Towards Universal Self-supervised Model for Speech Processing

Hung-yi Lee

https://speech.ee.ntu.edu.tw/~hylee/
**Self-supervised Learning Framework**

**Phase 1: Pre-train**

- Mask the input signals and then reconstruct them.
- Predict the targets obtained without human efforts.
- Contrastive learning

Unlabeled Data → Pre-trained Model

- Task-agnostic

(representations)

(not complete survey)
Mockingjay
mimic sound it hears

Learn to reconstruct

Some of the input are masked

masked
masked
Self-supervised Learning Framework

Phase 1: Pre-train

PASE+  APC  NPC  Mockingjay  DeCoAR  Wav2vec  HuBERT

Just name a few ...

Unlabeled Data

Pre-trained Model

representations
**Self-supervised Learning Framework**

**Phase 2: Downstream**

A downstream task to be solved (e.g., ASR)

- **Labelled data**
- **Pre-trained Model**
- **Downstream Model**

“How are you?”

Input: Unlabelled data

Output: Predictions for **Downstream Model**

- **Downstream Model** takes the **Pre-trained Model** as input and outputs predictions.

**Labelled data** is used to train the **Pre-trained Model**.

**Diagram**:
- The **Labelled data** is fed into the **Pre-trained Model**.
- The **Pre-trained Model** outputs predictions that are then used to train the **Downstream Model**.
- The **Downstream Model** takes the **Labelled data** as input and outputs predictions.

**Example**:

- Input: “How are you?”
- **Pre-trained Model** outputs predictions.
- **Downstream Model** takes the predictions and outputs the final prediction: “I’m fine, thank you.”
Specialist? Universal?

Just name a few ...

PASE+  APC  NPC  Mockingjay  DeCoAR  Wav2vec  HuBERT

They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?
How are you?
They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?

- I believe they are specialist.
- To be good at ASR, a model learns to extract content and ignore speaker.
- Hence, super good on ASR Poor performance on speaker related tasks.

My two cents (one year ago)
SUPERB
Speech processing Universal PERformance Benchmark

PASE+  APC  NPC  Mockingjay  DeCoAR  Wav2vec  HuBERT

ASR  Query-by-example  Speaker Identification  Speaker verification  Spoken Slot Filling

Phoneme recognition  Keyword spotting  Emotion  Speaker Diarization

https://arxiv.org/abs/2105.01051
SUPERB
Speech processing Universal PERformance

SUPERB: Speech processing Universal PERformance Benchmark

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhota, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, Hung-yi Lee

Presented at INTERSPEECH 2021
# Introduction of Contestants

<table>
<thead>
<tr>
<th>Method</th>
<th>Network</th>
<th>#Params</th>
<th>Stride</th>
<th>Input</th>
<th>Corpus</th>
<th>Pretraining</th>
<th>Official Github</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBANK</td>
<td>-</td>
<td>7.83M</td>
<td>0</td>
<td>waveform</td>
<td>-</td>
<td>multi-task</td>
<td>santi-pdp / pase</td>
</tr>
<tr>
<td>PASE+</td>
<td>SincNet, 7-Conv, 1-QRNN</td>
<td>4.66M</td>
<td>10ms</td>
<td>waveform</td>
<td>LS 50 hr</td>
<td></td>
<td>iamyuanchung / APC</td>
</tr>
<tr>
<td>APC</td>
<td>3-GRU</td>
<td>19.38M</td>
<td>10ms</td>
<td>waveform</td>
<td>LS 360 hr</td>
<td>F-G</td>
<td>iamyuanchung / VQ-APC</td>
</tr>
<tr>
<td>VQ-APC</td>
<td>3-GRU</td>
<td>21.33M</td>
<td>10ms</td>
<td>waveform</td>
<td>LS 960 hr</td>
<td>F-G + VQ</td>
<td>Alexander-H-Liu / NPC</td>
</tr>
<tr>
<td>NPC</td>
<td>4-Conv, 4-Masked Conv</td>
<td>85.12M</td>
<td>10ms</td>
<td>waveform</td>
<td>LS 360 hr</td>
<td>M-G + VQ</td>
<td>s3prl / s3prl</td>
</tr>
<tr>
<td>Mockingjay</td>
<td>12-Trans</td>
<td>89.84M</td>
<td>10ms</td>
<td>waveform</td>
<td>LS 960 hr</td>
<td>time M-G</td>
<td>s3prl / s3prl</td>
</tr>
<tr>
<td>TERA</td>
<td>12-Trans</td>
<td>32.54M</td>
<td>10ms</td>
<td>waveform</td>
<td>LL 60k hr</td>
<td>F-C</td>
<td>awslabs / speech-representations</td>
</tr>
<tr>
<td>DeCoAR 2.0</td>
<td>5-Conv, 1-LSTM</td>
<td>3.15M</td>
<td>10ms</td>
<td>waveform</td>
<td>LS 960 hr</td>
<td>F-C + VQ</td>
<td>pytorch / fairseq</td>
</tr>
<tr>
<td>modified CPC</td>
<td>7-Conv 12-Trans</td>
<td>95.04M</td>
<td>20ms</td>
<td>waveform</td>
<td>LS 960 hr</td>
<td>M-C + VQ</td>
<td>pytorch / fairseq</td>
</tr>
<tr>
<td>wav2vec</td>
<td>7-Conv 24-Trans</td>
<td>317.38M</td>
<td>20ms</td>
<td>waveform</td>
<td>LL 60k hr</td>
<td>M-C + VQ</td>
<td>pytorch / fairseq</td>
</tr>
<tr>
<td>wav2vec 2.0 Base</td>
<td>7-Conv 12-Trans</td>
<td>94.68M</td>
<td>20ms</td>
<td>waveform</td>
<td>LL 60k hr</td>
<td>M-P + VQ</td>
<td>pytorch / fairseq</td>
</tr>
<tr>
<td>wav2vec 2.0 Large</td>
<td>7-Conv 24-Trans</td>
<td>316.61M</td>
<td>20ms</td>
<td>waveform</td>
<td>LL 60k hr</td>
<td>M-P + VQ</td>
<td>pytorch / fairseq</td>
</tr>
<tr>
<td>HuBERT Base</td>
<td>7-Conv 12-Trans</td>
<td>84.10M</td>
<td>10ms</td>
<td>waveform</td>
<td>LS 960 hr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HuBERT Large</td>
<td>7-Conv 24-Trans</td>
<td>32.54M</td>
<td>10ms</td>
<td>waveform</td>
<td>LS 960 hr</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **G**: reconstructing the input
- **P**: token prediction
- **C**: contrastive learning
- **VQ**: quantization
- **F**: predicting future information
- **M**: input masking
Tasks - Content

