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

Solene 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 Esteve[2], Benjamin Lecouteux[1], François Portet[1], Solange Rossato[1], Fabien Ringeval[1], Didier Schwab[1] and Laurent Besacier[1,5]

[1] LIG-Grenoble, [2] LIA-Avignon, [3] Atos-Echirolles, [4] ESPCI-Paris,
[5] Naver Labs Europe-Grenole

Outline

- Speech representations (for affect modeling)
- Self-supervised learning for speech
- LeBenchmark: A reproducible framework for SSL from speech
- Conclusion

Speech Representations (for Affect Modeling)

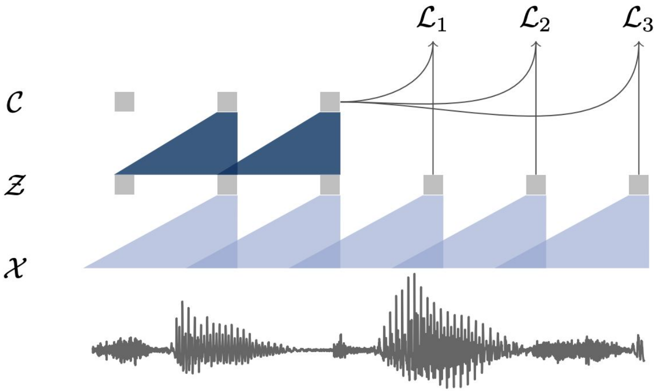
- Compact set describing human perception of sounds (e.g., log Mel filterbanks)
- Extension with long-term suprasegmental descriptors (e.g., prosody, voice quality)
- Distributional representations with Bags of Audio Words and Fisher Vectors (gradients of the log-likelihood of the data w.r.t. GMM’s parameters)
- Data-driven feature extraction with learnable convolutional filter banks (CNNs)
- Exploit knowledge from computer vision (ImageNet) to describe spectrograms (Deep Spectrum)
- Self-supervised learning: representations are learnt while resolving an unsupervised task
    - Do not require labels and can explore a large amount of data
    - Speech: predict occluded parts of a sentence
    - Vision: make representations invariant to augmentations

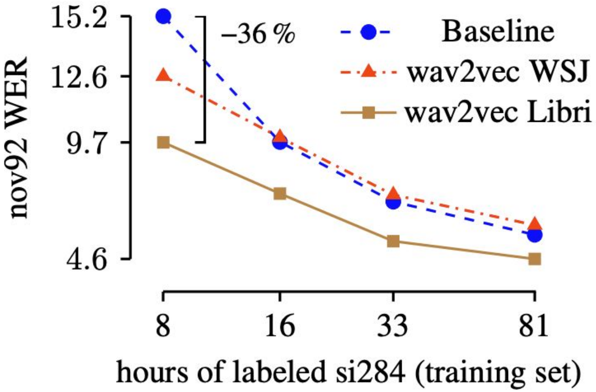
Self-Supervised Learning for Speech: wav2vec

- Learn latent speech audio representations with Contrastive Predicting Coding
  - Encode speech signal with two stacked CNNs
  - Predict whether future frames are real or distractors
  - Simplified loss (binary cross entropy)
  - Improved performance in ASR tasks

WAV2VEC: UNSUPERVISED PRE-TRAINING FOR SPEECH RECOGNITION

Steffen Schneider, Alexei Baevski, Ronan Collobert, Michael Auli
Facebook AI Research



