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

電信碩二 謝文崴 r11942078

Introduction

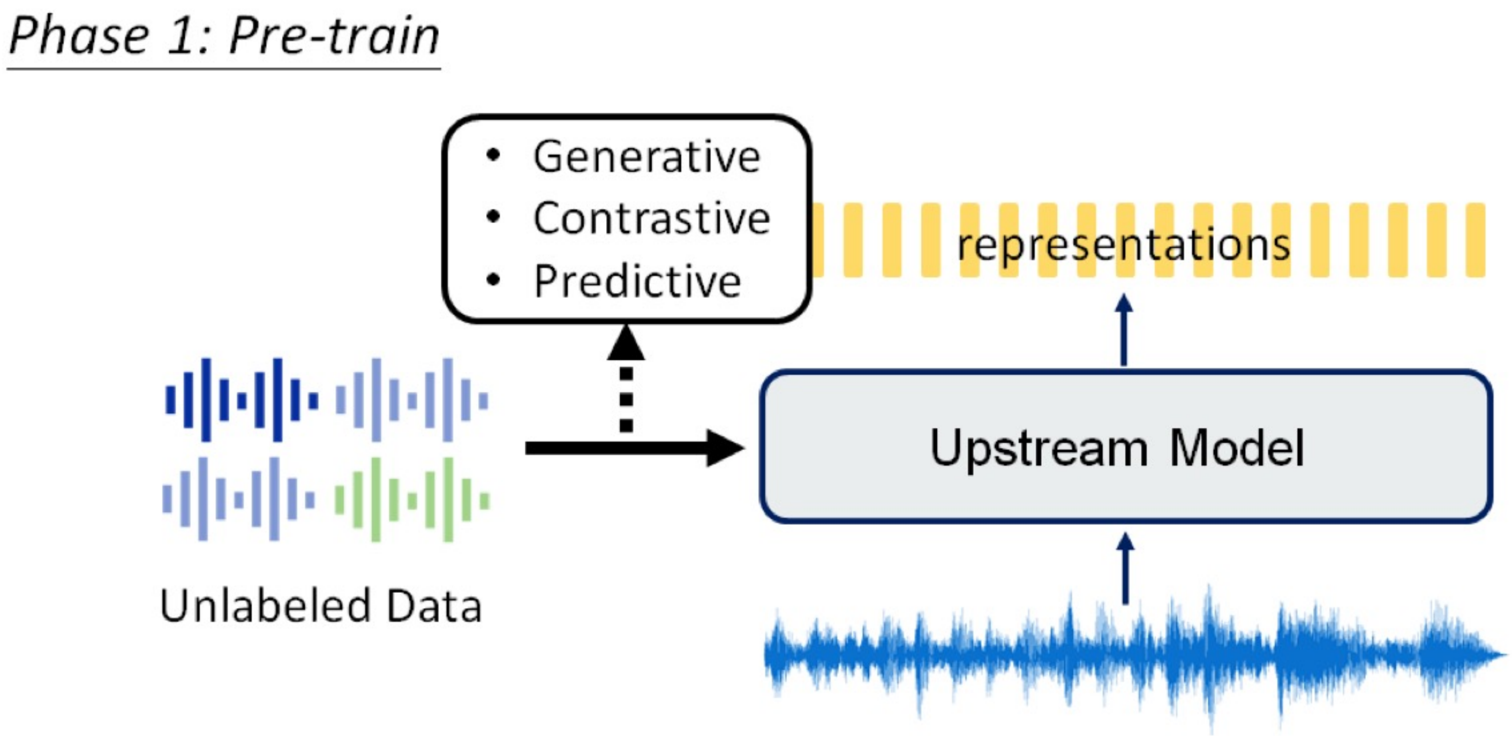


Fig. 1: Framework for using self-supervised representation learning in downstream applications

