CNRS GDR Groupement de recherche
TAL Traitement automatique des langues

Nantes, September 2023

Self-supervised Learning approaches for Spoken Language Processing

Ha Nguyen

LABORATOIRE INFORMATIQUE D'AVIGNON

AVIGNON UNIVERSITÉ

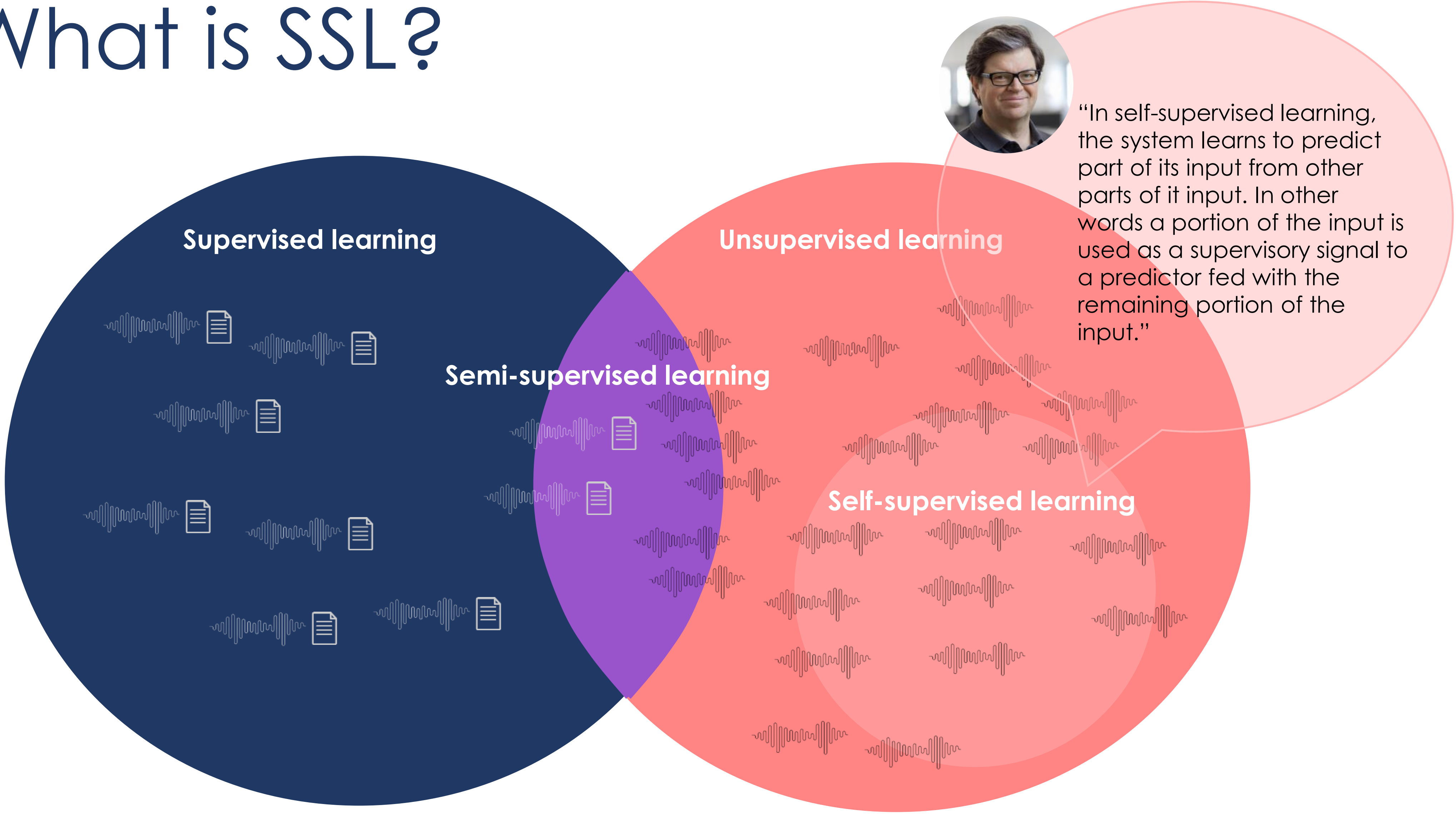
LeBenchmark

Outline

I. Introduction
II. SSL models
III. Wav2vec2.0
IV. LeBenchmark
V. Conclusion

Introduction

What is SSL?



Supervised learning
Unsupervised learning
Semi-supervised learning
Self-supervised learning
“In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input.”

What’s good about SSL?

Exploit the potential of unlabeled data

- Goal: capture implicit speech representation directly from speech
- Targets are computed from the signal itself
- Objective: pretext tasks

Pretraining process

↓

Use in downstream tasks?

- Feature extractor
- Fine-tuned

SSL models

SSL models

SpeechT5
Ao, Junyi, et al., 2022

data2vec
Baevski, Alexei, et al., 2022

BigSSL
Zhang, Yu, et al., 2022

UniSpeech-SAT
Chen, Sanyuan, et al. , 2022b

WavLM
Chen, Sanyuan, et al. , 2022a

XLS-R
Babu, Arun, et al., 2021

HuBERT
Hsu, Wei-Ning, et al., 2021

UniSpeech
Wang, Chengyi, et al., 2021

w2v-BERT
Chung, Yu-An, et al., 2021

w2v-Conformer
Zhang, Yu, et al., 2021

XLSR-53
Conneau, Alexis, et al., 2020

wav2vec 2.0
Baevski, Alexei, et al., 2020

Mockingjay
Liu, Andy T., et al., 2020

vq-wav2vec
Baevski, Alexei, et al., 2019

wav2vec
Schneider, Steffen, et al., 2019

2019 2020 2021 2022

SSL models

Convolutional Neural Networks (CNN)
Transformer (Vaswani, Ashish, et al., 2017)
Conformer (Gulati, Anmol, et al., 2020)

wav2vec
Schneider, Steffen, et al., 2019

vq-wav2vec
Baevski, Alexei, et al., 2019

Mockingjay
Liu, Andy T., et al., 2020

wav2vec 2.0
Baevski, Alexei, et al., 2020

XLSR-53
Conneau, Alexis, et al., 2020

w2v-Conformer
Zhang, Yu, et al., 2021

w2v-BERT
Chung, Yu-An, et al., 2021

UniSpeech
Wang, Chengyi, et al., 2021

HuBERT
Hsu, Wei-Ning, et al., 2021

XLS-R
Babu, Arun, et al., 2021

WavLM
Chen, Sanyuan, et al. , 2022a

UniSpeech-SAT
Chen, Sanyuan, et al. , 2022b

BigSSL
Zhang, Yu, et al., 2022

data2vec
Baevski, Alexei, et al., 2022

SpeechT5
Ao, Junyi, et al., 2022

2019 2020 2021 2022

SSL models

Contrastive predictive coding (Oord, Aaron van den, et al., 2018)
Mask Language Modeling objective
Multi-task objectives
Regression

SpeechT5
Ao, Junyi, et al., 2022
data2vec
Baevski, Alexei, et al., 2022
BigSSL
Zhang, Yu, et al., 2022
UniSpeech-SAT
Chen, Sanyuan, et al. , 2022b
WavLM
Chen, Sanyuan, et al. , 2022a
XLS-R
Babu, Arun, et al., 2021
HuBERT
Hsu, Wei-Ning, et al., 2021
UniSpeech
Wang, Chengyi, et al., 2021
w2v-BERT
Chung, Yu-An, et al., 2021
w2v-Conformer
Zhang, Yu, et al., 2021
XLSR-53
Conneau, Alexis, et al., 2020
wav2vec 2.0
Baevski, Alexei, et al., 2020
Mockingjay
Liu, Andy T., et al., 2020
vq-wav2vec
Baevski, Alexei, et al., 2019
wav2vec
Schneider, Steffen, et al., 2019