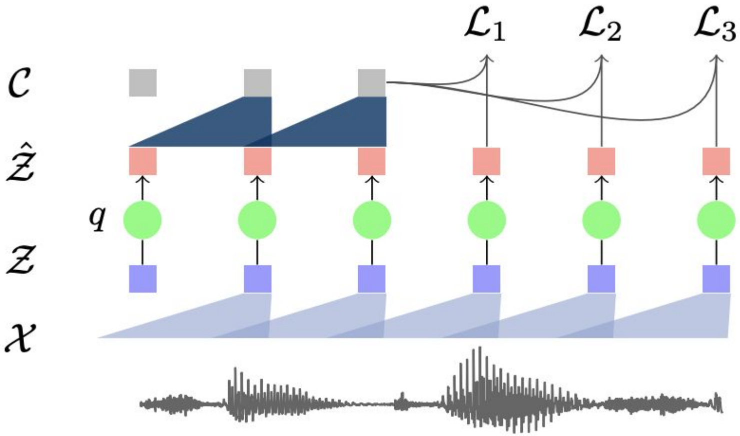


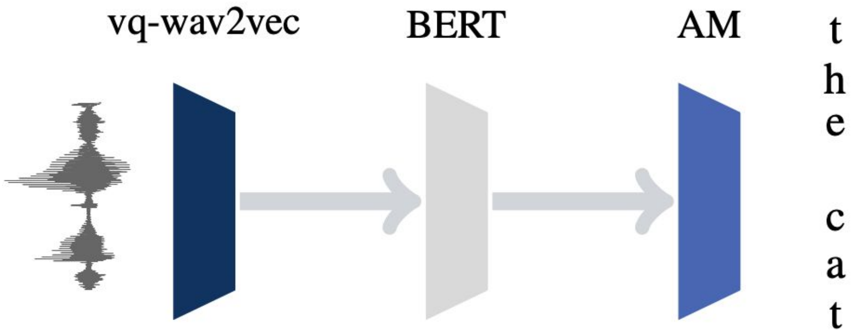
Self-Supervised Learning for Speech: vq-wav2vec

- Learn discrete latent speech representations with CPC
  - Identify an inventory of latent discrete speech representations with Vector Quantisation
  - Context representations learnt on top of speech units
  - VQ enables build NLP models with Seq2Seq
  - Vq-wav2vec: context in latent space prediction
  - Vq-vae: context in data reconstruction

VQ-WAV2VEC: SELF-SUPERVISED LEARNING OF DISCRETE SPEECH REPRESENTATIONS

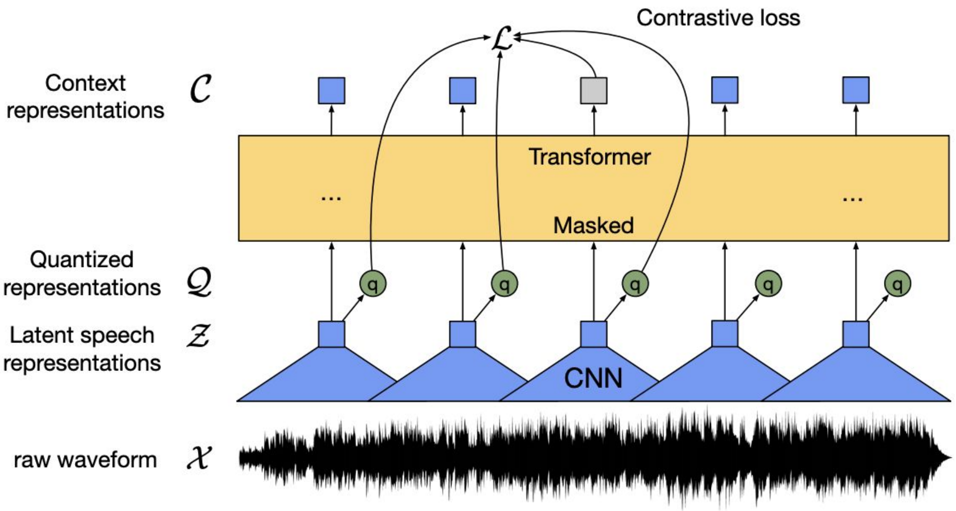
Alexei Baevski*△ Steffen Schneider*▽† Michael Auli△
△ Facebook AI Research, Menlo Park, CA, USA
▽ University of Tübingen, Germany





Self-Supervised Learning for Speech: wav2vec 2.0

- Jointly learn an inventory of speech units and a context representation with Transformer
  - Encode the raw waveform with a CNN (25 ms speech audio)
  - Transformer builds a representation for the entire sequence
  - Masked prediction task performed on discrete vocabulary of speech (Gumbel softmax VQ)



wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Alexei Baevski Henry Zhou Abdelrahman Mohamed Michael Auli