**Phoneme Recognition**
- Transcribe
- /b/ /d/ /f/ /g/ ...

**Keyword Spotting**
- Classify
- Left / Right / Go ...

**ASR**
- Transcribe
- I want to pet a cat

**Query-by-Example**

**Spoken Document**

**Spoken Query**
Tasks – Speaker

**Speaker Identification**

**Speaker Verification**

**Speaker Diarization**

Utterance A  Utterance B

A & B same speaker? Yes / No

Great … The … of …Yep. … factors.

A X A

One … is, uh, … of … has … B
Tasks – Semantic

Intent Classification

classify

Intent classes

Slot Filling

extract

Slot type
from_location

Slot value
to_location

Taipei

New York

Tasks – Emotion

Emotion Recognition

classify

Happy/Angry/Sad ...

I fly from Taipei to New York

Please refer to the paper for more details.
Game Start!

Round 1
Rules in Round 1

How are you?

I only put two out of ten downstream models for simplicity.
Rules in Round 1

The network architecture of a downstream model is predefined.

Keep it simple

e.g., Linear layer

The network architecture of a downstream model is predefined.

Labelled data

Pre-trained Model

fixed

Last Layer Output

Downstream Model 1

“How are you?”

Speaker 42

e.g., 2-layer LSTM

Downstream Model 2

ASR

Labelled data

SID

Labelled data
Rules in Round 1 – Downstream

- Phoneme Recognition: linear layer
- Keyword Spotting: linear layer
- Speech Recognition: 2-layer LSTM
- Query-by-example: none
- Speaker Identification: linear layer
- Speaker Verification: the same as x-vector
- Speaker Diarization: 1-layer LSTM
- Intent Classification: linear layer
- Slot Filling: 2-layer LSTM

Keep it simple
Why so constrained?

“How are you?”

Speaker 42

Downstream Model 1

Downstream Model 2

Limited capacity

Universal features

Share across all the tasks

Pre-trained Model

fixed
Why so constrained?

“How are you?”

Speaker 42

“Downstream Model 1”

“Downstream Model 2”

“Pre-trained Model”

Universal features

Easy to build new applications!