2019 2020 2021 2022

SSL models

Hugging Face

02/09/2023

Models 5,835 wav2vec

new Full-text search

Sort: Trending

- jonatasgrosman/wav2vec2-large-xlsr-53-english
  Automatic Speech Recognition • Updated Mar 25 • 61.1M • 261
- facebook/wav2vec2-base-960h
  Automatic Speech Recognition • Updated Nov 14, 2022 • 587k • 146
- audeering/wav2vec2-large-robust-12-ft-emotion-msp-d...
  Audio Classification • Updated 5 days ago • 133k • 37
- hafidikhsan/Wav2vec2-large-robust-Pronounciation-Ev...
  Audio Classification • Updated Jun 26 • 39 • 3
- AndrewMcDowell/wav2vec2-xls-r-1b-arabic
  Automatic Speech Recognition • Updated Feb 1, 2022 • 16 • 1
- Arnold/wav2vec2-large-xlsr-hausa2-demo-colab
  Automatic Speech Recognition • Updated Feb 15, 2022 • 7 • 2
- Harveenchadha/vakyansh-wav2vec2-hindi-him-4200
  Automatic Speech Recognition • Updated Jan 29, 2022 • 1.62k • 1
- facebook/wav2vec2-large-robust
  Updated Nov 5, 2021 • 2.87k • 16
- facebook/wav2vec2-xls-r-2b
  Updated Aug 10, 2022 • 686 • 17
- facebook/wav2vec2-xls-r-300m
  Updated Aug 10, 2022 • 15.1k • 42
- kresnik/wav2vec2-large-xlsr-korean
  Automatic Speech Recognition • Updated Jul 3 • 4.59k • 22
- speechbrain/emotion-recognition-wav2vec2-IEMOCAP
  Audio Classification • Updated Jul 23 • 42.5k • 49
- qinyue/wav2vec2-large-xlsr-53-chinese-zn-cn-aishell1
  Automatic Speech Recognition • Updated Aug 3, 2022 • 9 • 7
- wbbbbb/wav2vec2-large-chinese-zh-cn
  Automatic Speech Recognition • Updated Jul 18, 2022 • 5.81k • 26
- speechbrain/asr-wav2vec2-commonvoice-14-zh-CN
  Automatic Speech Recognition • Updated 18 days ago • 7 • 1
- Umong/wav2vec2-large-mms-1b-bengali
  Automatic Speech Recognition • Updated 5 days ago • 258 • 1

How to evaluate SSL models???

Based solely on the performance of the downstream tasks
There are too many models, how to compare them???

SUPERB

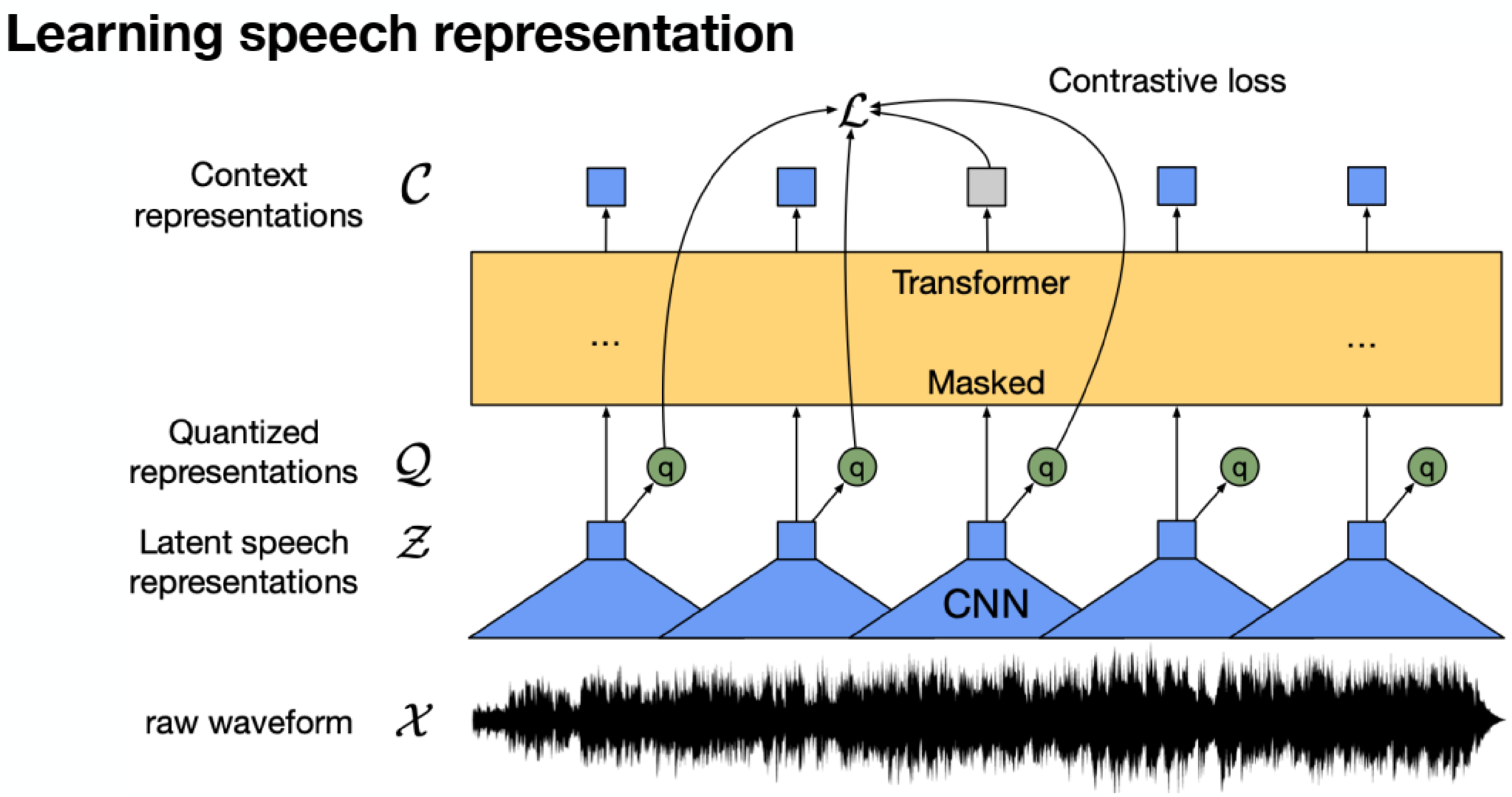
Speech processing **U**niversal **PER**formance **B**enchmark

