
3    # (batch_size, max_sequence_length, hidden_size)
```

```
>>> all_representations[“c”].dim()
3    # wav2vec: (batch_size, max_sequence_length, hidden_size)
```

Pre-trained model collection

- Lots of checkpoint variants!

TERA	APC	wav2vec 2.0	HuBERT
tera_100hr	apc_100hr	wav2vec2_960	hubert_960
tera_360hr	apc_360hr	wav2vec2_ll60k	hubert_ll60k
tera_960hr	apc_960hr	...	...
...	...		

Pre-trained model collection

List all available checkpoints

```
>>> dir(s3prl.hub)
```

 # or

```
>>> torch.hub.list("s3prl/s3prl")
```

```
["apc", "apc_360hr", "apc_960hr", "tera_100hr",
"hubert", "hubert_large_ll60k", …, "wav2vec2"]
```

Get model by name

```
>>> model = s3prl.hub.tera_100hr().cuda()
```

Understand your checkpoint

https://github.com/s3prl/s3prl/tree/master/s3prl/upstream#upstream-information

Downstream benchmarking & SUPERB

Downstream Benchmarking & SUPERB

Tasks in Mockingjay, TERA

- phoneme frame classification
- MOSEI sentiment
- speaker frame/utterance classification

SUPERB

- phoneme recognition
- speaker ID
- emotion recognition
- intent classification
- keyword spotting
- ASR
- ASV
- query by example
- slot filling
- speaker diarization

SUPERB-SG

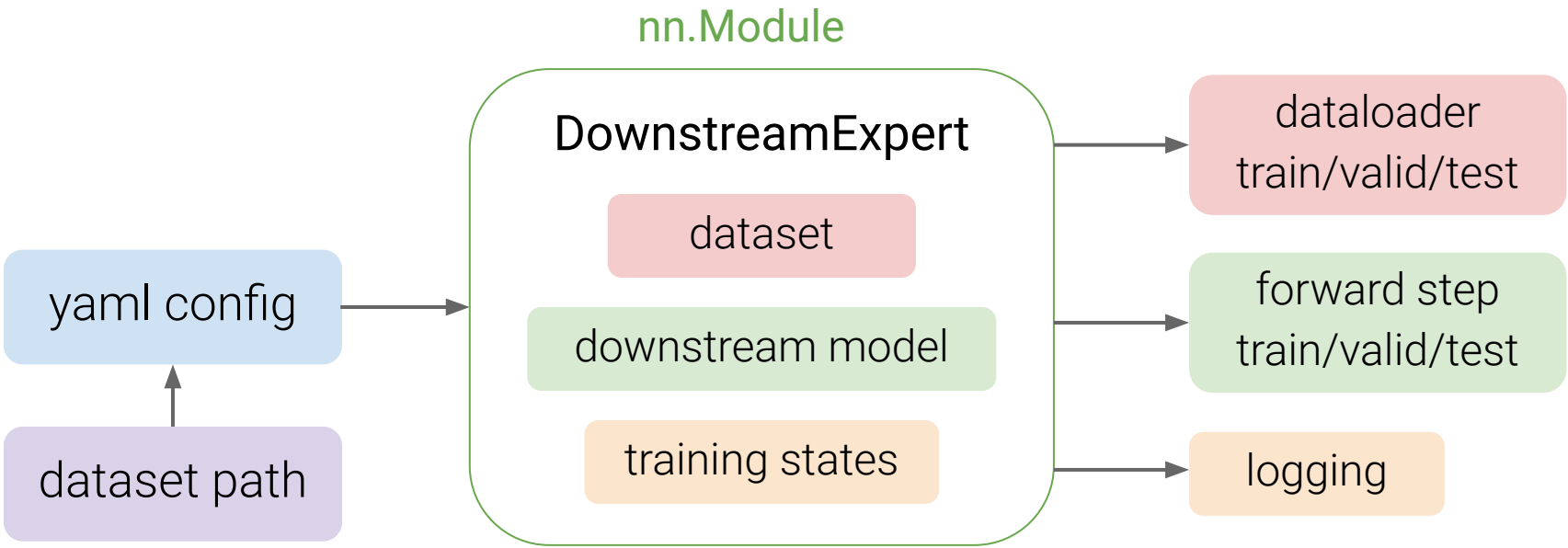
- speech translation
- speech enhancement
- source separation
- out-of-domain ASR
- voice conversion

others

- atis
- audio snips

Downstream benchmarking & SUPERB

- Each downstream task is a `nn.Module`



Downstream benchmarking & SUPERB

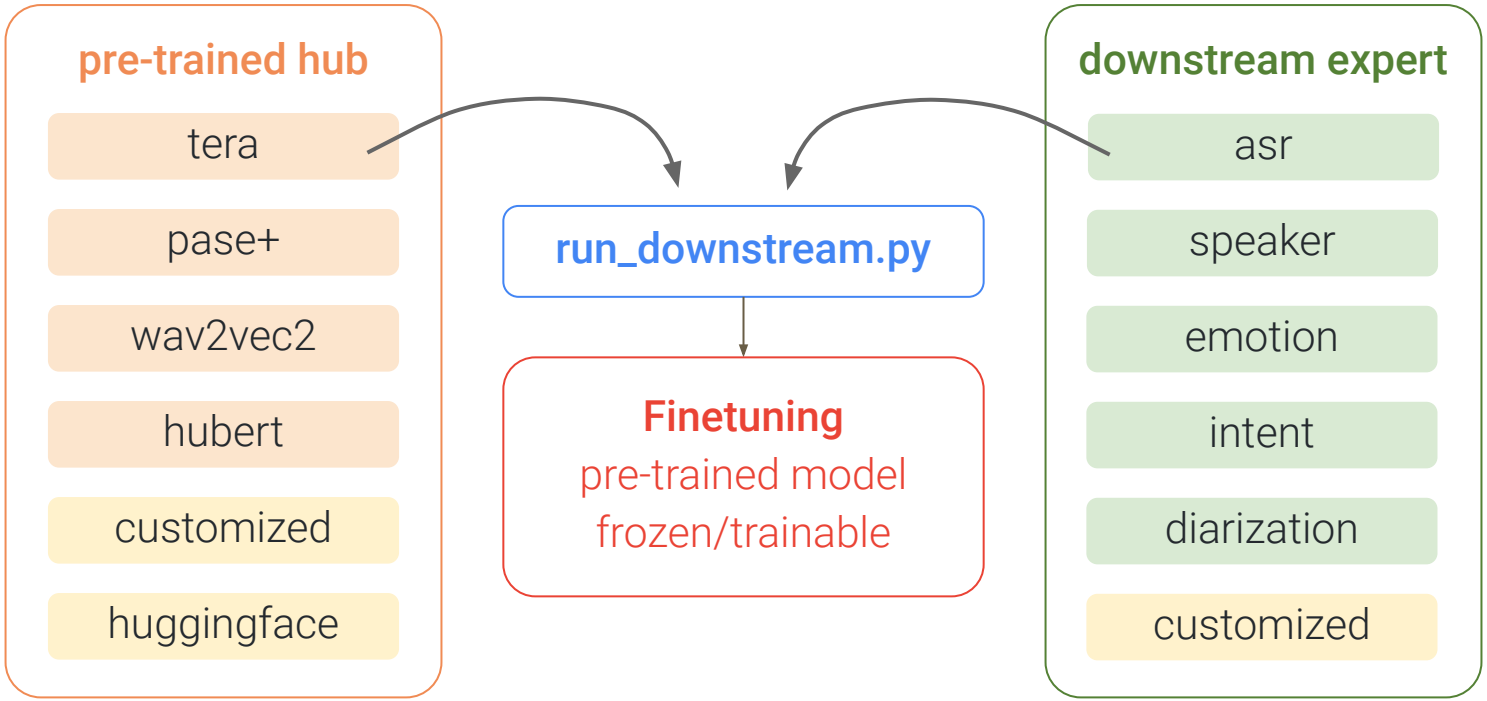
```
from collections import defaultdict
from s3prl.downstream import asr