This sounds too good to be true......
## Results of Round 1

<table>
<thead>
<tr>
<th></th>
<th>Content</th>
<th>Speaker</th>
<th>Semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PR</td>
<td>KS</td>
<td>ASR</td>
</tr>
<tr>
<td>fbank</td>
<td>82.01</td>
<td>8.63</td>
<td>15.21</td>
</tr>
<tr>
<td>PASE+</td>
<td>58.88</td>
<td>82.37</td>
<td>16.61</td>
</tr>
<tr>
<td>APC</td>
<td>41.85</td>
<td>91.04</td>
<td>15.09</td>
</tr>
<tr>
<td>VQ-APC</td>
<td>42.86</td>
<td>90.52</td>
<td>15.37</td>
</tr>
<tr>
<td>NPC</td>
<td>52.67</td>
<td>88.54</td>
<td>14.69</td>
</tr>
<tr>
<td>Mockingjay</td>
<td>80.01</td>
<td>82.67</td>
<td>15.94</td>
</tr>
<tr>
<td>TERA</td>
<td>47.53</td>
<td>88.09</td>
<td>12.44</td>
</tr>
<tr>
<td>modified CPC</td>
<td>41.66</td>
<td>92.02</td>
<td>13.57</td>
</tr>
<tr>
<td>wav2vec</td>
<td>32.39</td>
<td>94.09</td>
<td>11.3</td>
</tr>
<tr>
<td>vq-wav2vec</td>
<td>53.49</td>
<td>92.28</td>
<td>12.69</td>
</tr>
<tr>
<td>wav2vec 2.0 base</td>
<td>28.37</td>
<td>92.31</td>
<td>6.32</td>
</tr>
<tr>
<td>HuBERT base</td>
<td>6.85</td>
<td>95.98</td>
<td>4.93</td>
</tr>
</tbody>
</table>

### Pre-trained Models

- fbank
- PASE+
- APC
- VQ-APC
- NPC
- Mockingjay
- TERA
- modified CPC
- wav2vec
- vq-wav2vec
- wav2vec 2.0 base
- HuBERT base
### Results of Round 1

<table>
<thead>
<tr>
<th></th>
<th>Content</th>
<th>Speaker</th>
<th>Semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PR</td>
<td>KS</td>
<td>ASR</td>
</tr>
<tr>
<td>fbank</td>
<td>82.01</td>
<td>8.63</td>
<td>15.21</td>
</tr>
<tr>
<td>PASE+</td>
<td>58.88</td>
<td>82.37</td>
<td></td>
</tr>
<tr>
<td>APC</td>
<td>41.85</td>
<td>91.04</td>
<td>15.09</td>
</tr>
<tr>
<td>VQ-APC</td>
<td>42.86</td>
<td>90.52</td>
<td></td>
</tr>
<tr>
<td>NPC</td>
<td>52.67</td>
<td>88.54</td>
<td>14.69</td>
</tr>
<tr>
<td>Mockingjay</td>
<td>80.01</td>
<td>82.67</td>
<td></td>
</tr>
<tr>
<td>TERA</td>
<td>47.53</td>
<td>88.09</td>
<td>12.44</td>
</tr>
<tr>
<td>modified CPC</td>
<td>41.66</td>
<td>92.02</td>
<td>13.57</td>
</tr>
<tr>
<td>wav2vec</td>
<td>32.39</td>
<td>94.09</td>
<td>11.3</td>
</tr>
<tr>
<td>vq-wav2vec</td>
<td>53.49</td>
<td>92.28</td>
<td>12.69</td>
</tr>
<tr>
<td>wav2vec 2.0 base</td>
<td>28.37</td>
<td>92.31</td>
<td>6.32</td>
</tr>
<tr>
<td>HuBERT base</td>
<td>6.85</td>
<td>95.98</td>
<td>4.93</td>
</tr>
</tbody>
</table>

- Pre-trained models outperform fbank across many tasks.
- But they are not good at automatic speaker verification (ASV)?
We do not show the results of wav2vec 2.0 `large` and HuBERT `large` here because they do not perform well in round 1.

In round 1, we have not released the power of pre-trained models.
Game Start!

Round 2
Rules in Round 2

All the upstream models use the same downstream models.

Downstream Model 1

Downstream Model 2

Pre-trained Model

“How are you?”

Speaker 42

Keep it simple
e.g., Linear layer

e.g., 2-layer LSTM

ASR

Labelled data

SID

Labelled data

Last Layer Output

fixed
The feature weights are joined learned with the downstream task.
## Results of Round 2