Cosine similarity

Context representation

Discrete latent speech representation

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

Negative samples

Temperature

LeBenchmark: A reproducible framework for SSL

- Motivations
  - SSL enables exploring huge unlabeled data for both NLP and image processing
  - Pioneering work successfully improved performance on downstream tasks (ASR)
  - Lack of common benchmarks and language-specific models

- What we did
  - Gathered a large and heterogeneous collection of French utterances (read, spontaneous)
  - Trained SSL models on collections of 1k and 3k hours of French speech
  - Assessed performance on French language with several tasks using Jean Zay cluster
    - Speech Recognition (ASR)
    - Spoken Language Understanding (SLU)
    - Speech Translation (AST)
    - Emotion Recognition (AER)

LeBenchmark: A reproducible framework for SSL

Table 1: *Statistics for the speech corpora used to train SSL models according to gender information (male / female / unknown). The small dataset (1k hours) is from MLS only, and the medium dataset (2.9k hours) is from all of them; duration: hour(s):minute(s).*

Corpus	# Utterances	Duration	# Speakers	Mean Utt. Duration	Speech type
African Accented French [8]	**16,402** 373 / 102 / 15,927	**18:56** – / – / 18:56	**232** 48 / 36 / 148	**4 s** – / – / –	Read
Att-Hack [9]	**36,339** 16,564 / 19,775 / 0	**27:02** 12:07 / 14:54 / 0:00	**20** 9 / 11 / 0	**2.7 s** 2.6 s / 2.7 s / –	Acted Emotional
CaFE [10]	**936** 468 / 468 / 0	**1:09** 0:32 / 0:36 / 0:00	**12** 6 / 6 / 0	**4.4 s** 4.2 s / 4.7 s / –	Acted Emotional
CFPP2000* [11] [12]	**12,574** 203 / 1,686 / 10,685	**20:20** 0:16 / 2:35 / 17:28	**50** 2 / 4 / 44	**5.8 s** 4.9 s / 5.5 s / 5.9 s	Spontaneous
ESLO2 [13], [14]	**62,918** 30,440 / 32,147 / 331	**34:12** 17:06 / 16:57 / 0:09	**190** 68 / 120 / 2	**1.9 s** 2.0 s / 1.9 s / 1.7 s	Spontaneous
EPAC** [15]	**623,250** 465,859 / 157,391 / 0	**1,626:02** 1,240:10 / 385:52 / 0:00	**1,935** – / – / –	**9 s** – / – / –	Radio Broadcasts
GEMEP [16]	**1,236** 616 / 620 / 0	**0:50** 0:24 / 0:26 / 0:00	**10** 5 / 5 / 0	**2.5 s** 2.4 s / 2.5 s / –	Acted Emotional
MLS French [17]	**263055** 124,590 / 138,465 / 0	**1,096:43** 520:13 / 576:29 / 0:00	**178** 80 / 98 / 0	**15.0 s** 15.0 s / 15.0 s / –	Read
MPF [18], [19]	**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
PORTMEDIA (French) [20]	**19,627** 9,294 / 10,333 / 0	**38:59** 19:08 / 19:50 / 0:00	**193** 84 / 109 / 0	**7.1 s** 7.4 s / 6.9 s / –	Acted telephone dialogue
TCOF (Adult s) [21]	**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
ALL	**1,114,586** 664,110 / 380,399 / 70,077	**2,937:18** 1,824:42 / 1034:54 / 77:22	-	-	-

*version without the CEFC corpus v2.1, 02/2021; **speakers are not uniquely identified.

