Self-Supervised Speech Processing

Gérard Chollet (CNRS-SAMOVAR)

Yingzhi Wang (Zaion)

gerard.chollet@telecom-sudparis.eu

ywang@zaion.ai

Intelligent Voice® zaion Une école Institut Mines-Télécom TELECOM SudParis cnrs Depuis 80 ans, nos connaissances bâtissent de nouveaux mondes speechmorphing

2020

End-to-End Machine Learning for Speech Processing: Speech-to-Speech, Speech-to-Text and Text-to-Speech

https://hal.telecom-paris.fr/hal-02937687/file/E2E2Plenary.pdf

Gérard Chollet (CNRS-SAMOVAR)
and Colleagues from IV, Zaion, CNRS/IMT and SMI

gerard.chollet@telecom-sudparis.eu

2018

How to protect privacy in Cloud Computing?

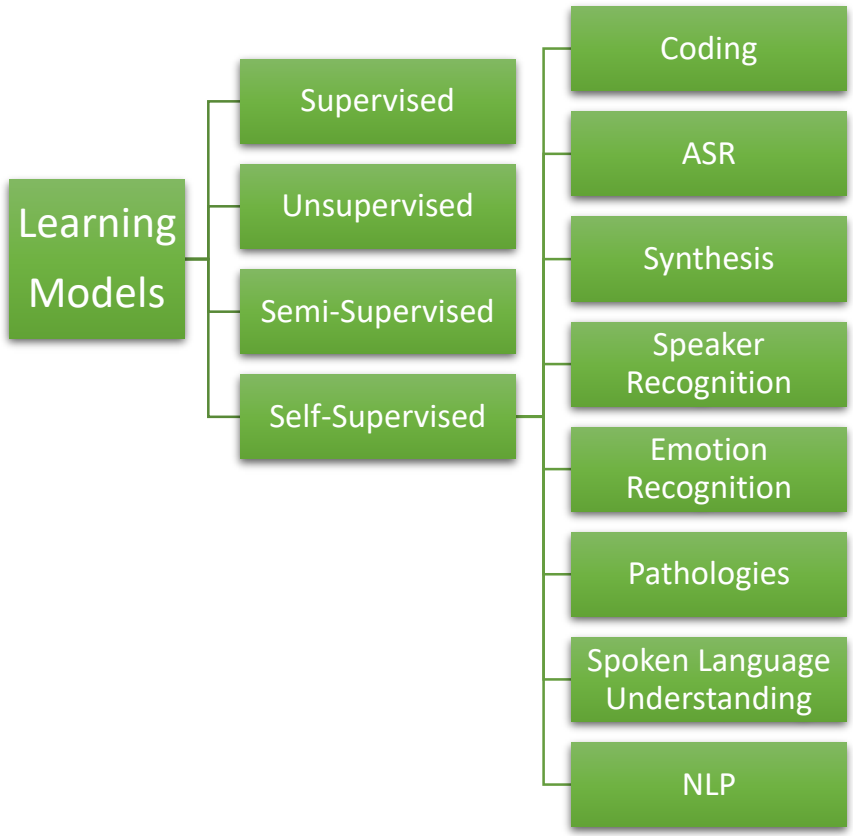
Gérard Chollet (IV & CNRS/IMT)
Bhiksha Raj (CMU)

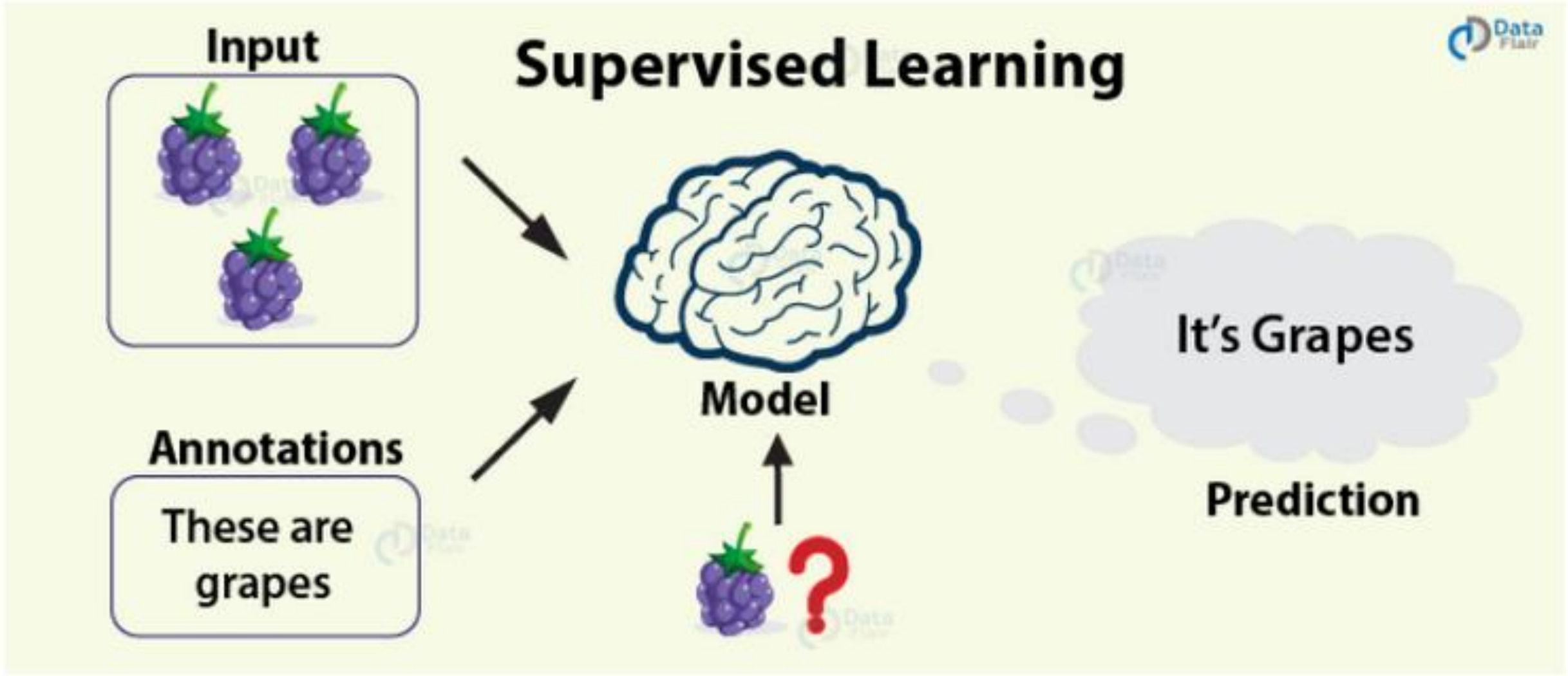
2017

Privacy Preserving Speech Processing

Gérard Chollet (IV & CNRS-IMT)