LeBenchmark

Wav2vec 2.0

Wav2vec 2.0

Baevski, Alexei, et al., 2020



Wav2vec 2.0

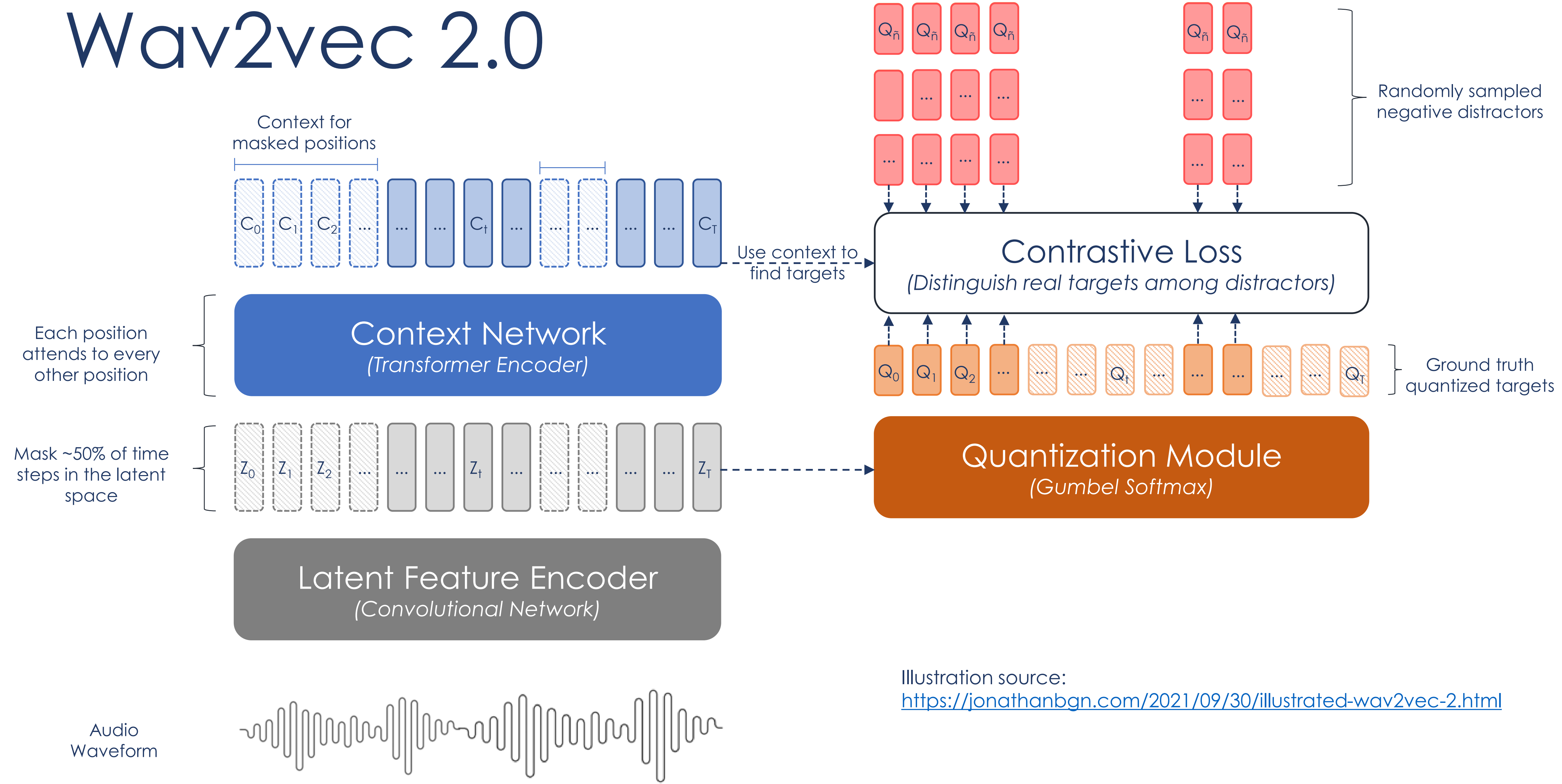
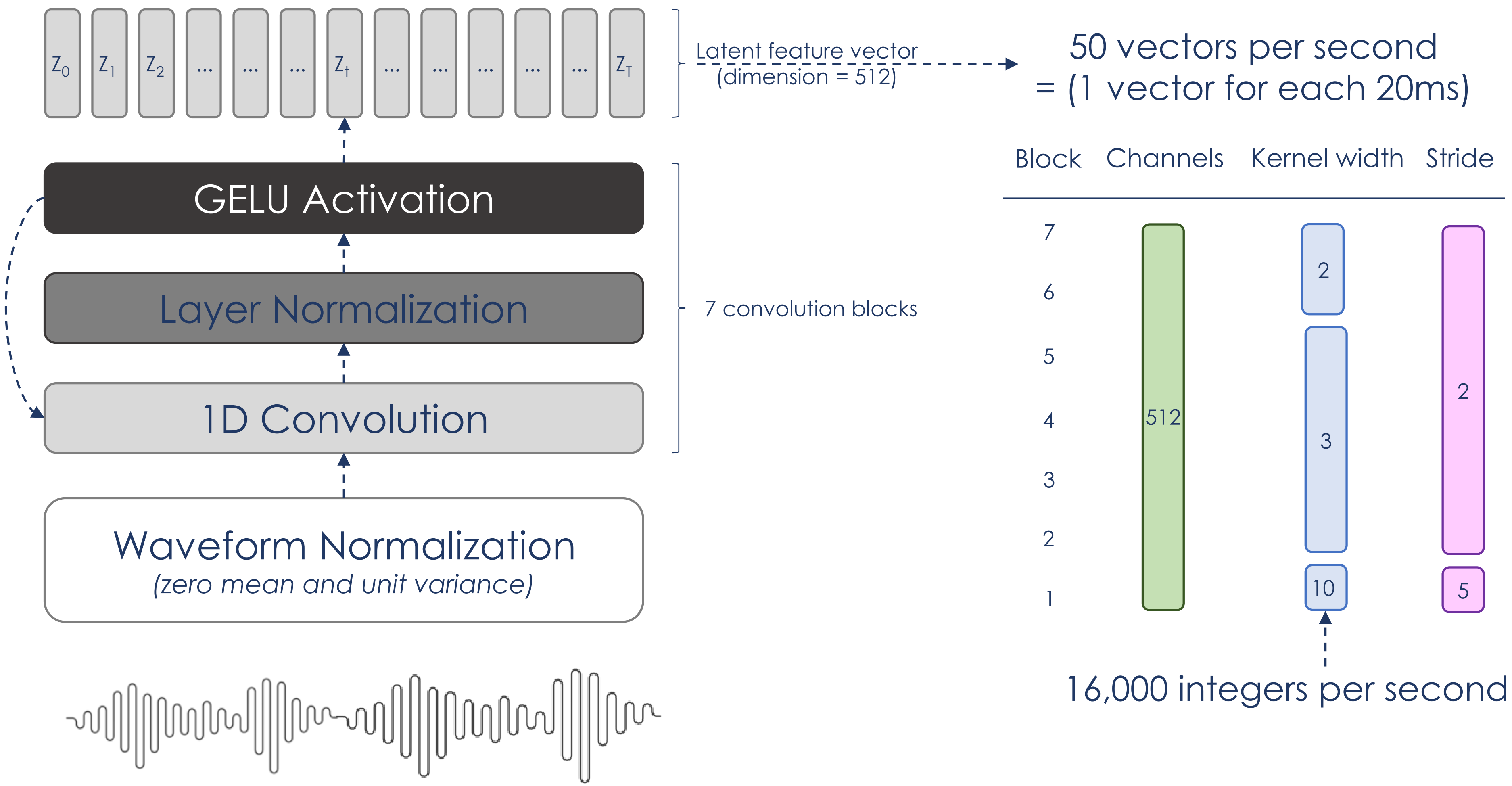


Illustration source:
https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html

Wav2vec 2.0

Latent feature encoder

Illustration source:
https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html



Wav2vec2.0

Quantization module

- Wav2vec 2.0 aims to learn discrete representations (discrete speech units)

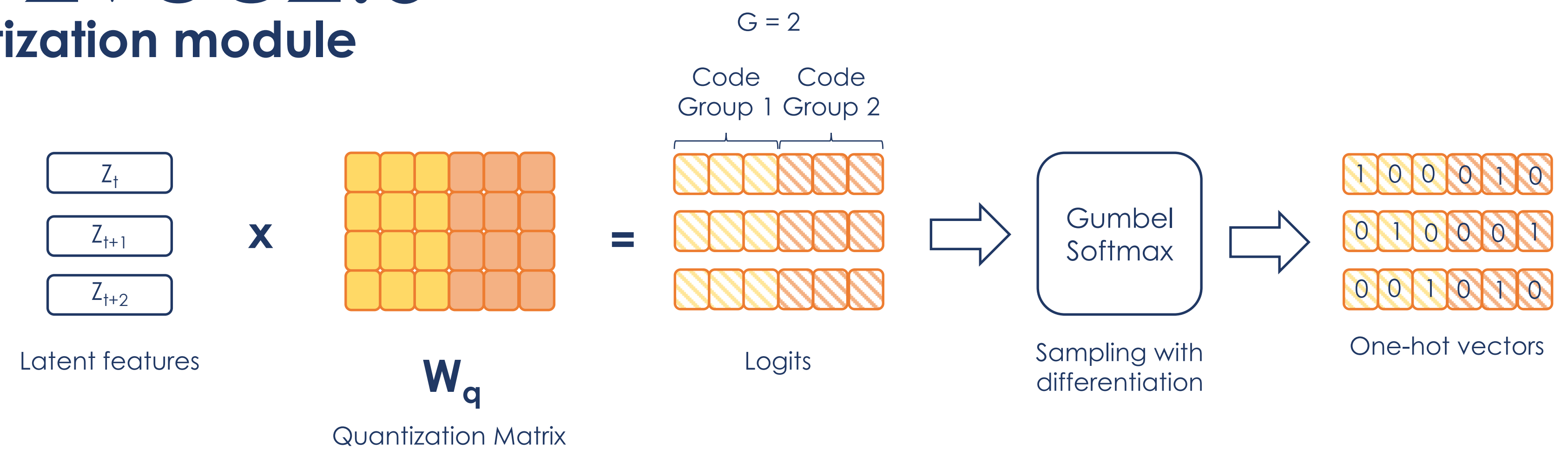
  Latent feature vectors ⇨ Discrete values
- wav2vec model captures some information that could be related to phones (subphones, triphones…)
- Two codebooks with 320 possible words in each group are concatenated:

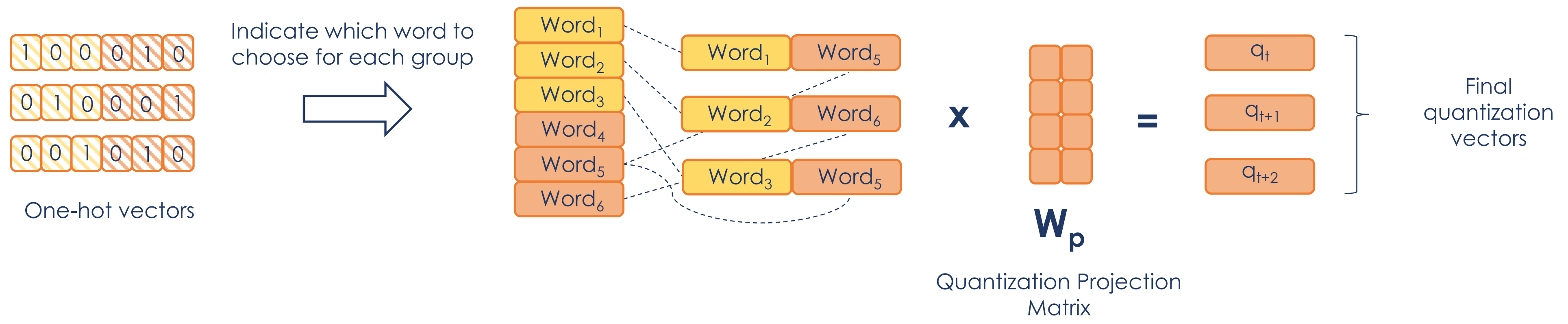
  320 x 320 = **102,400 speech units**
- Automatically learnt by product quantization and Gumbel Softmax

Illustration source:
https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html

Wav2vec2.0

Quantization module





Wav2vec 2.0

Context Network (Transformer)

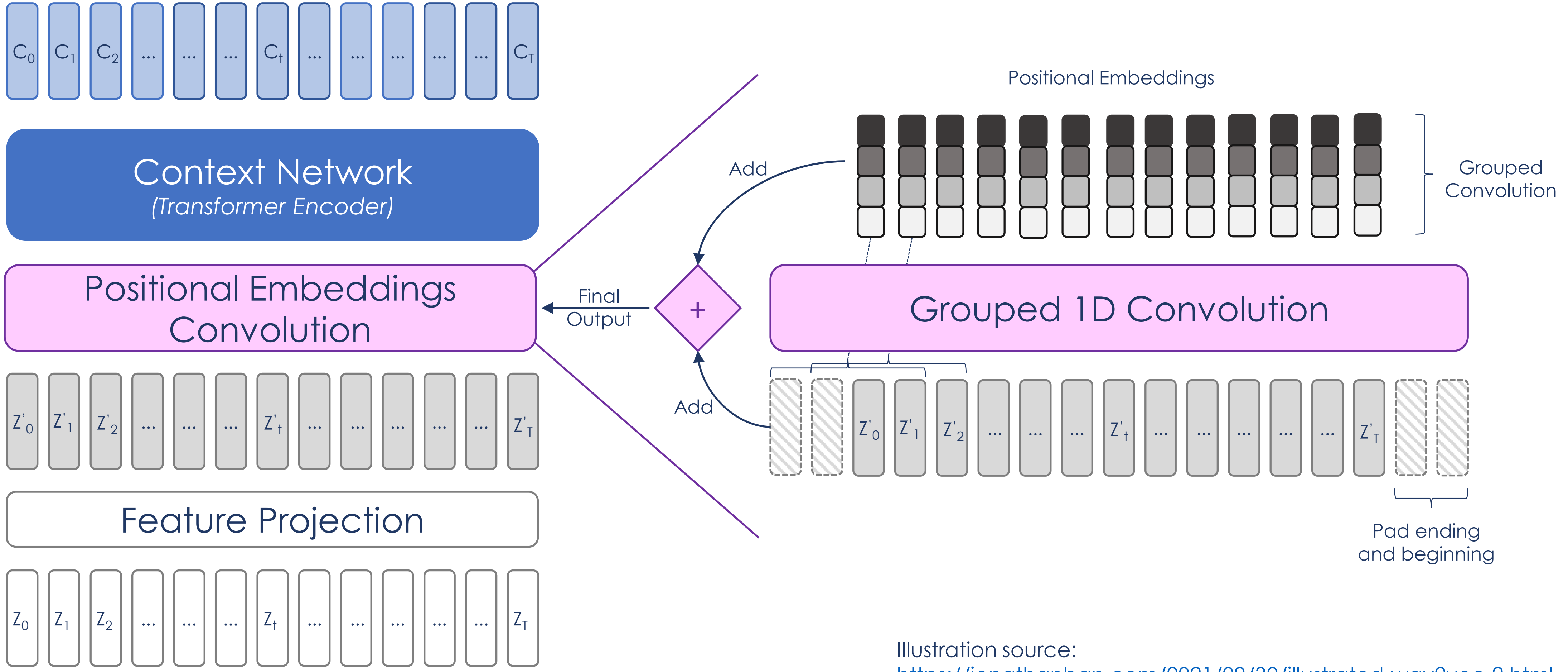


Illustration source:
https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html