LeBenchmark: A reproducible framework for SSL

- Automatic Speech Recognition
  - Datasets: Common Voice (477h), ETAPE (36h), EPAC (17.5k vocabulary)
  - Systems
    - Hybrid DNN-HMM: TDNN-F, 2 tri-gram LMs
    - End-to-end: SpeechBrain toolkit (encoder/decoder with attention)

Table 2: *ASR results (WER,%) on the ETAPE corpus for hybrid DNN-HMM acoustic models with TDNN-F topology.*

Language Model	**ETAPE**		**ESTER-1.2 + EPAC**	
Features	**Dev**	**Test**	**Dev**	**Test**
hires MFCC	39.28	40.89	35.60	37.73
W2V2-Fr-M-*large*	**32.19**	**33.87**	**28.53**	**30.77**
W2V2-En-*large*	39.93	42.30	36.18	38.75
XLSR-53-*large*	36.36	38.19	32.81	35.17

Table 3: *End-to-end ASR results (WER,%) on Common Voice and ETAPE corpora. (∗) means the training algorithm did not converge to a WER smaller than 100%.*

Corpus	**CommonVoice**		**ETAPE**	
Features	**Dev**	**Test**	**Dev**	**Test**
MFB	**20.19**	**23.40**	**54.55**	**56.17**
W2V2-Fr-M-*large*	20.23	24.06	55.56	57.04
W2V2-En-*large*	34.07	37.29	98.79	99.10
XLSR-53-*large*	30.07	32.72	(∗)	(∗)

LeBenchmark: A reproducible framework for SSL

- Spoken Language Understanding
  - Dataset: MEDIA corpus (56h)
  - System: end-to-end model with a pyramidal LSTM encoder (Fairseq)

Token decoding (Word Error Rate %)

[39] Seq	spectrogram	**29.42**	**28.71**
Kheops⊕Basic	spectrogram	36.25	37.12
Kheops⊕LSTM	spectrogram	**35.37**	**35.98**
Kheops⊕Basic	W2V2-En-*base*	19.80	21.78
Kheops⊕Basic	W2V2-En-*large*	24.44	26.96
Kheops⊕Basic	W2V2-Fr-S-*base*	23.11	25.22
Kheops⊕Basic	W2V2-Fr-S *large*	18.48	19.92
Kheops⊕Basic	W2V2-Fr-M-*base*	14.97	16.37
Kheops⊕Basic	W2V2-Fr-M *large*	**11.77**	**12.85**
Kheops⊕Basic	XLSR-53-*large*	14.98	15.74

SLU decoding (Concept Error Rate %)

[39] Seq	spectrogram	28.11	27.52
[39] XT	spectrogram	**23.39**	**24.02**
Kheops⊕Basic	spectrogram	39.66	40.76
Kheops⊕Basic +token	spectrogram	34.38	34.74
Kheops⊕LSTM +SLU	spectrogram	**33.63**	**34.76**
Kheops⊕LSTM	W2V2-En-*base*	26.31	26.11
Kheops⊕LSTM	W2V2-En-*large*	28.38	28.57
Kheops⊕LSTM	W2V2-Fr-S-*base*	26.16	26.69
Kheops⊕LSTM	W2V2-Fr-S *large*	22.53	23.03
Kheops⊕LSTM	W2V2-Fr-M-*base*	22.56	22.24
Kheops⊕LSTM	W2V2-Fr-M-*large*	**18.54**	**18.62**
Kheops⊕LSTM	XLSR-53-*large*	20.34	19.73

LeBenchmark: A reproducible framework for SSL

- Speech-to-text Translation
  - French as source language in two multilingual corpora (CoVoST-2, TEDx)
  - Target languages: English (TEDx: 50h, CoVoST2: 180h), Spanish (38h), Portuguese (25h)
  - System: Transformer (Fairseq S2T toolkit); block of linear-ReLU used before CNNs

Table 5: *BLEU on dev/valid and test sets of CoVoST-2 (CV2) and multilingual TEDx (mTEDx).*