<table>
<thead>
<tr>
<th></th>
<th>Content PR</th>
<th>Content KS</th>
<th>Content ASR</th>
<th>Content QbE</th>
<th>Content SID</th>
<th>Content ASV</th>
<th>Content SD</th>
<th>Semantic IC</th>
<th>Semantic SF</th>
<th>Semantic ER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>fbank</strong></td>
<td>82.01</td>
<td>8.63</td>
<td>15.21</td>
<td>0.0058</td>
<td>8.50E-04</td>
<td>9.56</td>
<td>10.05</td>
<td>9.1</td>
<td>69.64</td>
<td>35.39</td>
</tr>
<tr>
<td><strong>PASE+</strong></td>
<td>58.87</td>
<td>82.54</td>
<td>16.62</td>
<td>0.0072</td>
<td>37.99</td>
<td>11.61</td>
<td>8.68</td>
<td>29.82</td>
<td>62.14</td>
<td>57.86</td>
</tr>
<tr>
<td><strong>APC</strong></td>
<td>41.98</td>
<td>91.01</td>
<td>14.74</td>
<td>0.0310</td>
<td>60.42</td>
<td>8.56</td>
<td>10.53</td>
<td>74.69</td>
<td>70.46</td>
<td>59.33</td>
</tr>
<tr>
<td><strong>VQ-APC</strong></td>
<td>41.08</td>
<td>91.11</td>
<td>15.21</td>
<td>0.0251</td>
<td>60.15</td>
<td>8.72</td>
<td>10.45</td>
<td>74.48</td>
<td>68.53</td>
<td>59.66</td>
</tr>
<tr>
<td><strong>NPC</strong></td>
<td>43.81</td>
<td>88.96</td>
<td>13.91</td>
<td>0.0246</td>
<td>55.92</td>
<td>9.40</td>
<td>9.34</td>
<td>69.44</td>
<td>72.79</td>
<td>59.08</td>
</tr>
<tr>
<td><strong>Mockingjay</strong></td>
<td>70.19</td>
<td>83.67</td>
<td>15.48</td>
<td>6.60E-04</td>
<td>32.29</td>
<td>11.66</td>
<td>10.54</td>
<td>34.33</td>
<td>61.59</td>
<td>50.28</td>
</tr>
<tr>
<td><strong>TERA</strong></td>
<td>49.17</td>
<td>89.48</td>
<td>12.16</td>
<td>0.0013</td>
<td>57.57</td>
<td>15.89</td>
<td>9.96</td>
<td>58.42</td>
<td>67.50</td>
<td>56.27</td>
</tr>
<tr>
<td><strong>DeCoAR 2.0</strong></td>
<td>14.93</td>
<td>94.48</td>
<td>9.07</td>
<td>0.0406</td>
<td>74.42</td>
<td>7.16</td>
<td>6.59</td>
<td>90.80</td>
<td>83.28</td>
<td>62.47</td>
</tr>
<tr>
<td><strong>modified CPC</strong></td>
<td>42.54</td>
<td>91.88</td>
<td>13.53</td>
<td>0.0326</td>
<td>39.63</td>
<td>12.86</td>
<td>10.38</td>
<td>64.09</td>
<td>71.19</td>
<td>60.96</td>
</tr>
<tr>
<td><strong>wav2vec</strong></td>
<td>31.58</td>
<td>95.59</td>
<td>11.00</td>
<td>0.0485</td>
<td>56.56</td>
<td>7.99</td>
<td>9.90</td>
<td>84.92</td>
<td>76.37</td>
<td>59.79</td>
</tr>
<tr>
<td><strong>vq-wav2vec</strong></td>
<td>33.48</td>
<td>93.38</td>
<td>12.80</td>
<td>0.0410</td>
<td>38.80</td>
<td>10.38</td>
<td>9.93</td>
<td>85.68</td>
<td>77.68</td>
<td>58.24</td>
</tr>
<tr>
<td><strong>wav2vec 2.0 base</strong></td>
<td>5.74</td>
<td>96.23</td>
<td>4.79</td>
<td>0.0233</td>
<td>75.18</td>
<td>6.02</td>
<td>6.08</td>
<td>92.35</td>
<td>88.30</td>
<td>63.43</td>
</tr>
<tr>
<td><strong>wav2vec 2.0 large</strong></td>
<td>4.75</td>
<td>96.66</td>
<td>3.10</td>
<td>0.0489</td>
<td>86.14</td>
<td>5.65</td>
<td>5.62</td>
<td>95.28</td>
<td>87.11</td>
<td>65.64</td>
</tr>
<tr>
<td><strong>HuBERT base</strong></td>
<td>5.41</td>
<td>96.30</td>
<td>4.79</td>
<td>0.0736</td>
<td>81.42</td>
<td>5.11</td>
<td>5.88</td>
<td>98.34</td>
<td>88.53</td>
<td>64.92</td>
</tr>
<tr>
<td><strong>HuBERT large</strong></td>
<td>3.53</td>
<td>95.29</td>
<td>2.94</td>
<td>0.0353</td>
<td>90.33</td>
<td>5.98</td>
<td>5.75</td>
<td>98.76</td>
<td>89.81</td>
<td>67.62</td>
</tr>
</tbody>
</table>
### Results of Round 2