config[“path”] = LIBRISPEECH_PATH
expert = asr(**config).cuda()
records = defaultdict(list)
for batch in expert.get_dataloader(“train”):
    batch = batch.cuda()
    loss = expert(“train”, batch, records)

    if is_logging_step:
        expert.log_records(“train”, records, tensorboard)
```

Use Trainer in other toolkit

- PyTorch Lightning
- Huggingface
- Espnet
- Speechbrain

Downstream benchmarking & SUPERB

- An unified downstream finetuning script (DDP supported)



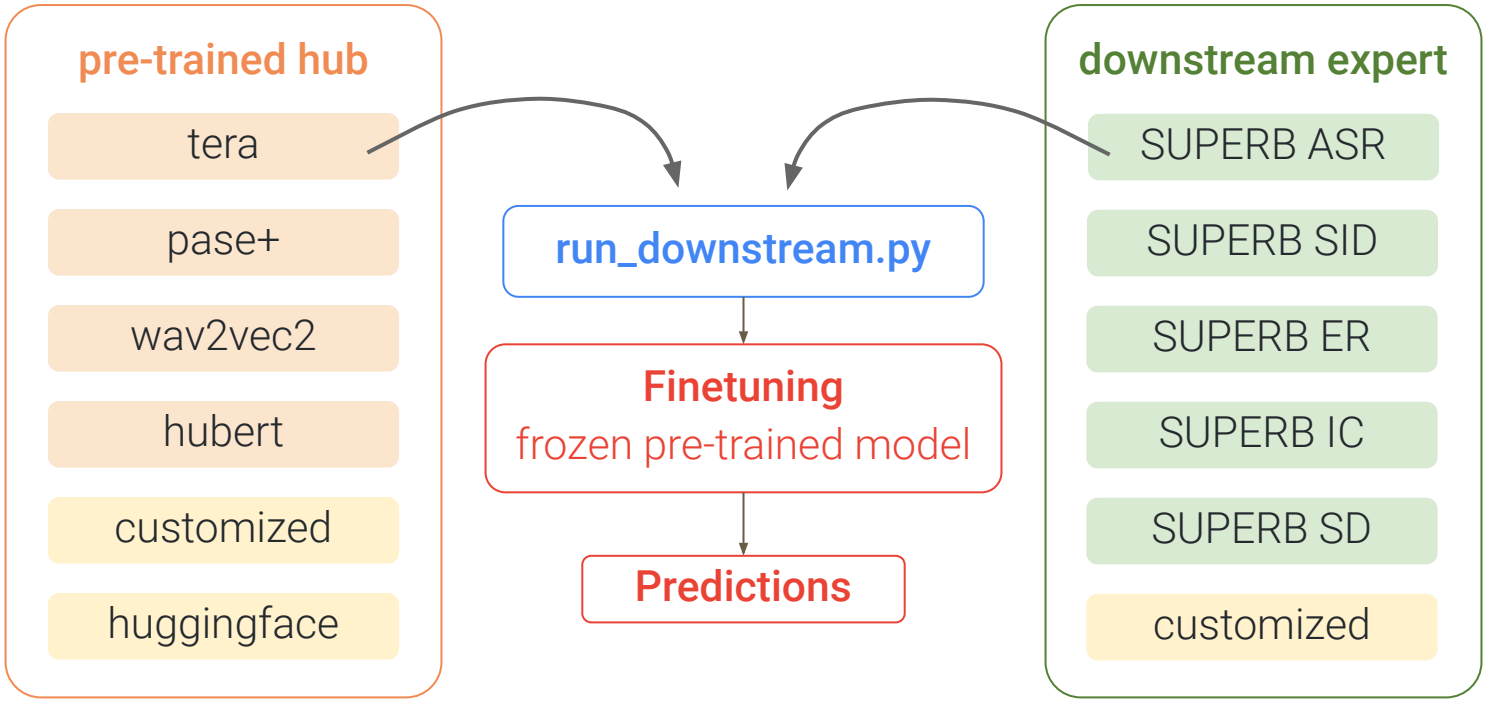
Downstream benchmarking & SUPERB

- Where to start?
  - General usage of the unified finetuning script
    - https://github.com/s3prl/s3prl/blob/master/s3prl/downstream/README.md
  - Task-specific usage
    - https://github.com/s3prl/s3prl/blob/master/s3prl/downstream/docs/superb.md#asr-automatic-speech-recognition
    - Prepare the dataset according to the README
    - Change the dataset path in the config file
    - Train!