Introduction

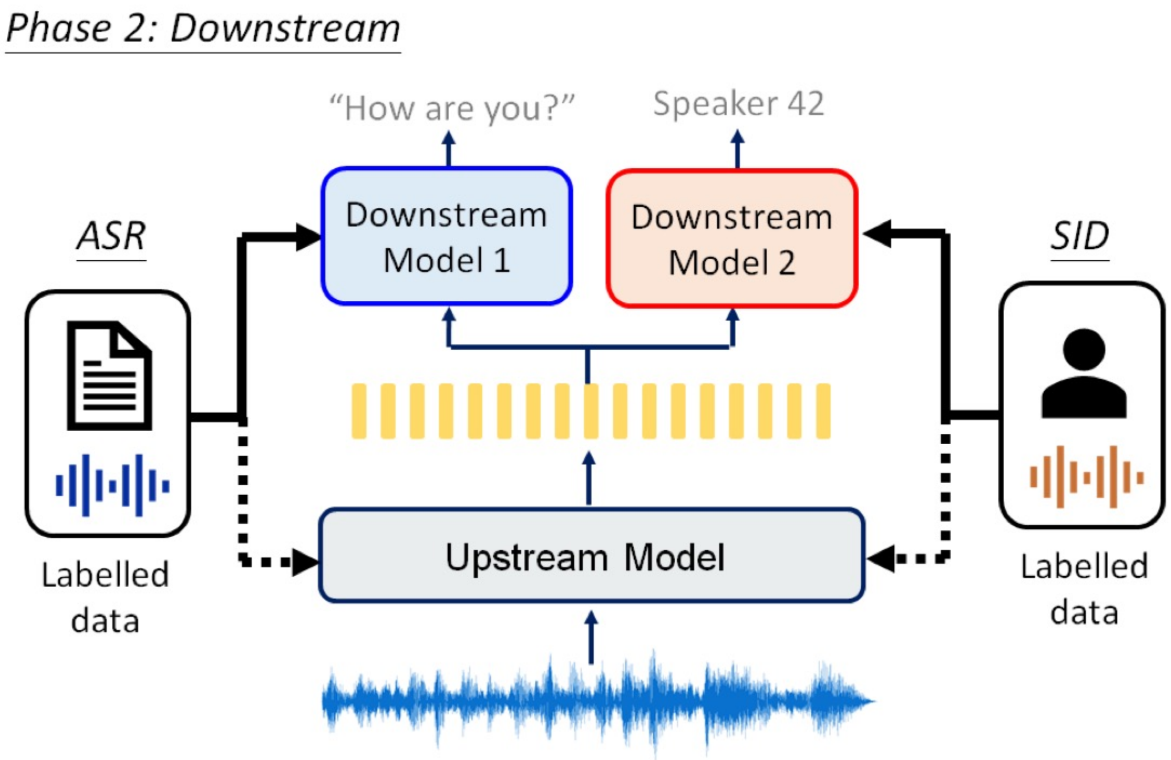


Fig. 1: Framework for using self-supervised representation learning in downstream applications

Introduction

- learn speech representations that capture low-level acoustic events, lexical knowledge, all the way to syntactic and semantic information

Introduction

- Since spoken utterances contain much richer information than the corresponding text transcriptions—e.g., speaker identity, style, emotion, surrounding noise, and communication channel noise—it is important to learn representations that disentangle these factors of variation.