<table>
<thead>
<tr>
<th></th>
<th>Content</th>
<th>Speaker</th>
<th>Semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PR</td>
<td>KS</td>
<td>ASR</td>
</tr>
<tr>
<td>fbank</td>
<td>82.01</td>
<td>8.63</td>
<td>15.21</td>
</tr>
<tr>
<td>PASE+</td>
<td>58.87</td>
<td>82.54</td>
<td>8.63</td>
</tr>
<tr>
<td>APC</td>
<td>41.98</td>
<td>91.01</td>
<td>14.74</td>
</tr>
<tr>
<td>VQ-APC</td>
<td>41.08</td>
<td>91.11</td>
<td>15.21</td>
</tr>
<tr>
<td>NPC</td>
<td>43.81</td>
<td>88.96</td>
<td>13.91</td>
</tr>
<tr>
<td>Mockingjay</td>
<td>70.19</td>
<td>83.67</td>
<td>15.21</td>
</tr>
<tr>
<td>TERA</td>
<td>49.17</td>
<td>89.48</td>
<td>12.16</td>
</tr>
<tr>
<td>DeCoAR 2.0</td>
<td>14.93</td>
<td>94.48</td>
<td>9.07</td>
</tr>
<tr>
<td>modified CPC</td>
<td>42.54</td>
<td>91.88</td>
<td>13.53</td>
</tr>
<tr>
<td>wav2vec</td>
<td>31.58</td>
<td>95.59</td>
<td>11.00</td>
</tr>
<tr>
<td>vq-wav2vec</td>
<td>33.48</td>
<td>93.38</td>
<td>12.80</td>
</tr>
<tr>
<td>wav2vec 2.0 base</td>
<td>5.74</td>
<td>96.23</td>
<td>4.79</td>
</tr>
<tr>
<td>wav2vec 2.0 large</td>
<td>4.75</td>
<td>96.66</td>
<td>3.10</td>
</tr>
<tr>
<td>HuBERT base</td>
<td>5.41</td>
<td>96.30</td>
<td>4.79</td>
</tr>
<tr>
<td>HuBERT large</td>
<td>3.53</td>
<td>95.29</td>
<td>2.94</td>
</tr>
</tbody>
</table>

- Pre-trained learning outperforms fbank in most cases.
### Results of Round 2

<table>
<thead>
<tr>
<th></th>
<th>Content</th>
<th>Speaker</th>
<th>Semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IC</td>
<td>SF</td>
<td>ER</td>
</tr>
<tr>
<td>fbank</td>
<td>9.1</td>
<td>69.64</td>
<td>35.39</td>
</tr>
<tr>
<td>PASE+</td>
<td>29.82</td>
<td>72.79</td>
<td>59.08</td>
</tr>
<tr>
<td>APC</td>
<td>74.69</td>
<td>70.46</td>
<td>59.33</td>
</tr>
<tr>
<td>VQ-APC</td>
<td>74.48</td>
<td>70.66</td>
<td>62.76</td>
</tr>
<tr>
<td>NPC</td>
<td>69.44</td>
<td>72.79</td>
<td>59.08</td>
</tr>
<tr>
<td>Mockingjay</td>
<td>34.33</td>
<td>50.28</td>
<td></td>
</tr>
<tr>
<td>TERA</td>
<td>58.42</td>
<td>62.27</td>
<td></td>
</tr>
<tr>
<td>DeCoAR 2.0</td>
<td>90.80</td>
<td>83.28</td>
<td>62.47</td>
</tr>
<tr>
<td>modified CPC</td>
<td>64.09</td>
<td>71.19</td>
<td>60.96</td>
</tr>
<tr>
<td>wav2vec</td>
<td>84.92</td>
<td>76.37</td>
<td>59.79</td>
</tr>
<tr>
<td>vq-wav2vec</td>
<td>85.68</td>
<td>77.68</td>
<td>58.24</td>
</tr>
<tr>
<td>wav2vec 2.0 base</td>
<td>92.35</td>
<td>88.30</td>
<td>63.43</td>
</tr>
<tr>
<td>wav2vec 2.0 large</td>
<td>95.28</td>
<td>87.11</td>
<td>65.64</td>
</tr>
<tr>
<td>HuBERT base</td>
<td>98.34</td>
<td>88.53</td>
<td>64.92</td>
</tr>
<tr>
<td>HuBERT large</td>
<td>98.76</td>
<td>89.81</td>
<td>67.62</td>
</tr>
</tbody>
</table>

- Several pre-trained models are all-around.
Analysis of the Weights

The feature weights are joined learned with the downstream task.
Layer Weights
– Phoneme Recognition

(The weights are normalized by the representation’s norms.)
Layer Weights
- Speaker Verification

![Graph showing layer weights across different layers and models.](image)
Specialist? Universal?

Just name a few ...

They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?

They are universal!

.... but how can task-agnostic self-supervised learning achieve that?