Input features	**Dev/Valid data**				**Test data**			
	CV2	**mTEDx**			**CV2**	**mTEDx**		
	en	**en**	**es**	**pt**	**en**	**en**	**es**	**pt**
MFB	**23.37**	1.14	0.84	0.49	**22.66**	1.33	0.98	0.68
W2V2-En-*base*	19.24	0.90	0.65	0.43	18.19	0.88	0.34	0.27
W2V2-En-*large*	17.07	0.75	0.61	0.45	16.45	0.85	0.67	0.32
W2V2-Fr-S-*base*	19.86	2.64	0.49	0.50	19.04	1.66	0.67	0.61
W2V2-Fr-S-*large*	19.62	5.12	4.62	**2.06**	18.61	2.97	3.19	**2.25**
W2V2-Fr-M-*base*	19.47	6.98	1.87	0.63	18.32	6.37	1.99	0.54
W2V2-Fr-M-*large*	20.17	**9.35**	**7.72**	1.58	19.35	**6.76**	**6.63**	1.63
W2V2-Fr-VP-*base*	18.44	0.81	0.45	0.56	17.40	0.89	0.58	0.75
W2V2-Fr-VP-*large*	20.72	7.43	4.66	0.43	19.88	5.39	3.62	0.49
XLSR-53-*large*	20.54	0.59	0.41	0.49	19.93	0.44	0.62	0.29

LeBenchmark: A reproducible framework for SSL

- Automatic Emotion Recognition
  - Datasets: RECOLA (4h), AlloSat (37h)
  - Task: time-continuous prediction of affective dimensions (arousal, valence, satisfaction)
  - System: linear layer + tanh, GRU, performance: concordance correlation coefficient

Corpus		**RECOLA**		**AlloSat**
Model	**Feature**	**Arousal**	**Valence**	**Satisfaction**
Linear-Tanh	MFB	0.192	0.075	0.065
Linear-Tanh	W2V2-Fr-M-*base*	**0.385**	**0.090**	**0.193**
Linear-Tanh	XLSR-53-*large*	0.155	0.024	0.093
GRU-32	MFB	0.654	0.252	0.437
GRU-32	W2V2-Fr-M-*base*	**0.767**	**0.376**	**0.507**
GRU-32	XLSR-53-*large*	0.605	0.320	0.446
GRU-64	MFB	0.712	0.307	0.400
GRU-64	W2V2-Fr-M-*base*	**0.760**	**0.352**	**0.507**
GRU-64	XLSR-53-*large*	0.585	0.280	0.434

Conclusion

- We trained SSL Wav2Vec 2.0 models for French on large and diverse collection of speech
- SSL models seem benefinitial for lower resource tasks (SLU, AST/TEDx, AER) or simple architectures (AER)
- SSL models do not improve compared to MFCC for end-to-end ASR (no fine-tuning)
- Models and scripts available online
  - Github: https://github.com/LeBenchmark
  - Hugginface: https://huggingface.co/LeBenchmark

- Ongoing work
  - Extension of the collection of speech data (7.7k hours at the moment)
  - Perform fine-tuning of the wav2vec models
    - Pursue unsupervised training on task data
    - Perform end-to-end supervised training on ASR
    - Perform end-to-end supervised training on task data
  - Jointly learn a model that predicts masked speech units and text units

Bibliography

[1] Ringeval, Fabien, et al. “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions”, 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, 2013.

[2] Khorram, Soheil, Melvin McInnis, and Emily Mower Provost. "Jointly aligning and predicting continuous emotion annotations." IEEE Transactions on Affective Computing (2019).

[3] Ravanelli, Mirco, and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet." IEEE Spoken Language Technology Workshop (SLT). IEEE (2018).

[4] Ringeval, Fabien, et al. “AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition”, Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019.

[5] Han, Jing, et al. “Bags in bag: Generating context-aware bags for tracking emotions from speech”, Interspeech 2018. ISCA, 2018.

[6] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).

[7] Latif, Siddique, et al. "Deep representation learning in speech processing: Challenges, recent advances, and future trends." arXiv preprint arXiv:2001.00378 (2020).

[8] Baevski, Alexei, et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations”, arXiv preprint arXiv:2006.11477, 2020.

[9] Chung, Yu-An, and James Glass. "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech." arXiv preprint arXiv:1803.08976 (2018).