Wav2vec 2.0

Contrastive Loss

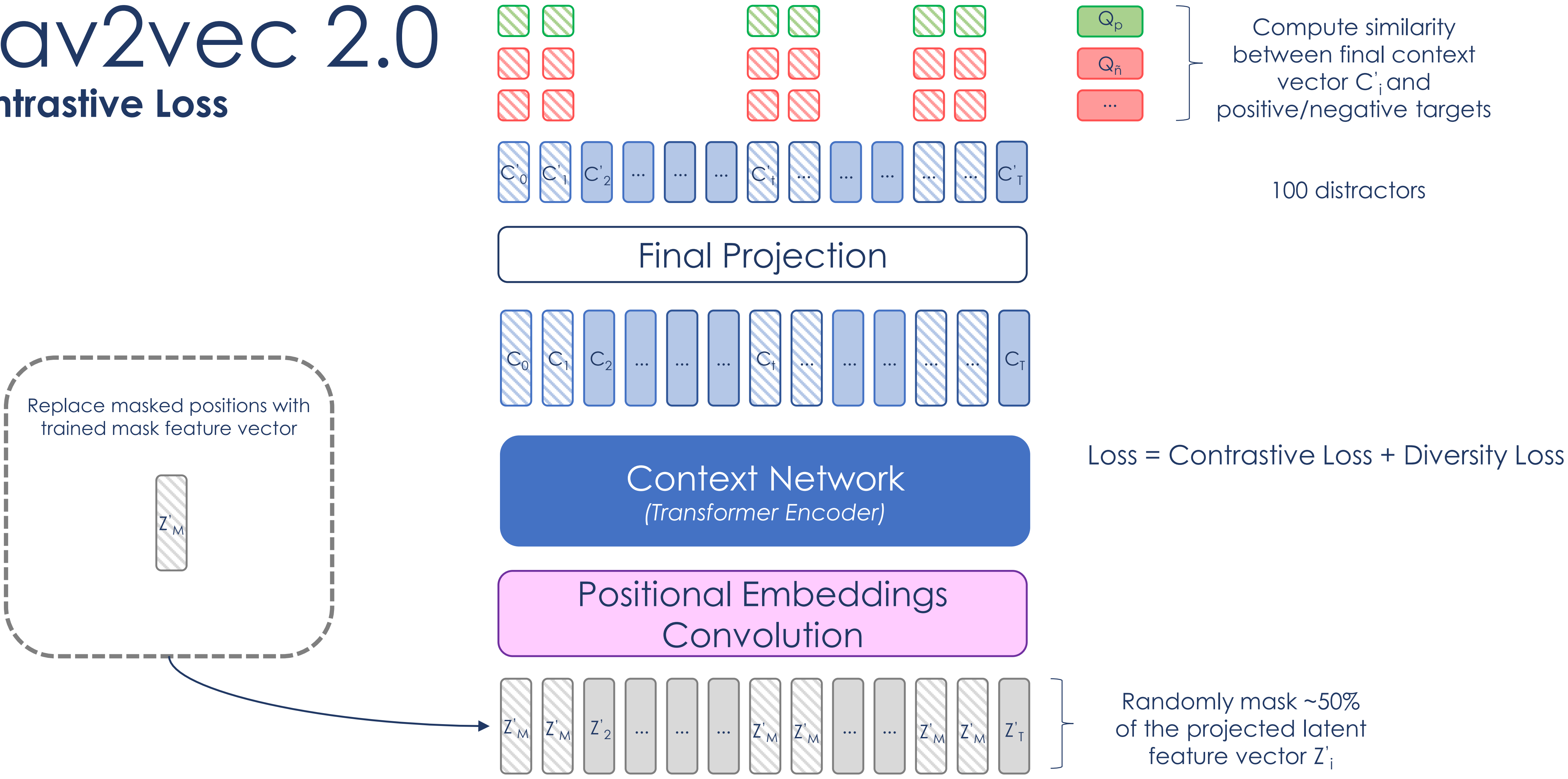


Illustration source:
https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html

LeBenchmark

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech

Solène Evain[1,*], *Ha Nguyen*[1,2,*], *Hang Le*[1,*], *Marcely Zanon Boito*[1,*], *Salima Mdhaffar*[2,*], *Sina Alisamir*[1,3], *Ziyi Tong*[1], *Natalia Tomashenko*[2], *Marco Dinarelli*[1,*], *Titouan Parcollet*[2,*], *Alexandre Allauzen*[4], *Yannick Estève*[2], *Benjamin Lecouteux*[1], *François Portet*[1], *Solange Rossato*[1], *Fabien Ringeval*[1], *Didier Schwab*[1] *and Laurent Besacier*[1,5]

[1]Univ. Grenoble Alpes, CNRS, Inria, G-INP, LIG, France
[2]LIA, Avignon Université, France
[3]Atos, Échirolles, France
[4]ESPCI, CNRS LAMSADE, PSL Research University, France
[5]Naver Labs Europe, France

yannick.esteve@univ-avignon.fr, francois.portet@univ-grenoble-alpes.fr

Task Agnostic and Task Specific Self-Supervised Learning from Speech with *LeBenchmark*

Solène Evain[1,*], **Ha Nguyen**[1,2,*], **Hang Le**[1,*], **Marcely Zanon Boito**[1,2,*], **Salima Mdhaffar**[2,*], **Sina Alisamir**[1,3,*], **Ziyi Tong**[1], **Natalia Tomashenko**[2,*], **Marco Dinarelli**[1,*], **Titouan Parcollet**[2,*], **Alexandre Allauzen**[4], **Yannick Estève**[2], **Benjamin Lecouteux**[1], **François Portet**[1], **Solange Rossato**[1], **Fabien Ringeval**[1], **Didier Schwab**[1], and **Laurent Besacier**[5]

[1]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France
[2]LIA, Avignon Université, France
[3]Atos, Échirolles, France
[4]ESPCI, CNRS LAMSADE, PSL Research University, France
[5]Naver Labs Europe, France
[*]Equal contributors

LeBechmark 2.0

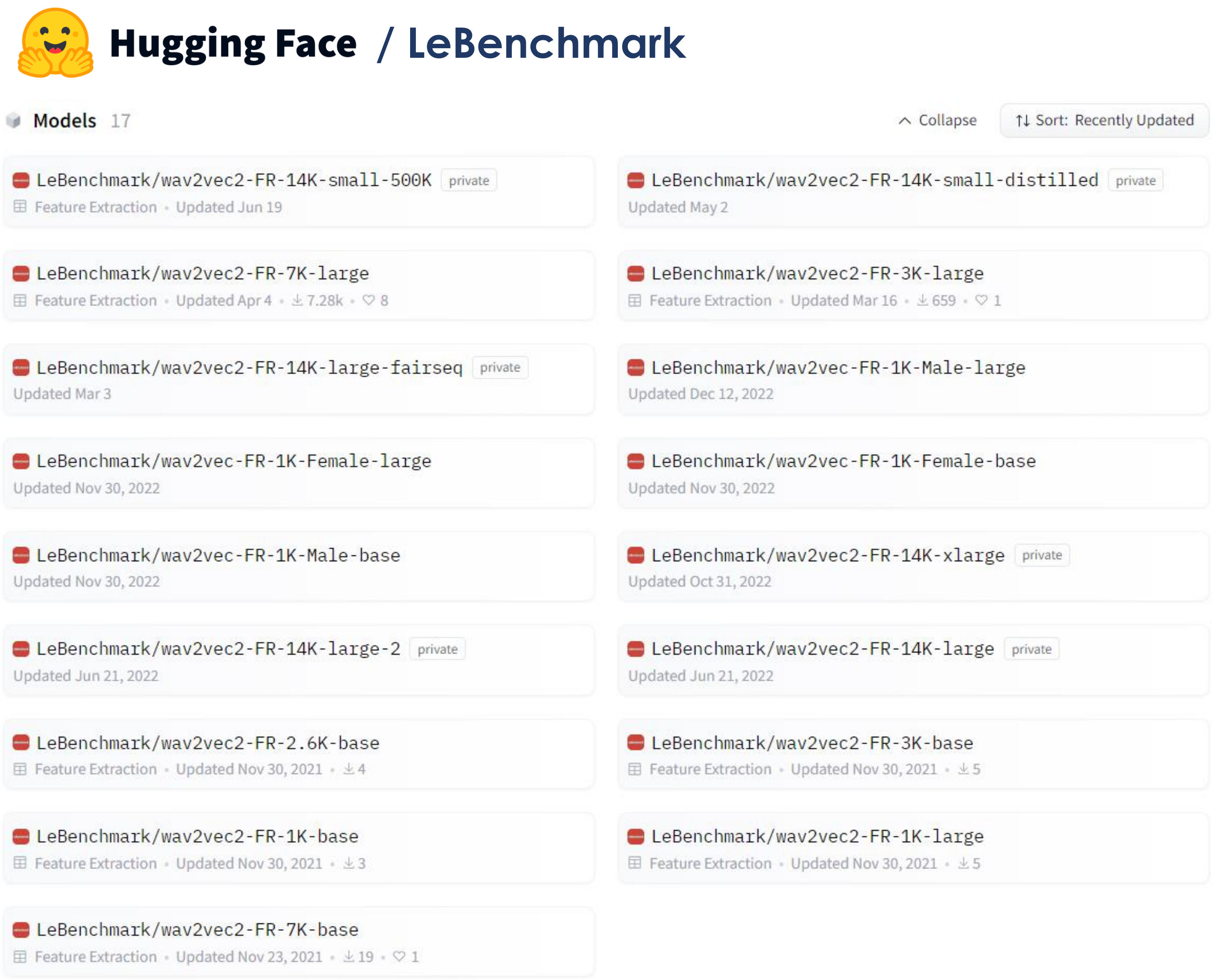
- More pre-training data
- More pre-trained SSL models
- More downstream tasks

LeBenchmark

Open-source and **reproducible** framework for assessing SSL from **French** speech data:

- Data collection
- SSL Pre-training
- Evaluation on downstream tasks

https://github.com/LeBenchmark/



Data collection

Different datasets:

- 1K hours
- 3K hours
- 7K hours
- **14K hours**

No.	Corpus$_{License}$	#Utterances	Duration	#Speakers	Mean Utt. Duration	Speech type
		Small dataset – 1K				
1	MLS French$_{CCBY4.0}$	**263,055** 124,590 / 138,465 / –	**1,096:43** 520:13 / 576:29 / –	**178** 80 / 98 / –	**15 s** 15 s / 15 s / –	Read
		Medium-clean dataset – 2.7K				
2	EPAC$_{NC}$	**623,250** 465,859 / 157,391 / –	**1,626:02** 1,240:10 / 385:52 / –	**Unk** – / – / –	**9 s** – / – / –	Radio Broadcasts
	2.7k dataset total	**886,305** 590,449 / 295,856 / -	**2,722:45** 1,760:23 / 962:21 / -	–	–	–
		Medium dataset – 3K				
3	African Accented French$_{Apache2.0}$	**16,402** 373 / 102 / 15,927	**18:56** – / – / 18:56	**232** 48 / 36 / 148	**4 s** – / – / –	Read
4	Att-Hack$_{CCBYNCND}$	**36,339** 16,564 / 19,775 / –	**27:02** 12:07 / 14:54 / –	**20** 9 / 11 / –	**2.7 s** 2.6 s / 2.7 s / –	Acted Emotional
5	CaFE$_{CCNC}$	**936** 468 / 468 / –	**1:09** 0:32 / 0:36 / –	**12** 6 / 6 / –	**4.4 s** 4.2 s / 4.7 s / –	Acted Emotional
6	CFPP2000$_{CCBYNCS\ A}$	**9853** 166 / 1,184 / 8,503	**16:26** 0:14 / 1:56 / 14:16	**49** 2 / 4 / 43	**6 s** 5 s / 5 s / 6 s	Spontaneous
7	ESLO2$_{NC}$	**62,918** 30,440 / 32,147 / 331	**34:12** 17:06 / 16:57 / 0:09	**190** 68 / 120 / 2	**1.9 s** 2 s / 1.9 s / 1.7 s	Spontaneous
8	GEMEP$_{NC}$	**1,236** 616 / 620 / –	**0:50** 0:24 / 0:26 / –	**10** 5 / 5 / –	**2.5 s** 2.4 s / 2.5 s / –	Acted Emotional
9	MPF	**19,527** 5,326 / 4,649 / 9,552	**19:06** 5:26 / 4:36 / 9:03	**114** 36 / 29 / 49	**3.5 s** 3.7 s / 3.6 s / 3.4 s	Spontaneous
10	PORTMEDIA$_{NC}$ (French)	**19,627** 9,294 / 10,333 / –	**38:59** 19:08 / 19:50 / –	**193** 84 / 109 / –	**7.1 s** 7.4 s / 6.9 s / –	Acted telephone dialogue
11	TCOF$_{CCBYNCSA}$ (Adults)	**58,722** 10,377 / 14,763 / 33,582	**53:59** 9:33 / 12:39 / 31:46	**749** 119 / 162 / 468	**3.3 s** 3.3 s / 3.1 s / 3.4 s	Spontaneous
	Medium dataset total	**1,111,865** 664,073 / 379,897 / 67,895	**2,933:24** 1,824:53 / 1,034:15 / 74:10	–	–	–
		Large dataset – 7K				
12	MaSS	**8,219** 8,219 / – / –	**19:40** 19:40 / – / –	**Unk** – / – / –	**8.6 s** 8.6 s / – / –	Read
13	NCCFr$_{NC}$	**29,421** 14,570 / 13,922 / 929	**26:35** 12:44 / 12:59 / 00:50	**46** 24 / 21 / 1	**3 s** 3 s / 3 s / 3 s	Spontaneous
14	Voxpopuli$_{CC0}$ *Unlabeled*	**568,338** – / – / –	**4,532:17** – / – / 4,532:17	**Unk** – / – / –	**29 s** – / – / –	Professional speech
15	Voxpopuli$_{CC0}$ transcribed	**76.281** – / – / –	**211:57** – / – / 211:57	**327** – / – / –	**10 s** – / – / –	Professional speech
	Large dataset total	**1,814,242** 682,322 / 388,217 / 99,084	**7,739:22** 1,853:02 / 1,041:07 / 4,845:07	–	–	–
		Extra Large dataset – 14K				
16	Audiocite.net$_{CC-BY}$	**817,295** 425 033 / 159 691 / 232 571	**6,698:35** 3477:24 / 1309:49 / 1911:21	**130** 35 / 32 / 63	**29 s** 29 s / 29 s / 29 s	Read
17	Niger-Mali Audio Collection CCBYNCND	**38,332** 18 546 / 19 786 / –	**111:01** 52:15 / 58:46 / –	**357** 192 / 165 / –	**10 s** 10 s / 10 s / –	Radio Broadcasts
	Extra Large dataset tota	**2,669,869** 1 125 901 / 567 694 / 331 655	**14,548:58** 5 382:41 / 2 409:42 / 6 756:28	–	–	–

LeBenchmark's SSL models

No.	Model	Pre-training data	Parameters count	Output Dimension	Updates	GPU Count	GPU Hours
1	1K-*base*	1,096 h	90M	768	200K	4	1,000
2	1K-*large*	1,096 h	330M	1,024	200K	32	3,700
3	2.7K-*base*	2,773 h	90M	768	500K	32	4,100
4	3K-*base*	2,933 h	90M	768	500K	32	4,100
5	3K-*large*	2,933 h	330M	1,024	500K	32	10,900
6	7K-*base*	7,739 h	90M	768	500K	64	7,900
7	7K-*large*	7,739 h	330M	1,024	500K	64	13,500
				LeBenchmark 2.0			
8	14K-*light*	14,000 h	26M	512	500K	32	5,000
9	14K-*large*	14,000 h	330M	1,024	1M	64	28,800
10	14K-*xlarge*	14,000 h	965M	1,280	1M	104	54,600

Downstream tasks

- Automatic Speech Recognition (ASR)
- Spoken Language Understanding (SLU)
- Automatic Speech-to-text Translation (AST)
- Automatic Emotion Recognition (AER)
- Automatic Speaker Verification (ASV)
- Syntactic Analysis (SA)

How to use SSL models???

- Feature extractor
  - Task agnostic pre-training
  - Task specific pre-training:
    - Fine-tuned in a SSL manner on the downstream task’s data
    - Fine-tuned on different tasks, for example, fine-tuning on ASR
- Fine-tuning with the downstream task’s model

  Normally using SSL models as speech encoders

Automatic Speech Translation Task

- mTEDx data xx-fr:
  - en-fr: 50 hours
  - es-fr: 38 hours
  - pt-fr: 25 hours
- En-base, En-large: English SSL models
- XLSR-53 multiligual SSL models of 53 languages
- MFB: filterbank features
- Scores are BLEU
- Task agnostic pre-training: using pre-trained SSL models off-the-shelf for extracting speech features for the AST task

No.	Features	Valid			Test		
		en	es	pt	en	es	pt
1	MFB	1.5±0.17	0.67±0.15	0.61±0.13	1.10±0.14	0.87±0.12	0.32±0.03
(a) Task agnostic pre-training							
2	En-base	5.54±0.27	1.30±0.17	0.54±0.11	5.20±0.28	1.47±0.15	0.38±0.05
3	En-large	4.11±0.25	1.67±0.20	0.32±0.03	3.56±0.22	2.29±0.18	0.43±0.05
4	1K-base	9.18±0.36	5.09±0.27	0.39±0.05	8.98±0.36	5.64±0.30	0.49±0.08
5	1K-large	15.31±0.46	13.74±0.43	8.29±0.34	14.46±0.46	14.77±0.46	9.37±0.38
6	2.7-base	15.09±0.49	13.27±0.43	4.72±0.27	14.69±0.48	14.04±0.43	5.51±0.28
7	3K-base	15.05±0.49	13.19±0.44	4.44±0.29	14.80±0.47	14.27±0.44	4.72±0.25
8	3K-large	17.94±0.51	16.40±0.49	8.64±0.34	18.00±0.51	18.12±0.48	9.55±0.36
9	7K-base	15.13±0.45	12.78±0.40	2.65±0.20	14.50±0.45	13.61±0.44	2.66±0.23
10	7K-large	19.23±0.54	17.59±0.49	9.68±0.37	19.04±0.53	18.24±0.49	10.98±0.41
11	XLSR-53-large	7.81±0.33	0.49±0.13	0.43±0.07	6.75±0.29	0.52±0.08	0.36±0.05