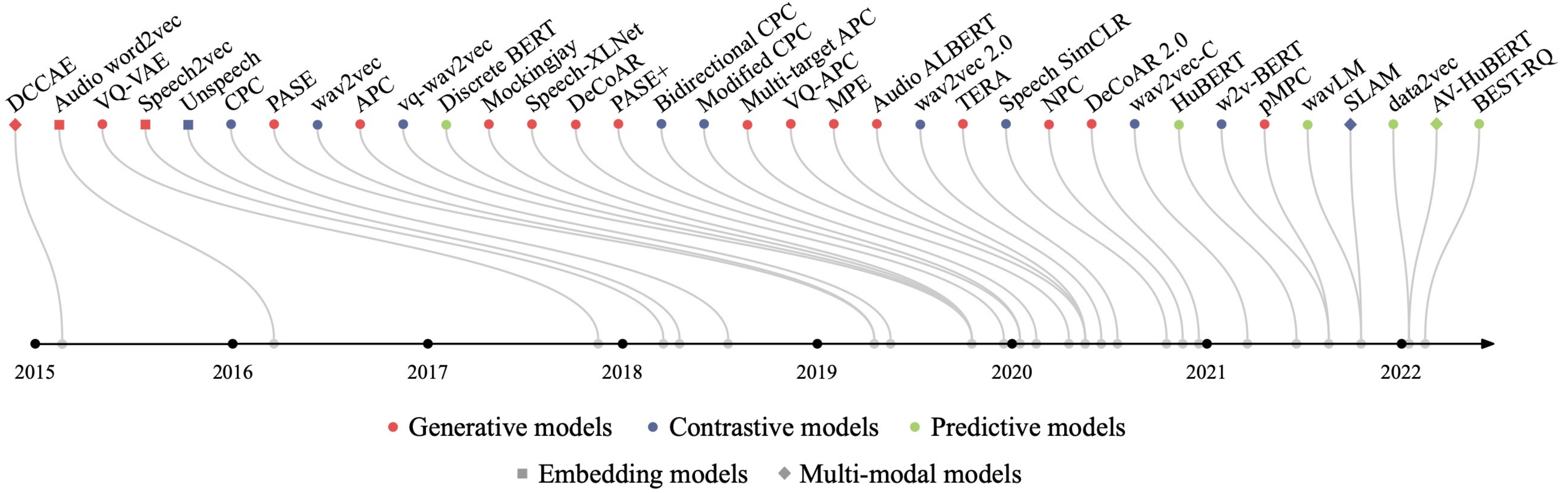
Introduction

- Learning feature hierarchies at the acoustic, lexical, and semantic levels supports applications with different requirements.

The characteristics of speech

- Speech is a sequence.
- Speech is a long sequence without segment boundaries.
- Speech is continuous.
- Speech processing tasks are diverse.

A selection of models



A. Generative approaches

- Motivation:

The pretext task is to generate, or reconstruct, the input data based on some limited view.

This includes predicting future inputs from past inputs, masked from unmasked, or the original from some other corrupted view.

A. Generative approaches

- Approaches:
  - Autoencoding: The pretext task is to reconstruct the given input.

$$\mathcal{L}_t = \underbrace{\log p(x_t | q_{[1,t]})}_{\text{encoder+decoder}} + \underbrace{\mathbf{MSE}\left(\mathrm{sg}\left[h_t\right], A\right)}_{\text{codebook}} + \underbrace{\alpha\ \mathbf{MSE}\left(h_t, \mathrm{sg}\left[A\right]\right)}_{\text{encoder}},$$

  - Autoregressive prediction:

$$\begin{aligned} H_{[1,t]} &= f(X_{[1,t]}), \\ \hat{x}_{t+c} &= g(h_t), \\ \mathcal{L}_t &= \|\hat{x}_{t+c} - x_{t+c}\|_1. \end{aligned}$$

A. Generative approaches

- Approaches:
  - Masked Reconstruction:

$$
\begin{aligned}
H &= f(m(X)),\\
\hat{x}_t &= g(h_t),\\
\mathcal{L}_t &= \|\hat{x}_t - x_t\|_1.
\end{aligned}
$$

  - More Generative Approaches:

    Using multiple targets, including the waveform, log power spectrum, mel cepstral coefficients (MFCCs), and prosody features.

A. Generative approaches

- Challenges:

A speech signal encodes more information than text, such as speaker identity and prosodic features, which makes it harder to generate.

B. Contrastive approaches

- Motivation:

Contrastive models learn representations by distinguishing a target sample (positive) from distractor samples (negatives) given an anchor representation.

B. Contrastive approaches

- Approaches:
  - CPC: Contrastive Predictive Coding uses a convolutional module f1(·) to produce localized representations z_t with a recurrent module f2(·) on top that outputs a contextualized representation h_t.

$$
\begin{aligned}
z_t &= f_1(X_{[t-u,t+u]} \\
H_{[1,t]} &= f_2(Z_{[1,t]}) \ , \\
\hat{z}_{t,k} &= g_k(h_t) \ .
\end{aligned}
$$

$$
\mathcal{L}_{t,k} = -\log\left(\frac{\exp(\hat{z}_{t,k}^{\mathrm{T}} z_{t+k})}{\sum_{i\in\mathcal{I}} \exp(\hat{z}_{t,k}^{\mathrm{T}} z_i)}\right)
$$