(I don’t have the answer now.)
Welcome to Joint the Game

https://superbbbenchmark.org/
<table>
<thead>
<tr>
<th>Method</th>
<th>Name</th>
<th>Description</th>
<th>URL</th>
<th>Rank</th>
<th>Score</th>
<th>Rank-P</th>
<th>Score-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>WavLM Large</td>
<td>Microsoft</td>
<td>M-P + VQ ...</td>
<td>⬤</td>
<td>18.8</td>
<td>1122</td>
<td>6.1</td>
<td>3.54</td>
</tr>
<tr>
<td>WavLM Base+</td>
<td>Microsoft</td>
<td>M-P + VQ ...</td>
<td>⬤</td>
<td>17.7</td>
<td>1106</td>
<td>12.7</td>
<td>11.68</td>
</tr>
<tr>
<td>WavLM Base</td>
<td>Microsoft</td>
<td>M-P + VQ ...</td>
<td>⬤</td>
<td>16</td>
<td>1019</td>
<td>11.45</td>
<td>10.76</td>
</tr>
<tr>
<td>HuBERT Large</td>
<td>paper</td>
<td>M-P + VQ</td>
<td></td>
<td>15.1</td>
<td>919</td>
<td>4.1</td>
<td>2.9</td>
</tr>
<tr>
<td>wav2vec 2.0 Large</td>
<td>paper</td>
<td>M-C + VQ</td>
<td></td>
<td>14.8</td>
<td>914</td>
<td>3.9</td>
<td>2.88</td>
</tr>
<tr>
<td>wav2vec 2.0 Large</td>
<td>paper</td>
<td>M-P + VQ</td>
<td></td>
<td>14.45</td>
<td>941</td>
<td>10.25</td>
<td>9.94</td>
</tr>
<tr>
<td>FaST-VGS+</td>
<td>Puyuan P...</td>
<td>FaST-VG...</td>
<td>⬤</td>
<td>12.9</td>
<td>809</td>
<td>5.9</td>
<td>3.72</td>
</tr>
<tr>
<td>wav2vec 2.0 Base</td>
<td>paper</td>
<td>M-C + VQ</td>
<td></td>
<td>11.85</td>
<td>818</td>
<td>8.7</td>
<td>8.61</td>
</tr>
<tr>
<td>DistillHuBERT</td>
<td>Heng-Jui</td>
<td>multi-task ...</td>
<td></td>
<td>11.1</td>
<td>717</td>
<td>15.6</td>
<td>30.54</td>
</tr>
<tr>
<td>DeCoAR 2.0</td>
<td>paper</td>
<td>M-G + VQ</td>
<td></td>
<td>10.5</td>
<td>722</td>
<td>8.5</td>
<td>8.03</td>
</tr>
<tr>
<td>wav2vec</td>
<td>paper</td>
<td>F-C</td>
<td></td>
<td>8.9</td>
<td>529</td>
<td>12.55</td>
<td>16.25</td>
</tr>
</tbody>
</table>
Toolkit – S3PRL

https://github.com/s3prl/s3prl
Let’s welcome the era of Pre-training.
Research in Progress based on Self-supervised Learning
More ……

• 1. Make Pre-trained Model Smaller
• 2. Attacking Pre-trained Model
• 3. Privacy Issue of Pre-trained Model
• 4. Data Bias vs. Pre-training
• 5. Unsupervised Speech Recognition
• 6. Spoken Question Answering
1. Make Pre-trained Model Smaller

Larger models usually lead to better performance.
Typical Knowledge Distillation

Each layer contains different information. Learning from the last layer is not sufficient.
1. Make Upstream Model Smaller

DistilHuBERT is better than the models with the same size.

https://arxiv.org/abs/2110.01900
2. Adversarial Attack

For all tasks

Downstream Model 1
Downstream Model 2
Downstream Model 3

single point of failure?

Pre-trained Model

https://arxiv.org/abs/2111.04330
Catastrophic performance degradation

https://arxiv.org/abs/2111.04330

Adding non-perceivable noise as different as possible

Pre-trained Model

Downstream Model 1
Downstream Model 2
Downstream Model 3

Pre-trained Model

Task-agnostic
### 2. Adversarial Attack

The directions of the arrows denote the directions towards the better performance of the task.