Automatic Speech Translation Task

- Task specific SSL pre-training: continue pre-training SSL models on the AST speech data then using them for extracting speech features for the AST task

No.	Features	Valid			Test		
		en	es	pt	en	es	pt
1	MFB	1.5±0.17	0.67±0.15	0.61±0.13	1.10±0.14	0.87±0.12	0.32±0.03
(a) Task agnostic pre-training							
2	En-base	5.54±0.27	1.30±0.17	0.54±0.11	5.20±0.28	1.47±0.15	0.38±0.05
3	En-large	4.11±0.25	1.67±0.20	0.32±0.03	3.56±0.22	2.29±0.18	0.43±0.05
4	1K-base	9.18±0.36	5.09±0.27	0.39±0.05	8.98±0.36	5.64±0.30	0.49±0.08
5	1K-large	15.31±0.46	13.74±0.43	8.29±0.34	14.46±0.46	14.77±0.46	9.37±0.38
6	2.7-base	15.09±0.49	13.27±0.43	4.72±0.27	14.69±0.48	14.04±0.43	5.51±0.28
7	3K-base	15.05±0.49	13.19±0.44	4.44±0.29	14.80±0.47	14.27±0.44	4.72±0.25
8	3K-large	17.94±0.51	16.40±0.49	8.64±0.34	18.00±0.51	18.12±0.48	9.55±0.36
9	7K-base	15.13±0.45	12.78±0.40	2.65±0.20	14.50±0.45	13.61±0.44	2.66±0.23
10	7K-large	19.23±0.54	17.59±0.49	9.68±0.37	19.04±0.53	18.24±0.49	10.98±0.41
11	XLSR-53-large	7.81±0.33	0.49±0.13	0.43±0.07	6.75±0.29	0.52±0.08	0.36±0.05
(b) Task specific pre-training (SSL pre-training on mTEDx data)							
12	3K-large	18.54±0.53	16.40±0.48	8.81±0.36	18.38±0.52	17.84±0.48	10.57±0.41
13	7K-large	19.65±0.55	17.53±0.47	9.35±0.36	19.36±0.54	18.95±0.53	10.94±0.38
14	XLSR-53-large	6.83±0.33	0.54±0.14	0.34±0.03	6.75±0.32	0.34±0.03	0.29±0.03

Automatic Speech Translation Task

- Task specific supervised pre-training: fine-tune pre-trained SSL models on the ASR task then using them for extracting speech features for the AST task

No.	Features	Valid			Test		
		en	es	pt	en	es	pt
1	MFB	1.5±0.17	0.67±0.15	0.61±0.13	1.10±0.14	0.87±0.12	0.32±0.03
(a) Task agnostic pre-training							
2	En-base	5.54±0.27	1.30±0.17	0.54±0.11	5.20±0.28	1.47±0.15	0.38±0.05
3	En-large	4.11±0.25	1.67±0.20	0.32±0.03	3.56±0.22	2.29±0.18	0.43±0.05
4	1K-base	9.18±0.36	5.09±0.27	0.39±0.05	8.98±0.36	5.64±0.30	0.49±0.08
5	1K-large	15.31±0.46	13.74±0.43	8.29±0.34	14.46±0.46	14.77±0.46	9.37±0.38
6	2.7-base	15.09±0.49	13.27±0.43	4.72±0.27	14.69±0.48	14.04±0.43	5.51±0.28
7	3K-base	15.05±0.49	13.19±0.44	4.44±0.29	14.80±0.47	14.27±0.44	4.72±0.25
8	3K-large	17.94±0.51	16.40±0.49	8.64±0.34	18.00±0.51	18.12±0.48	9.55±0.36
9	7K-base	15.13±0.45	12.78±0.40	2.65±0.20	14.50±0.45	13.61±0.44	2.66±0.23
10	7K-large	19.23±0.54	17.59±0.49	9.68±0.37	19.04±0.53	18.24±0.49	10.98±0.41
11	XLSR-53-large	7.81±0.33	0.49±0.13	0.43±0.07	6.75±0.29	0.52±0.08	0.36±0.05
(b) Task specific pre-training (SSL pre-training on mTEDx data)							
12	3K-large	18.54±0.53	16.40±0.48	8.81±0.36	18.38±0.52	17.84±0.48	10.57±0.41
13	7K-large	19.65±0.55	17.53±0.47	9.35±0.36	19.36±0.54	18.95±0.53	10.94±0.38
14	XLSR-53-large	6.83±0.33	0.54±0.14	0.34±0.03	6.75±0.32	0.34±0.03	0.29±0.03
(c) Task specific pre-training (fine-tuned for ASR on mTEDx data)							
15	3K-large	21.09±0.53	19.28±0.53	14.40±0.47	21.34±0.58	21.18±0.52	16.66±0.49
16	7K-large	21.41±0.51	20.32±0.49	15.14±0.48	21.69±0.58	21.57±0.52	17.43±0.52
17	XLSR-53-large	21.09±0.54	20.38±0.56	14.56±0.45	20.68±0.53	21.14±0.55	17.21±0.54

Automatic Speech Translation Task

- Task specific fine-tuning: fine-tune the pre-trained SSL models directly on the AST task

No.	Features	Valid			Test		
		en	es	pt	en	es	pt
(b) Task specific pre-training (SSL pre-training on mTEDx data)							
12	3K-large	18.54±0.53	16.40±0.48	8.81±0.36	18.38±0.52	17.84±0.48	10.57±0.41
13	7K-large	19.65±0.55	17.53±0.47	9.35±0.36	19.36±0.54	18.95±0.53	10.94±0.38
14	XLSR-53-large	6.83±0.33	0.54±0.14	0.34±0.03	6.75±0.32	0.34±0.03	0.29±0.03
(c) Task specific pre-training (fine-tuned for ASR on mTEDx data)							
15	3K-large	21.09±0.53	19.28±0.53	14.40±0.47	21.34±0.58	21.18±0.52	16.66±0.49
16	7K-large	21.41±0.51	20.32±0.49	15.14±0.48	21.69±0.58	21.57±0.52	17.43±0.52
17	XLSR-53-large	21.09±0.54	20.38±0.56	14.56±0.45	20.68±0.53	21.14±0.55	17.21±0.54
(d) Task specific fine-tuning directly on mTEDx data							
15	3K-large	17.6±0.51	15.1±0.45	8.6±0.34	16.9±0.47	15.6±0.46	9.7±0.37
16	7K-large	20.1±0.52	17.4±0.52	10.7±0.37	19.0±0.57	18.8±0.49	12.0±0.41
17	XLSR-53-large	15.6±0.49	15.6±0.45	8.4±0.31	12.5±0.47	15.8±0.44	9.1±0.36

What else can we do with SSL models???

- Leveraging text data to improve SSL speech models:

  Text data is more abundantly available ⇨ pre-trained text-based models have been long developed