B. Contrastive approaches

- Approaches:
  - wav2vec 2.0: The wav2vec 2.0 model combines contrastive learning with masking.

$$
\begin{aligned}
z_t &= f_1(X_{[t-u,t+u]}) \\
H &= f_2(m(Z)) \quad , \\
q_t &= g(z_t) \quad .
\end{aligned}
\qquad
\mathcal{L}_t = -\log\left(\frac{\exp(S_{\mathrm{c}}(h_t, q_t))}{\sum_{i\in\mathcal{I}} \exp(S_{\mathrm{c}}(h_t, q_i))}\right)
$$

B. Contrastive approaches

- Challenges:

Since speech input is smooth and lacks natural segmentation, it can be difficult to define a contrastive sampling strategy that is guaranteed to provide samples that always relate to the anchor as truly positives and negatives in a sound way.

C. Predictive approaches

• Motivation:

Similar to the contrastive approaches, but they do not employ a contrastive loss and instead use a loss function such as squared error and cross- entropy.

C. Predictive approaches

- Approaches:
  - Discrete BERT: Discrete units c_t are first extracted with the vq-wav2vec model f1(·) and then used as inputs and targets in a standard BERT model f2(·).

$$
\begin{aligned}
c_t &= f_1(X_{[t-u,t+u]}) \ , \\
H &= f_2(m(C)) \ , \\
\hat{c}_t &= g(h_t) \ .
\end{aligned}
$$

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

C. Predictive approaches

- Approaches:
  - HuBERT: Intuitively, the HuBERT model is forced to learn both an acoustic and a language model.

$$
\begin{aligned}
c_t &= g_1(X_{[t-w,t+w]}) \ , \\
z_t &= f_1(X_{[t-u,t+u]}) \ , \\
H &= f_2(m(Z)) \ , \\
\hat{c}_t &= g_2(h_t) \ ,
\end{aligned}
$$

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

$$
\mathcal{L} = \beta \mathcal{L}_m + (1-\beta)\mathcal{L}_u \ .
$$

C. Predictive approaches

- Challenges:

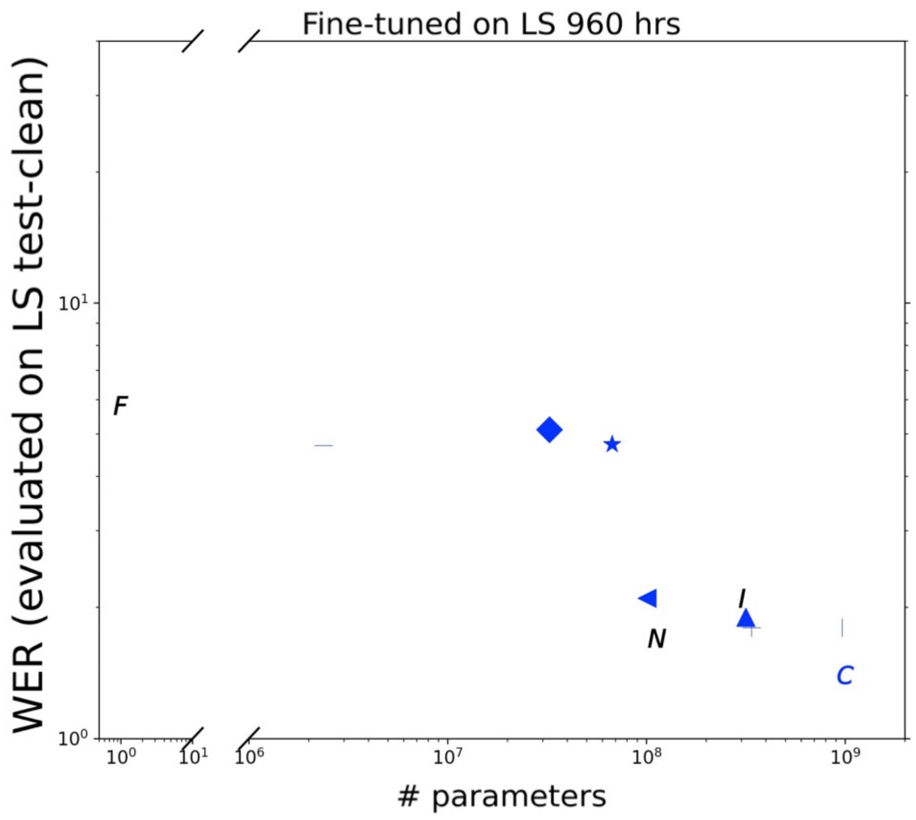
The iterative nature of pre-training for the HuBERT and wavLM could present a practical inconvenience when working with large volumes of data.