History / Overview



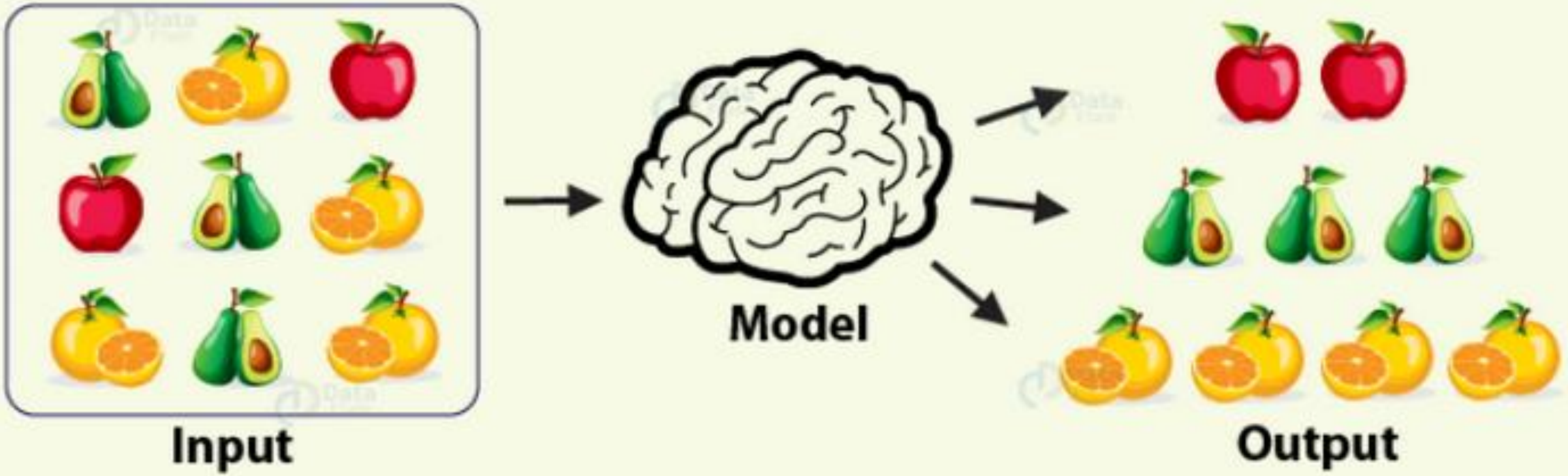


Supervised learning | Source

Needs labeled data !

https://neptune.ai/blog/self-supervised-learning

Unsupervised Learning

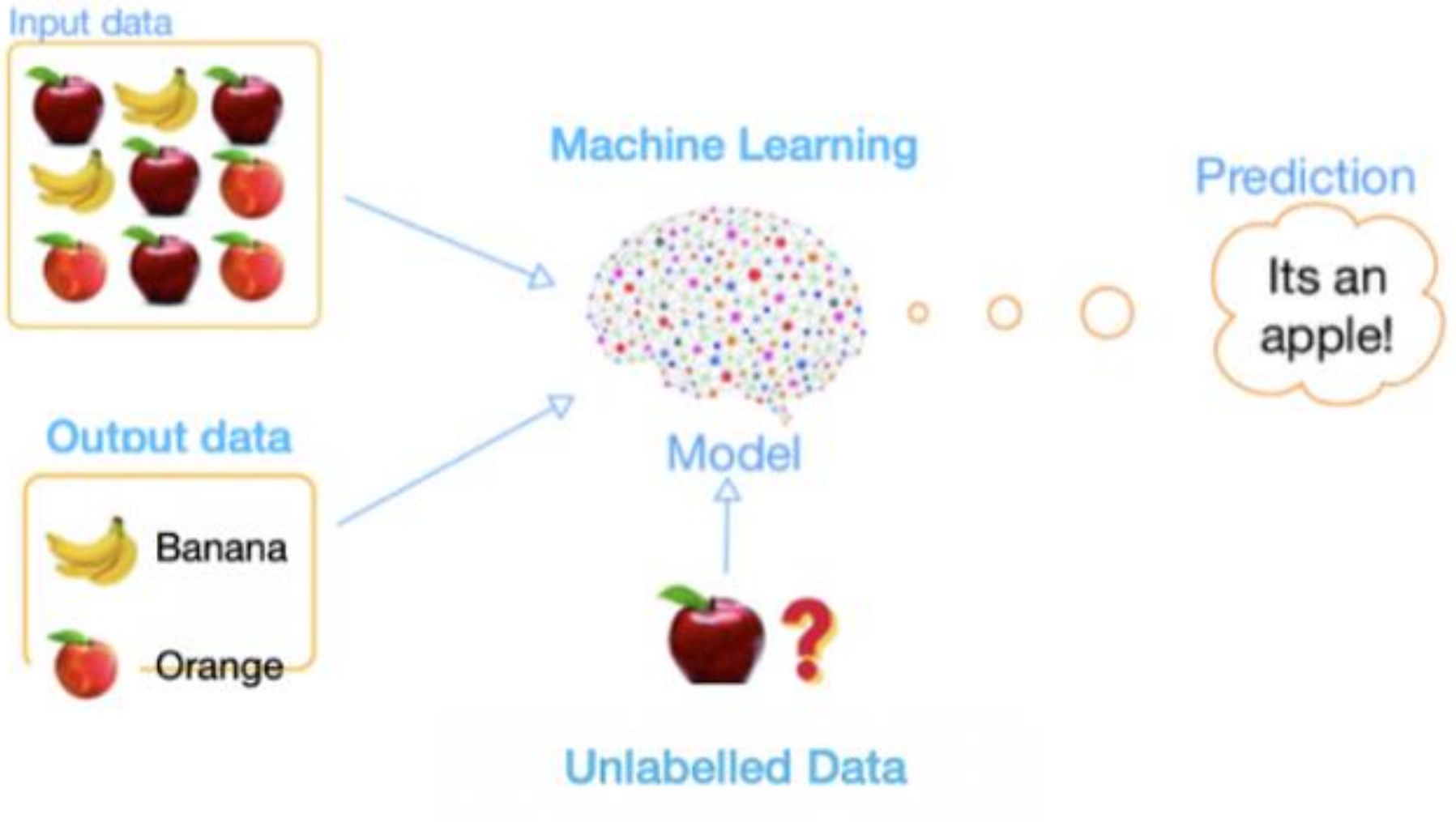


Unsupervised learning | Source

discovers patterns in data without pre-assigned labels

https://neptune.ai/blog/self-supervised-learning

Semi-supervised learning

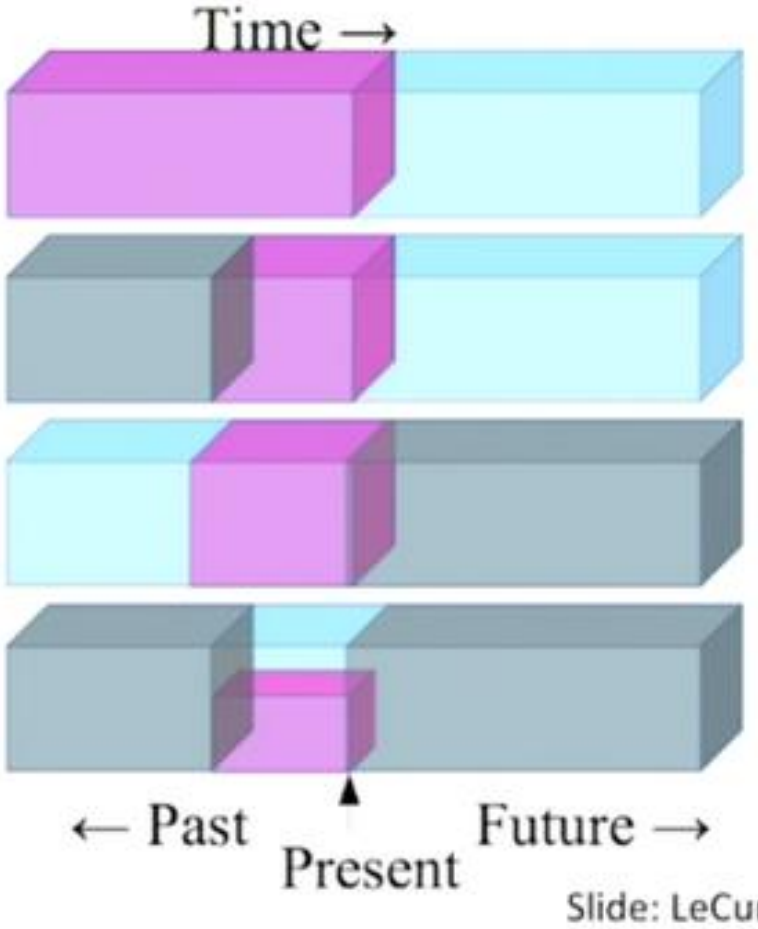


uses a small number of labeled samples to guide learning with a larger amount of unlabeled data

https://neptune.ai/blog/self-supervised-learning

What is self-supervised learning?