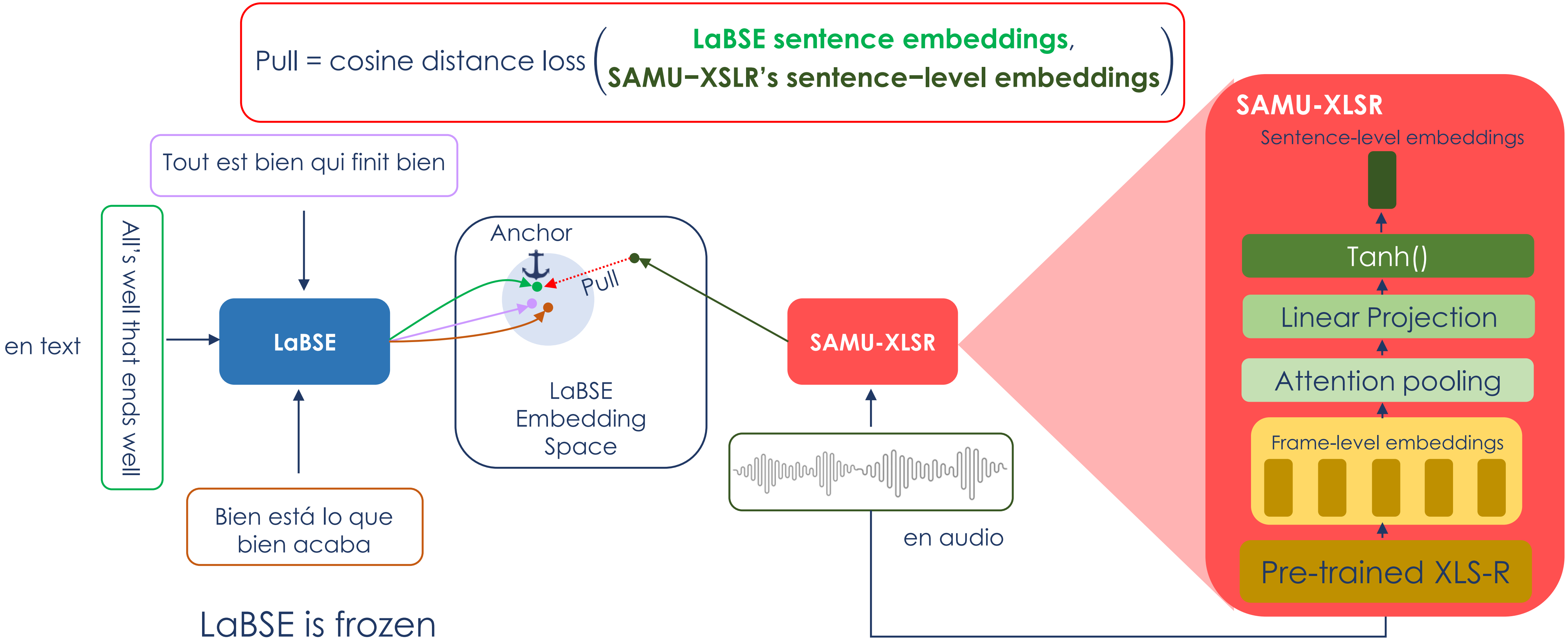
  - Project speech representations and text representations on the same embedding spaces
  - Using different objectives to “pull” these two types of embeddings together (L2, cosine similarity, etc.) (Han, Chi, et al., 2021, Agrawal, Bhuvan, et al., 2022, Khurana, Sameer, et al., 2022)

- SAMU-XLSR: **S**emantically-**A**ligned **M**ultimodal **U**tterance-level Cross-Lingual Speech Representation (Khurana, Sameer, et al., 2022)

SAMU-XLSR

Language-agnostic **B**ERT **S**entence **E**mbedding (Feng, Fangxiaoyu, et al., 2020)

XLS-R (Babu, Arun, et al., 2021)



LaBSE is frozen

SAMU-XLSR for AST

Language pair: Tamasheq – French (low-resource pair with very little AST labelled data (14 hours))
Baseline end-to-end AST model: wav2vec2.0 speech encoder + transformer decoder

🤗/ LIA-AvignonUniversity/IWSLT2022-tamasheq-only

facebook/mbart-large-50-many-to-many-mmt (Tang, Yuqing, et al., 2020)

No.	Model	dev	test
1	IWSLT2022-tamasheq-only + Transformer decoder	7.63	5.83
2	IWSLT2022-tamasheq-only + mBART decoder	9.46	7.4
3	SAMU-IWSLT2022-tamasheq-only + mBART decoder	12.6	9.7
4	SAMU-XLSR(53) + mBART decoder	12.5	7.9
5	SAMU-XLSR(60) + mBART decoder	19.1	14.2
6	SAMU-XLSR(100) + mBART decoder	19.3	13.5
7	SAMU-XLSR(100) + mBART decoder (IWLST23 best setup)	**21.4**	**16.5**

SAMU-XLSR

And there are more!!!

- Speech-to-text/speech translation retrieval
- Large-scale speech-text/speech data mining to create parallel speech-text/speech translation datasets
- etc.

Conclusion

Conclusion

- SSL is a very promising approach for Speech Processing
- If helps improve the performance of a wide range of downstream tasks
- Especially useful for low-resource languages
- It might take some resources to pre-train a SSL model
- But the model can be reused in so many tasks in so many different ways.

Thank you for listening!

References

Oord, Aaron van den, et al., 2018 "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).

Schneider, Steffen, et al, 2019 "wav2vec: Unsupervised pre-training for speech recognition." arXiv preprint arXiv:1904.05862 (2019).

Baevski, Alexei, et al., 2019 "vq-wav2vec: Self-supervised learning of discrete speech representations." arXiv preprint arXiv:1910.05453 (2019).

Liu, Andy T., et al., 2020 "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

Baevski, Alexei, et al., 2020 "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.

Conneau, Alexis, et al., 2020 "Unsupervised cross-lingual representation learning for speech recognition." arXiv preprint arXiv:2006.13979 (2020).

Zhang, Yu, et al., 2020 "Pushing the limits of semi-supervised learning for automatic speech recognition." arXiv preprint arXiv:2010.10504 (2020).

Chung, Yu-An, et al., 2021 "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training." 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021.

Wang, Chengyi, et al., 2021 "Unispeech: Unified speech representation learning with labeled and unlabeled data." International Conference on Machine Learning. PMLR, 2021.

References (2)

Hsu, Wei-Ning, et al., 2021 "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 3451-3460.

Babu, Arun, et al., 2021 "XLS-R: Self-supervised cross-lingual speech representation learning at scale." arXiv preprint arXiv:2111.09296 (2021).

Chen, Sanyuan, et al., 2022a "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1505-1518.

Chen, Sanyuan, et al., 2022b "Unispeech-sat: Universal speech representation learning with speaker aware pre-training." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

Zhang, Yu, et al., 2022 "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1519-1532.

Baevski, Alexei, et al., 2022 "Data2vec: A general framework for self-supervised learning in speech, vision and language." International Conference on Machine Learning. PMLR, 2022.

Ao, Junyi, et al., 2022 "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing." arXiv preprint arXiv:2110.07205 (2021).

Gulati, Anmol, et al., 2020 "Conformer: Convolution-augmented transformer for speech recognition." arXiv preprint arXiv:2005.08100 (2020).

Vaswani, Ashish, et al., 2017 "Attention is all you need." Advances in neural information processing systems 30 (2017).

References (3)

Khurana, Sameer, et al., 2022 "SAMU-XLSR: Semantically-aligned multimodal utterance-level cross-lingual speech representation." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1493-1504.

Feng, Fangxiaoyu, et al., 2020 "Language-agnostic BERT sentence embedding." arXiv preprint arXiv:2007.01852 (2020).

Tang, Yuqing, et al., 2020 "Multilingual translation with extensible multilingual pretraining and finetuning." arXiv preprint arXiv:2008.00401 (2020).

Agrawal, Bhuvan, et al., 2022 "Tie your embeddings down: Cross-modal latent spaces for end-to-end spoken language understanding." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

Han, Chi, et al., 2021 "Learning shared semantic space for speech-to-text translation." arXiv preprint arXiv:2105.03095 (2021).

LeBenchmark's SSL models

Estimates of the energy in kilowatt hour (kwH) and CO_2 equivalent in kilogram produced by the training of the *LeBenchmark 2.0* models.

No.	Model	Pre-training time *(hours)*	GPUs	Energy *(kWh)*	CO_2 *(kg)*
1	1K-*base*	250 h	4 Tesla V100	195.0	10.5
2	1K-*large*	925 h	4 Tesla V100	721.5	37.5
3	2.7K-*base*	128 h	32 Tesla V100	682.2	35.4
4	3K-*base*	128 h	32 Tesla V100	682.2	35.4
5	3K-*large*	341 h	32 Tesla V100	1,817.5	94.5
6	7K-*base*	123 h	64 Tesla V100	1,535.0	79.8
7	7K-*large*	211 h	64 Tesla V100	4,501.0	234
			LeBenchmark 2.0		
8	14K-*light*	156 h	32 Tesla V100	1,497.6	77.8
9	14K-*large*	436 h	64 Tesla V100	8,371.2	435
10	14K-*xlarge*	525 h	104 Tesla A100	16,511.2	859