<table>
<thead>
<tr>
<th></th>
<th>ASR</th>
<th>PR</th>
<th>KS</th>
<th>IC</th>
<th>SF</th>
<th>SID</th>
<th>ER</th>
<th>SD</th>
<th>ASV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>PER</td>
<td>Acc</td>
<td>Acc</td>
<td>F1</td>
<td>CER</td>
<td>Acc</td>
<td>Acc</td>
<td>Acc</td>
</tr>
<tr>
<td>(a)</td>
<td>36.66</td>
<td>41.99</td>
<td>61</td>
<td>52</td>
<td>88.62</td>
<td>18.47</td>
<td>77</td>
<td>75</td>
<td>88.2</td>
</tr>
<tr>
<td>(b)</td>
<td>8.73</td>
<td>6.74</td>
<td>94</td>
<td>83</td>
<td>95.54</td>
<td>9.31</td>
<td>89</td>
<td>93</td>
<td>95.06</td>
</tr>
<tr>
<td>(c)</td>
<td>0.54</td>
<td>0.96</td>
<td>97</td>
<td>95</td>
<td>99.13</td>
<td>1.61</td>
<td>95</td>
<td>97</td>
<td>98.2</td>
</tr>
<tr>
<td>(d)</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>(e)</td>
<td>58.79</td>
<td>40.59</td>
<td>64</td>
<td>61</td>
<td>73.94</td>
<td>36.75</td>
<td>69</td>
<td>74</td>
<td>87.53</td>
</tr>
<tr>
<td>(f)</td>
<td>2.50</td>
<td>3.04</td>
<td>97</td>
<td>98</td>
<td>98.63</td>
<td>2.22</td>
<td>89</td>
<td>91</td>
<td>95.02</td>
</tr>
<tr>
<td>(g)</td>
<td>0</td>
<td>0.41</td>
<td>99</td>
<td>99</td>
<td>98.81</td>
<td>1.47</td>
<td>94</td>
<td>100</td>
<td>98.18</td>
</tr>
<tr>
<td>(h)</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

w2v2 and HuBERT are self-supervised models.

**Without attack:** Only select the samples with the correct predictions (e.g., 0% WER for ASR, 0% PER for PR, etc.)
2. Adversarial Attack

The directions of the arrows denote the directions towards the better performance of the task.

<table>
<thead>
<tr>
<th></th>
<th>ASR WER ↓</th>
<th>PR PER ↓</th>
<th>KS Acc ↑</th>
<th>IC Acc ↑</th>
<th>SF F1 ↑</th>
<th>CER ↓</th>
<th>SID Acc ↑</th>
<th>ER Acc ↑</th>
<th>SD Acc ↑</th>
<th>ASV Acc ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) w2v2-w2v2</td>
<td>36.66</td>
<td>41.99</td>
<td>61</td>
<td>52</td>
<td>88.62</td>
<td>18.47</td>
<td>77</td>
<td>75</td>
<td>88.2</td>
<td>17.5</td>
</tr>
<tr>
<td>(b) HuBERT-w2v2</td>
<td>8.73</td>
<td>6.74</td>
<td>94</td>
<td>83</td>
<td>95.54</td>
<td>9.31</td>
<td>89</td>
<td>93</td>
<td>95.06</td>
<td>7.3</td>
</tr>
<tr>
<td>(c) gau-w2v2</td>
<td>0.54</td>
<td>0.96</td>
<td>97</td>
<td>95</td>
<td>99.13</td>
<td>1.61</td>
<td>95</td>
<td>97</td>
<td>98.2</td>
<td>2.6</td>
</tr>
<tr>
<td>(d) Clean-w2v2</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>(e) HuBERT-HuBERT</td>
<td>58.79</td>
<td>40.59</td>
<td>64</td>
<td>61</td>
<td>73.94</td>
<td>36.75</td>
<td>69</td>
<td>74</td>
<td>87.53</td>
<td>18.5</td>
</tr>
<tr>
<td>(f) w2v2-HuBERT</td>
<td>2.50</td>
<td>3.04</td>
<td>97</td>
<td>98</td>
<td>98.63</td>
<td>2.22</td>
<td>89</td>
<td>91</td>
<td>95.02</td>
<td>7.1</td>
</tr>
<tr>
<td>(g) gau-HuBERT</td>
<td>0</td>
<td>0.41</td>
<td>99</td>
<td>99</td>
<td>98.81</td>
<td>1.47</td>
<td>94</td>
<td>100</td>
<td>98.18</td>
<td>2.5</td>
</tr>
<tr>
<td>(h) Clean-HuBERT</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

w2v2 and HuBERT are self-supervised models.

**Without attack:** Only select the samples with the correct predictions (e.g., 0% WER for ASR, 0% PER for PR, etc.)

**Adding Gaussian noises:** Only a small impact on performance
2. Adversarial Attack

The directions of the arrows denote the directions towards the better performance of the task.