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



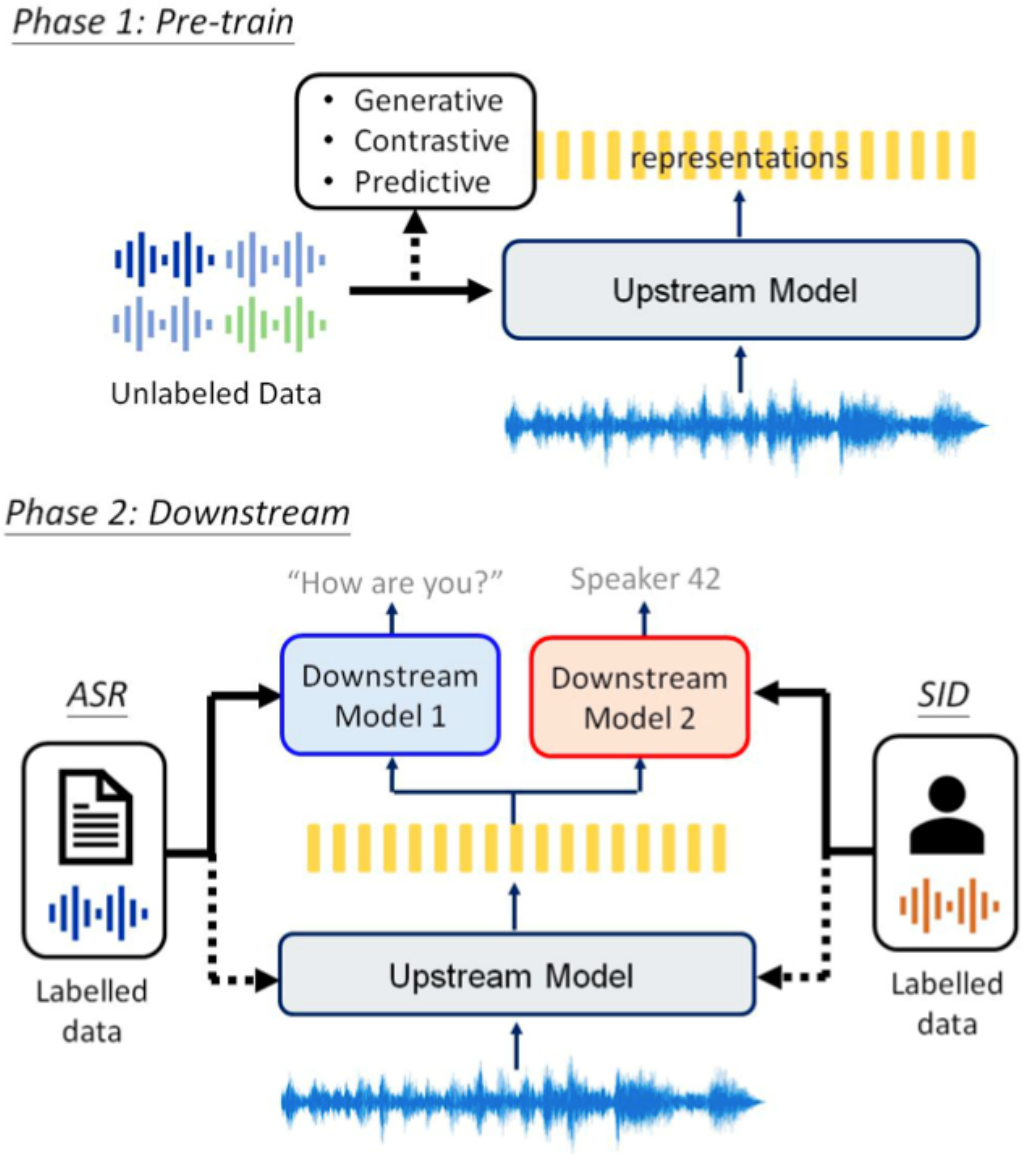
uses information from input data as the label to learn representations useful for downstream tasks