```
python run_downstream.py -u hubert -d asr -p ExpDir
```

SUPERB leaderboard submission https://superbbenchmark.org/

- An unified downstream finetuning script (DDP supported)



SUPERB leaderboard submission https://superbbenchmark.org/

Submissions we received in the past year

DeCoAR Team helped merge the model into S3PRL and we benchmarked it

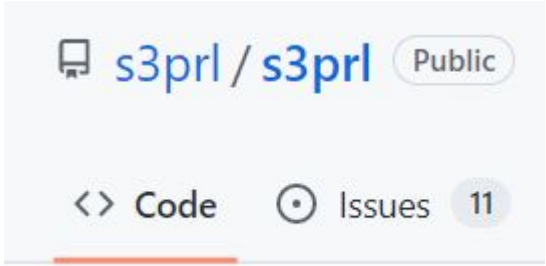
Method	Name	Description	URL	Rank ↑	Score ↑	PR ↓	KS ↑	IC ↑	SID ↑
WavLM Large	Microsoft	M-P + VQ …		19.9	1145	3.06	97.86	99.31	95.49
WavLM Base+	Microsoft	M-P + VQ …		18.7	1106	3.92	97.37	99	89.42
WavLM Base	Microsoft	M-P + VQ …		16.9	1019	4.84	96.79	98.63	84.51
HuBERT Large	paper	M-P + VQ	-	15.8	919	3.53	95.29	98.76	90.33
wav2vec 2.0 Large	paper	M-C + VQ	-	15.4	914	4.75	96.66	95.28	86.14
HuBERT Base	paper	M-P + VQ	-	15.25	941	5.41	96.3	98.34	81.42
LightHuBERT Small	LightHuBE…	Once-for-…		13.95	901	6.6	96.07	98.23	69.7
FaST-VGS+	Puyuan P…	FaST-VG…	-	13.15	809	7.76	97.27	98.97	41.34
wav2vec 2.0 Base	paper	M-C + VQ	-	12.35	818	5.74	96.23	92.35	75.18
DistilHuBERT	Heng-Jui …	multi-task …	-	11.2	717	16.27	95.98	94.99	73.54
DeCoAR 2.0	paper	M-G + VQ	-	10.6	722	14.93	94.48	90.8	74.42

Predictions

Submit! https://github.com/s3prl/s3prl/blob/master/s3prl/downstream/docs/superb.md#leaderboard-submission

Issues & Questions & Collaboration & Contact

- Any suggestion / feature request is welcome!
  - (We are under a huge restructure for the next major release)
- Directly open an issue on https://github.com/s3prl/s3prl
- Email
  - Shu-wen (Leo) Yang leo19941227@gmail.com and
  - Andy T. Liu tingweiandyliu@gmail.com and
  - Prof. Hung-yi Lee tlkagkb93901106@gmail.com
- Admin
  - Andy and Leo



Concluding Remarks

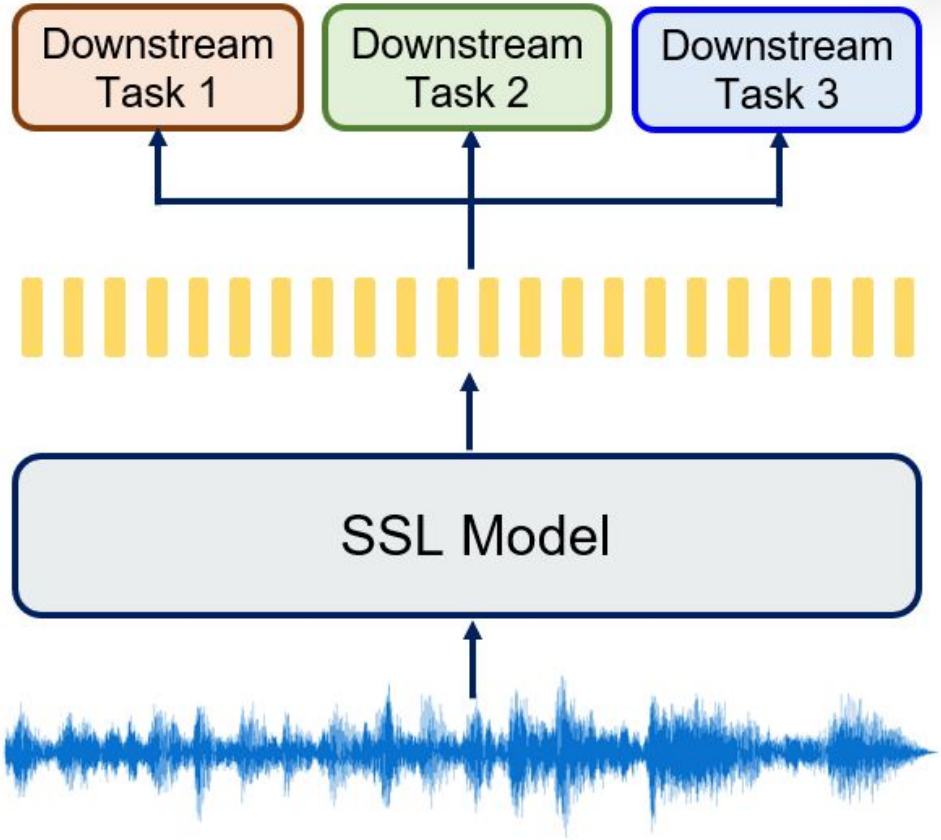
Hung-yi Lee

Concluding Remarks

- Review the history of representation learning
- Overview SSL pre-text tasks
- SSL with visual information
- Benchmarks of SSL models
- Analysis of SSL models
- From Representation Learning to Zero Resources
- Topics beyond accuracy
- S3PRL toolkit

Applications

Operating Systems



Let’s welcome the era of self-supervised Learning.

To Learn More ……

Self-Supervised Speech Representation Learning: A Review

Abdelrahman Mohamed*, Hung-yi Lee*, Lasse Borgholt*, Jakob D. Havtorn*, Joakim Edin, Christian Igel
Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, Shinji Watanabe

https://arxiv.org/abs/2205.10643

Tutorial website: https://sites.google.com/view/tutorial-ssl-speech