<table>
<thead>
<tr>
<th>Model</th>
<th>ASR WER ↓</th>
<th>PR PER ↓</th>
<th>KS Acc ↑</th>
<th>IC Acc ↑</th>
<th>SF F1 ↑</th>
<th>CER ↓</th>
<th>SID Acc ↑</th>
<th>ER Acc ↑</th>
<th>SD Acc ↑</th>
<th>DER ↓</th>
<th>ASV Acc ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) w2v2-w2v2</td>
<td>36.66</td>
<td>41.99</td>
<td>61</td>
<td>52</td>
<td>88.62</td>
<td>18.47</td>
<td>77</td>
<td>75</td>
<td>88.2</td>
<td>17.5</td>
<td>90</td>
</tr>
<tr>
<td>(b) HuBERT-w2v2</td>
<td>8.73</td>
<td>6.74</td>
<td>94</td>
<td>83</td>
<td>95.54</td>
<td>9.31</td>
<td>89</td>
<td>93</td>
<td>95.06</td>
<td>7.3</td>
<td>98</td>
</tr>
<tr>
<td>(c) gau-w2v2</td>
<td>0.54</td>
<td>0.96</td>
<td>97</td>
<td>95</td>
<td>99.13</td>
<td>1.61</td>
<td>95</td>
<td>97</td>
<td>98.2</td>
<td>2.6</td>
<td>100</td>
</tr>
<tr>
<td>(d) Clean-w2v2</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>(e) HuBERT-HuBERT</td>
<td>58.79</td>
<td>40.59</td>
<td>64</td>
<td>61</td>
<td>73.94</td>
<td>36.75</td>
<td>69</td>
<td>74</td>
<td>87.53</td>
<td>18.5</td>
<td>81</td>
</tr>
<tr>
<td>(f) w2v2-HuBERT</td>
<td>2.50</td>
<td>3.04</td>
<td>97</td>
<td>98</td>
<td>98.63</td>
<td>2.22</td>
<td>89</td>
<td>91</td>
<td>95.02</td>
<td>7.1</td>
<td>97</td>
</tr>
<tr>
<td>(g) gau-HuBERT</td>
<td>0</td>
<td>0.41</td>
<td>99</td>
<td>99</td>
<td>98.81</td>
<td>1.47</td>
<td>94</td>
<td>100</td>
<td>98.18</td>
<td>2.5</td>
<td>99</td>
</tr>
<tr>
<td>(h) Clean-HuBERT</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

**White-box attack**: the attack is very effective.

**Black-box attack**: not as effective as while-box attack
3. Privacy Issue

Unlabeled Speech

Customer has the right to delete them

Collected from real applications (e.g., from smart assistant)

Pre-train

Pre-trained Model

Recover?
3. Privacy Issue

- Membership Inference Attack

Given an utterance

Pre-trained Model

The input utterance is in training set or not?

With a native approach, we can get very high accuracy.

https://arxiv.org/abs/2111.05113
4. Would Biased Unlabeled Data become an Issue?

Don't speak too fast: The impact of data bias on self-supervised speech models

https://arxiv.org/abs/2110.07957
5. Unsupervised Speech Recognition

This can be achieved by Generative Adversarial Network (GAN).
**How is the results?**

- Unsupervised setting on TIMIT (text and audio are unpair, text is not the transcription of audio)
  - 63.6% PER (oracle boundaries) [Liu, et al., INTERSPEECH 2018]
  - 41.6% PER (automatic segmentation) [Yeh, et al., ICLR 2019]
  - 33.1% PER (automatic segmentation) [Chen, et al., INTERSPEECH 2019]

---

**The progress of supervised learning**

Unsupervised learning is as good as supervised learning 30 years ago.

---

Unsupervised ASR + Self-supervised Pre-training

Amount of labeled data used

Word error rate

https://ai.facebook.com/blog/wav2vec-unsupervised-speech-recognition-without-supervision/
5. Unsupervised Speech Recognition

Can they come from different domains?
5. Unsupervised Speech Recognition

https://arxiv.org/abs/2110.03509
6. Spoken Question Answering

- Spoken Question Answering (SQA)

Who is the U.S. president?

SQA

Without speech recognition

Joe Biden
6. Spoken Question Answering

Review

s = 2  e = 3

The answer is “\(d_2 d_3\)”. 

\[
\text{inner product} \quad \rightarrow 
\]

\[
\text{Softmax} 
\]

\[
\begin{array}{c}
0.1 \\
0.2 \\
0.7 \\
\end{array}
\]

\[
\text{Random Initialized} 
\]

\[
\begin{array}{c}
\text{[CLS]} \\
q_1 \\
q_2 \\
[SEP] \\
d_1 \\
d_2 \\
d_3 \\
\end{array}
\]

\[
\text{question} \quad \text{document} 
\]
6. Spoken Question Answering

How about spoken QA?

It does not work ..... 

No semantic information

Speech version of BERT

[CLS] [SEP] question [SEP] document
Pre-trained on text is required.

Speech version of BERT

[CLS]  [SEP]

question  document

Softmax

Ridiculous???
Recall these experiments ....

Why does BERT work?

No pre-train: **6.12** F1 score  Pre-training on text: **54.22** F1 score
More ......

• 1. Unsupervised Speech Recognition
• 2. Make Pre-trained Model Smaller
• 3. Attacking Pre-trained Model
• 4. Privacy Issue of Pre-trained Model
• 5. Data Bias vs. Pre-training
• 6. Spoken Question Answering