https://neptune.ai/blog/self-supervised-learning

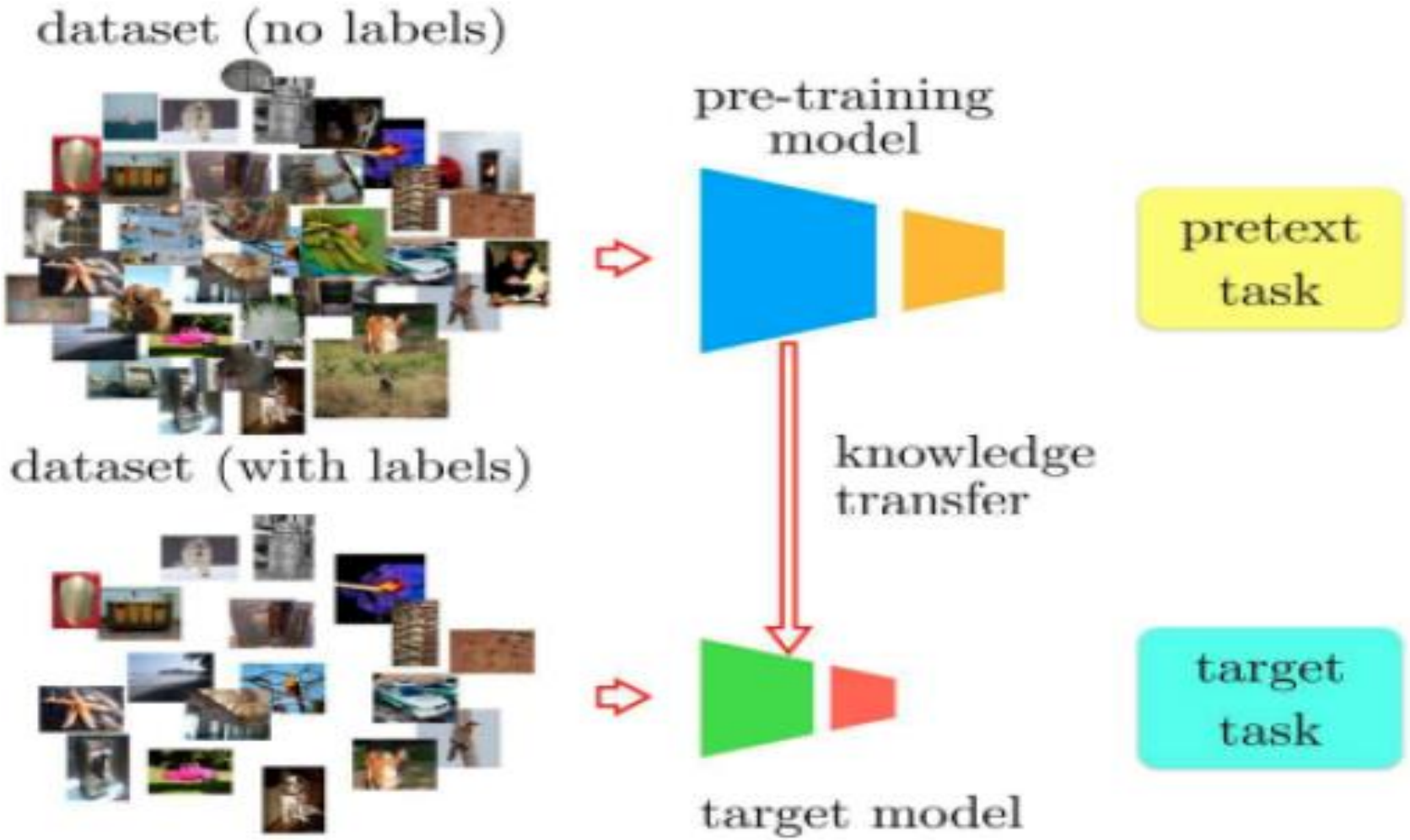
Self-supervised learning Framework

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



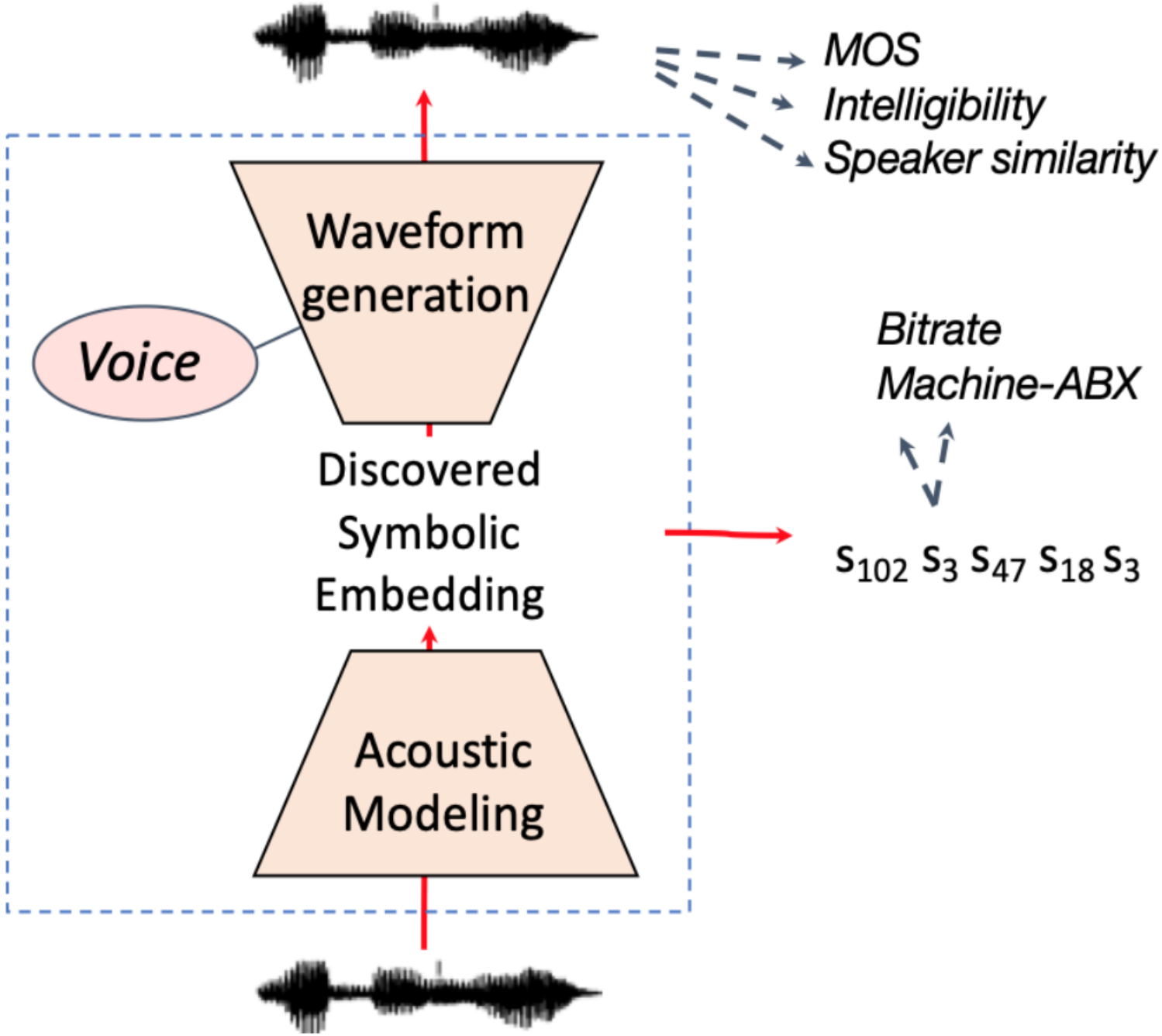
Vanilla Self-Supervised approach



https://neptune.ai/blog/self-supervised-learning

An example:

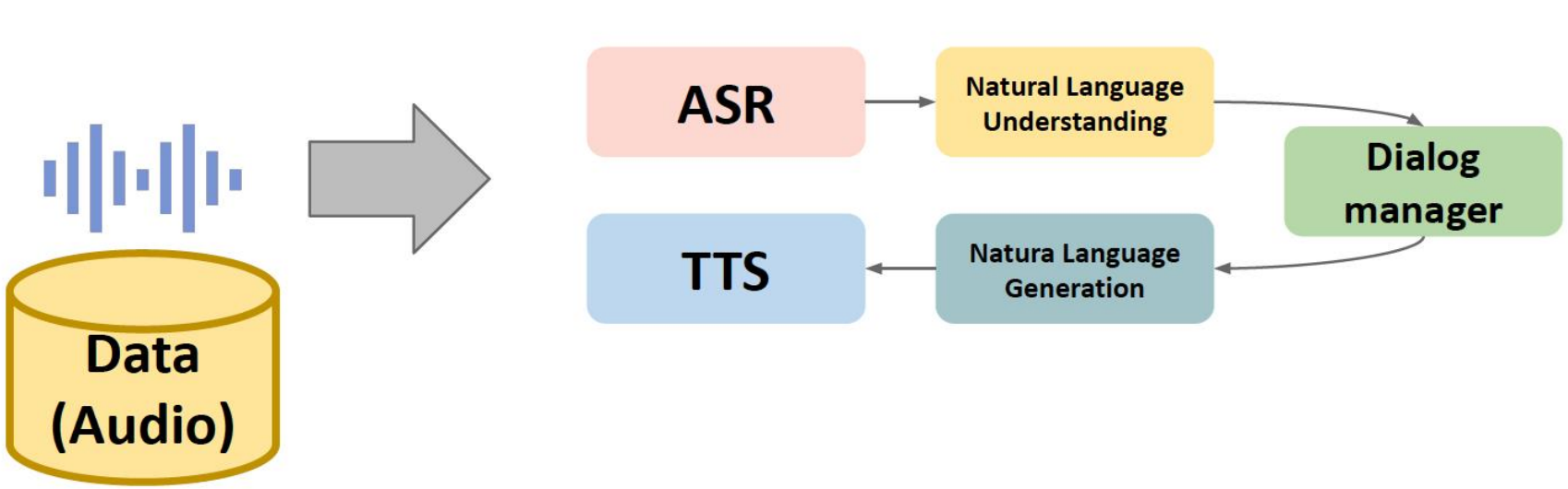
ZeroSpeech 2019: TTS without T



https://zerospeech.com/2019/

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

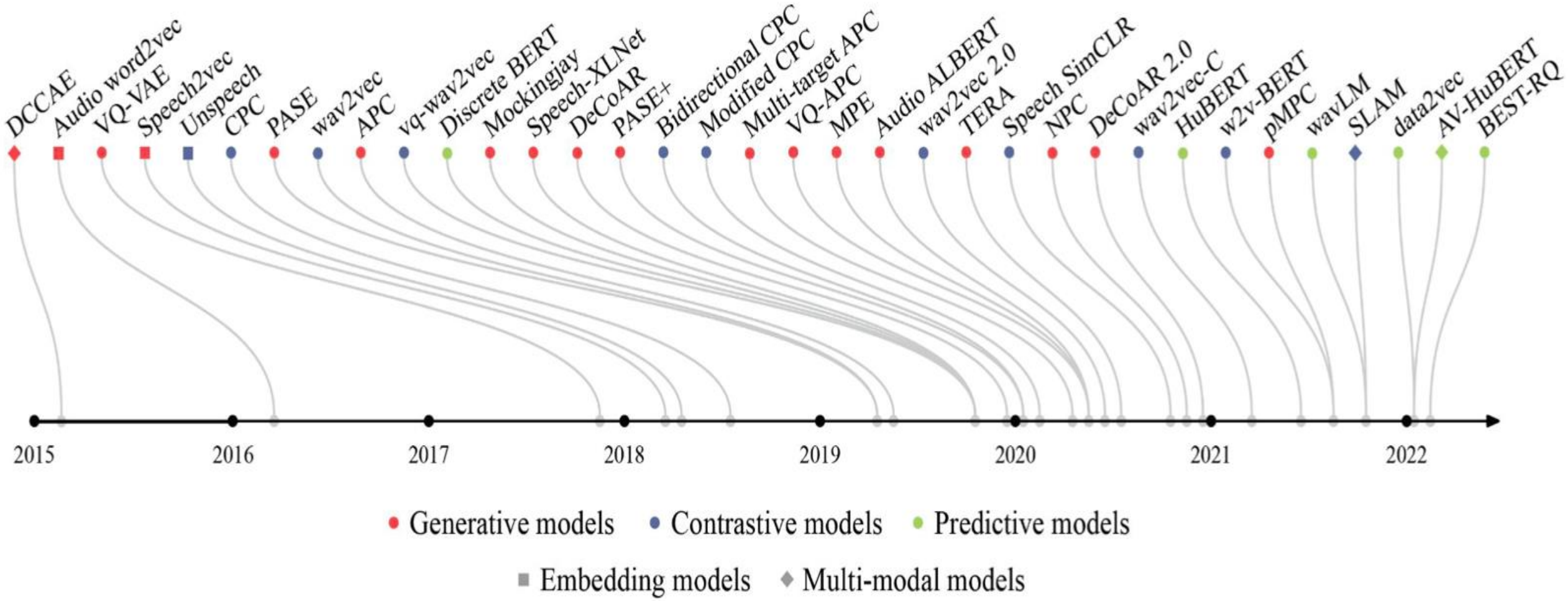


ICASSP 2022 SSL tutorial
https://2022.ieeeicassp.org/tutorials.php#tut22

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.

SSL methods TimeLine



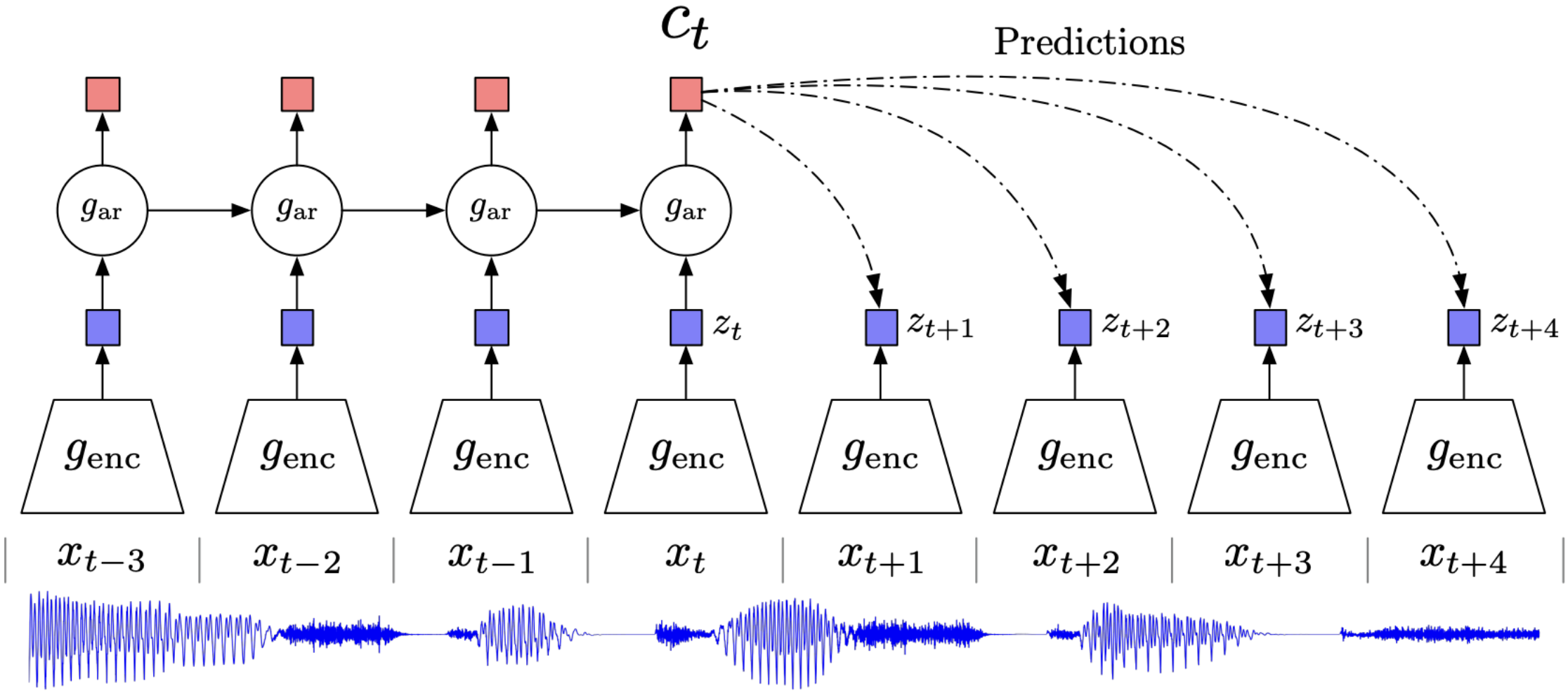
Representation Learning with Contrastive Predictive Coding