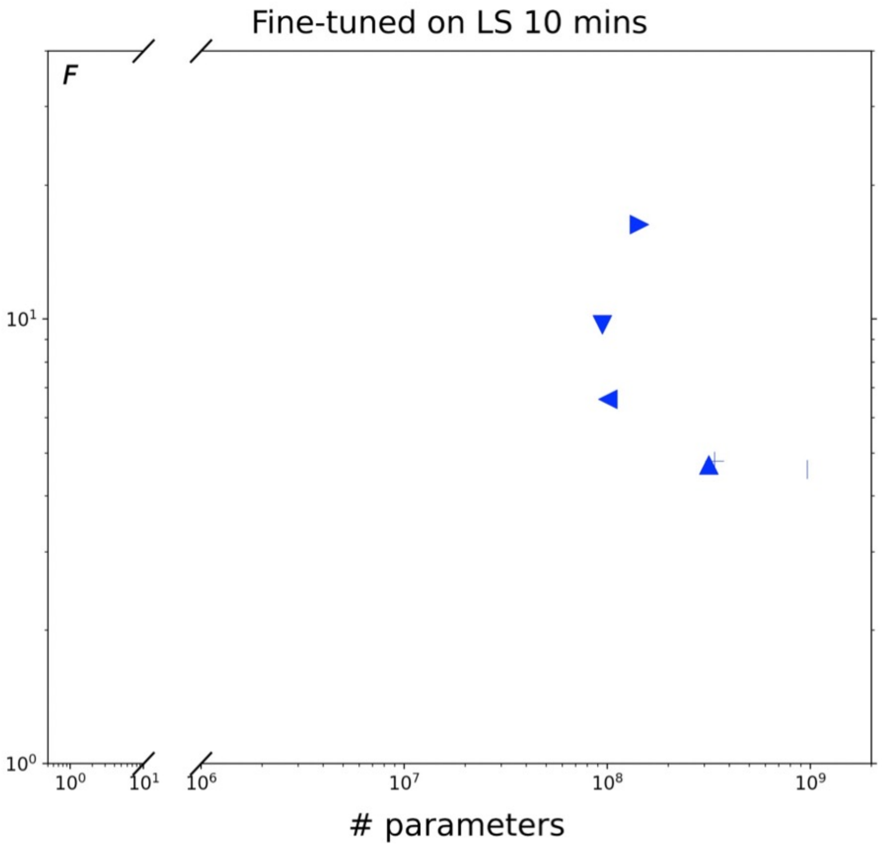
D. Learning from multi-modal data

- Multiple modalities are useful in many settings, where each modality provides information that is complementary to other modalities.

- In addition, learning from speech data with accompanying signals such as images or video can help learn representations that encode more semantic information.

Benchmark results





Benchmark results

TABLE V: Tasks where the state of the art is models with SSL pre-training.

Tasks	Dataset	non-SSL	SSL
ASR (WER $\downarrow$)	LS test-clean/other	2.1/4.0 [63]	1.4/2.6 [255]
IC (Acc $\uparrow$)	FSC	98.8 [237]	99.3 [219]
SID (Acc $\uparrow$)	VoxCeleb1	94.8 [256]	95.5 [131]
ASV (EER $\downarrow$)	VoxCeleb1	3.1 [257]	2.4 [258]
QbE (MTWV $\uparrow$)	QUESST (EN)	10.6 [259]	11.2 [219]

Future research directions

- Using the representation model.
- Increasing the efficiency of the representation model.
- Data-efficient approaches.
- Feature Disentanglement.
- Creating robust models.
- Capturing higher-level semantic information.
- Using text representation models to improve speech representation.

References

- https://arxiv.org/abs/2205.10643