Contrastive approaches

Aaron van den Oord
DeepMind
avdnoord@google.com

Yazhe Li
DeepMind
yazhe@google.com

Oriol Vinyals
DeepMind
vinyals@google.com



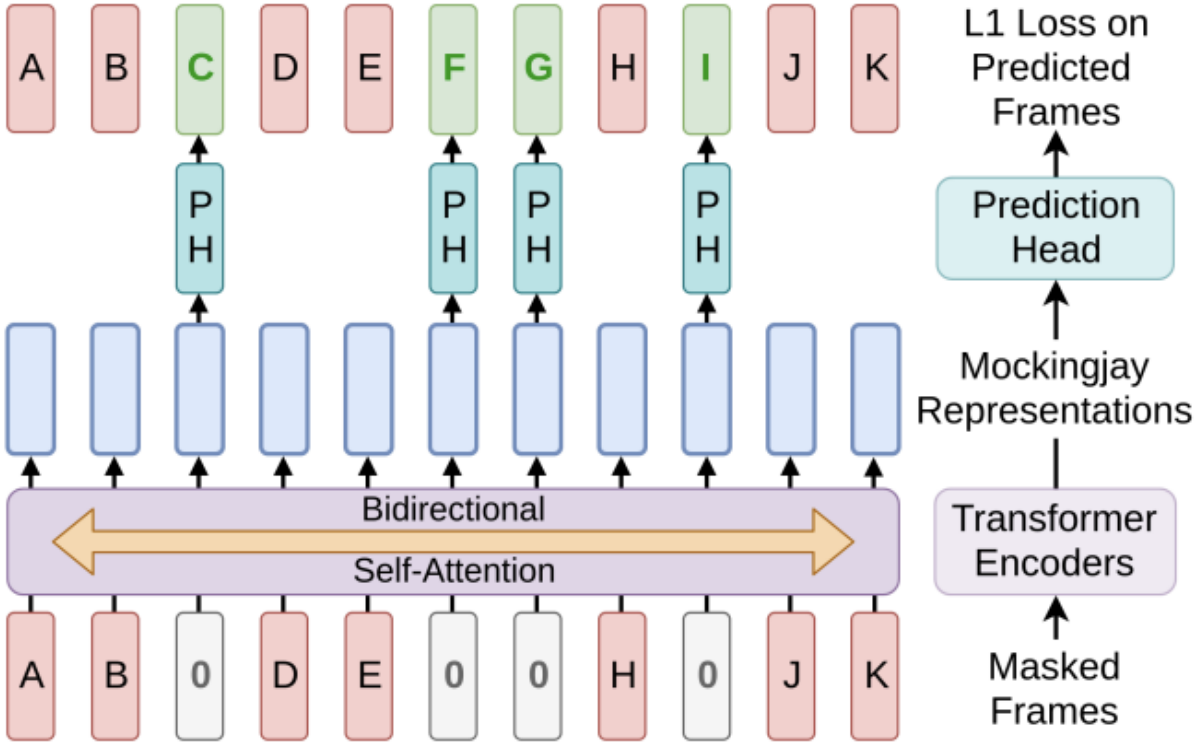
https://arxiv.org/pdf/1807.03748.pdf

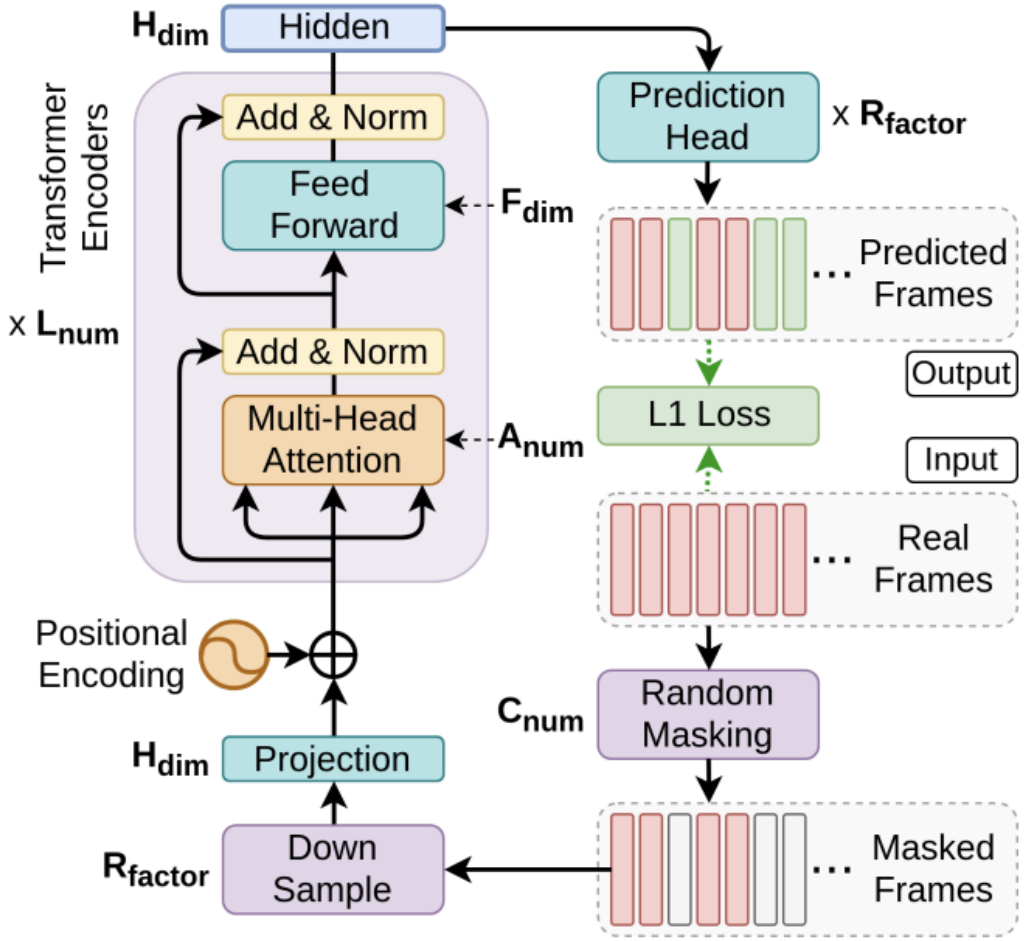
CPC contributions

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

MOCKINGJAY: UNSUPERVISED SPEECH REPRESENTATION LEARNING WITH DEEP BIDIRECTIONAL TRANSFORMER ENCODERS

National Taiwan University

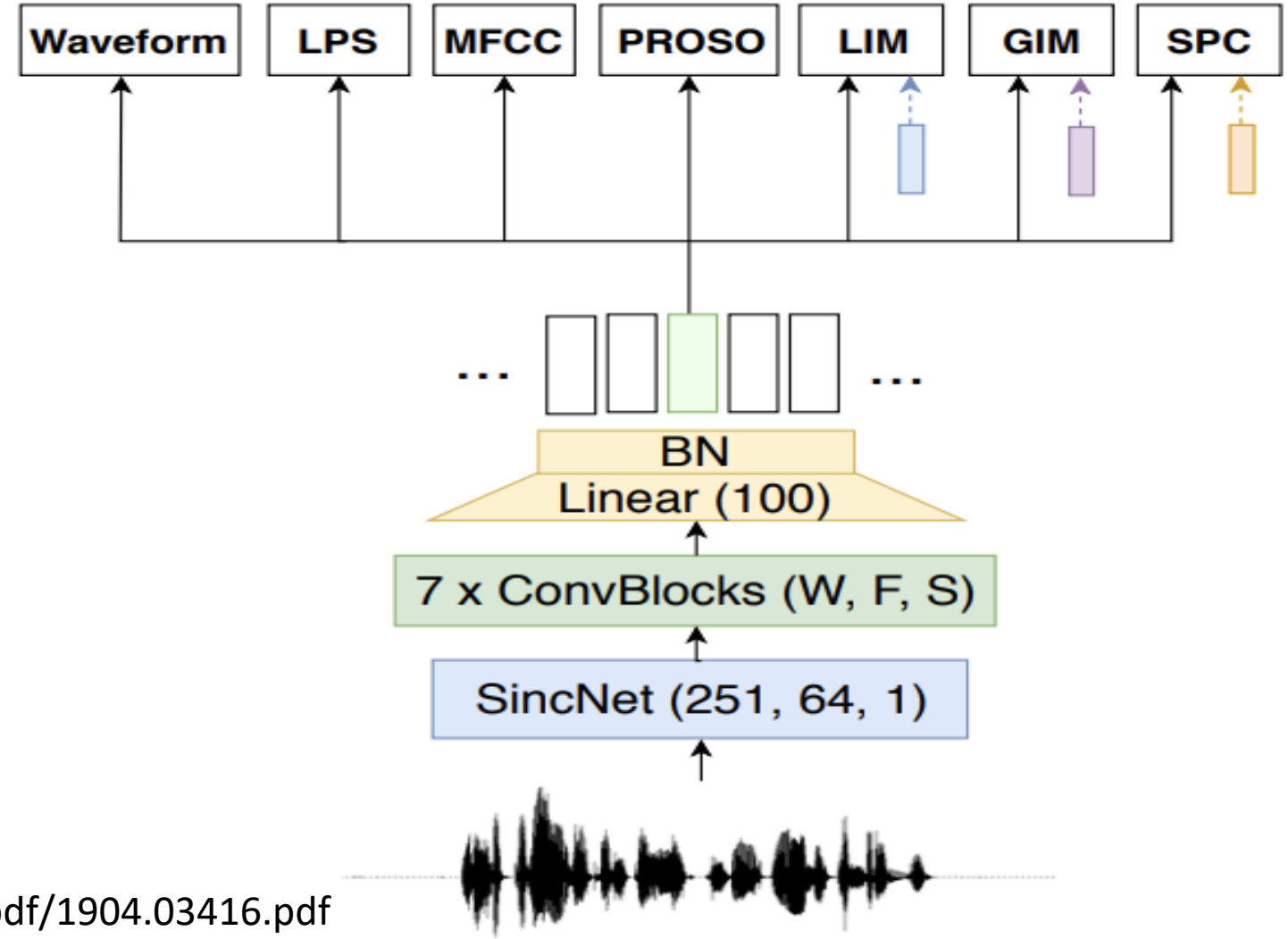




https://arxiv.org/pdf/1910.12638.pdf

Learning Problem-agnostic Speech Representations from Multiple Self-supervised Tasks

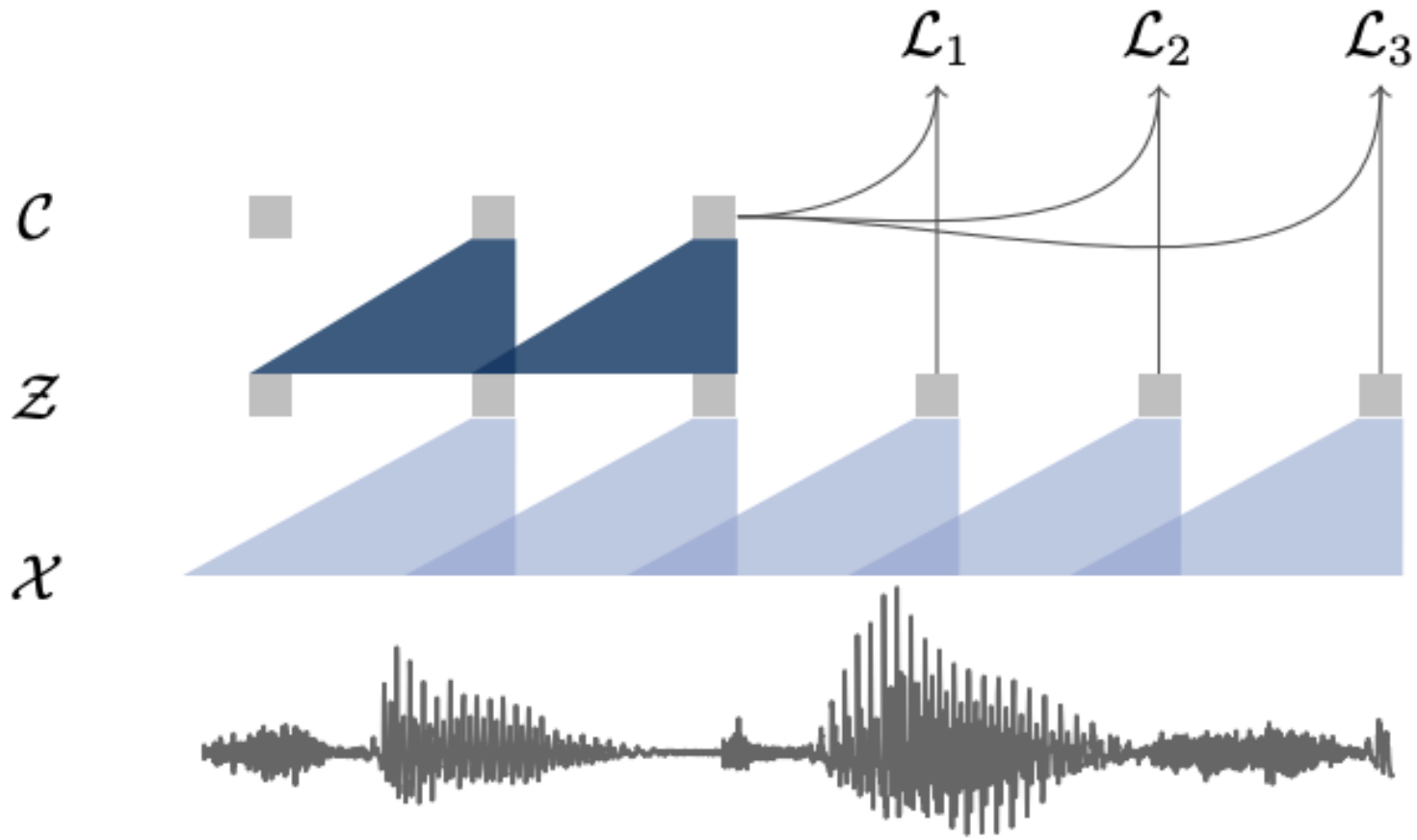
Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, Yoshua Bengio



https://arxiv.org/pdf/1904.03416.pdf

WAV2VEC: UNSUPERVISED PRE-TRAINING FOR SPEECH RECOGNITION

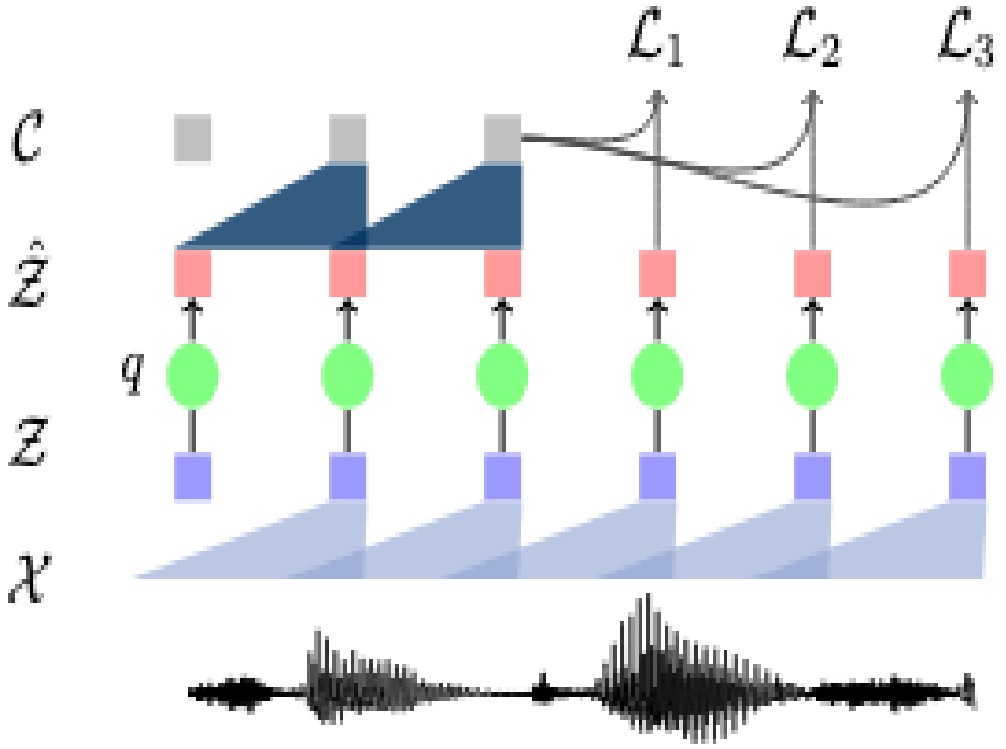
Steffen Schneider, Alexei Baevski, Ronan Collobert, Michael Auli
Facebook AI Research

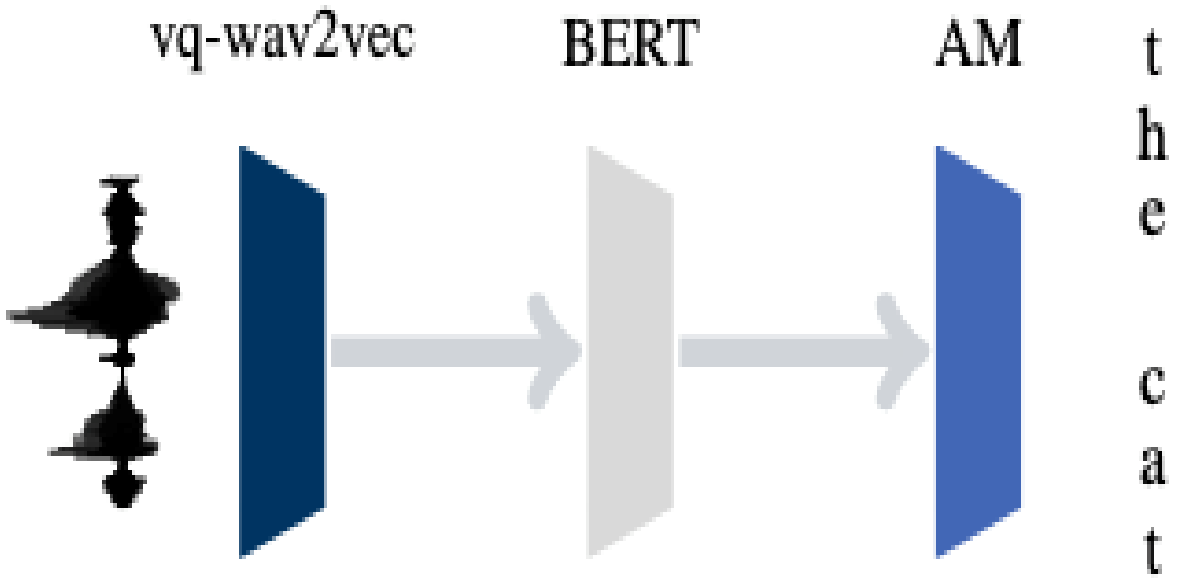


https://arxiv.org/pdf/1904.05862.pdf

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

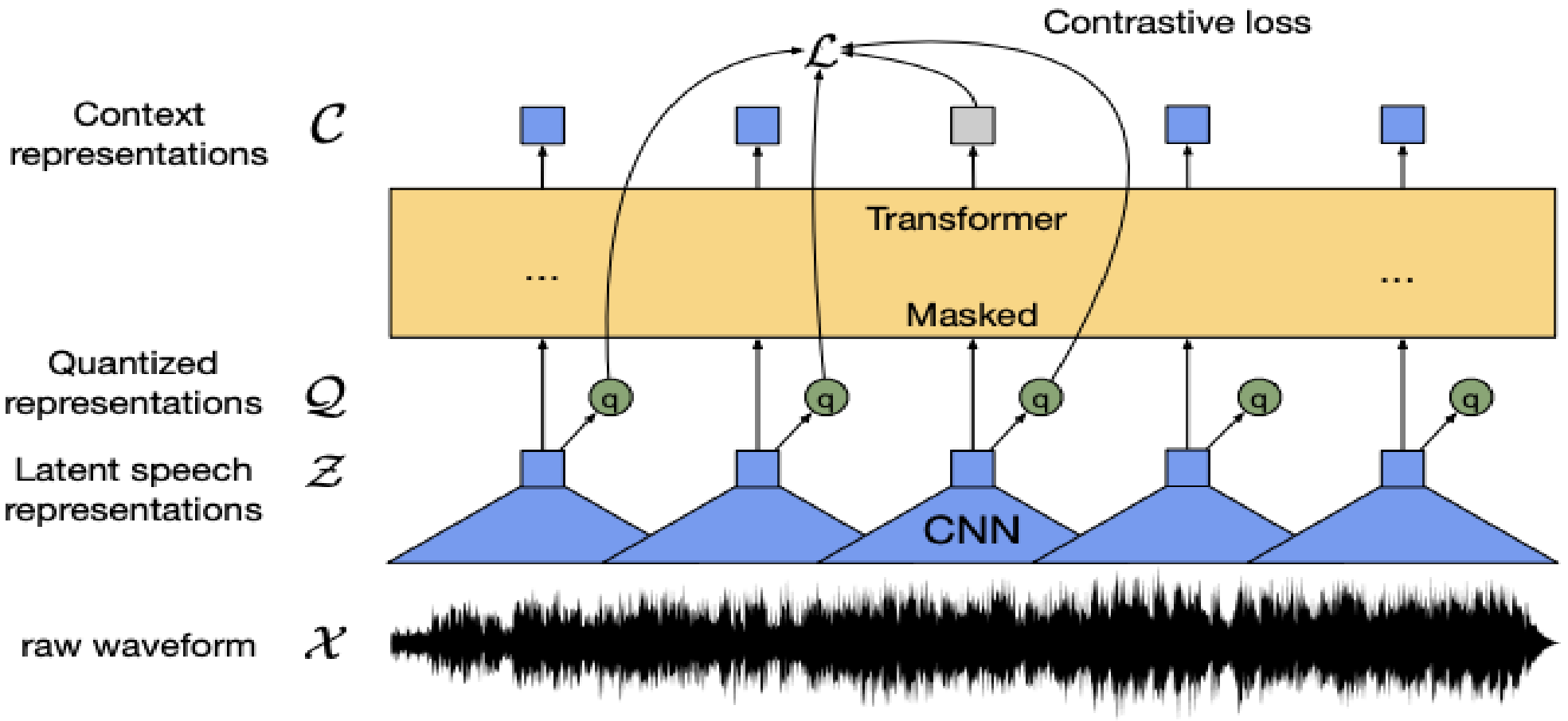




https://arxiv.org/pdf/1910.05453.pdf

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

Alexei Baevski Henry Zhou Abdelrahman Mohamed Michael Auli



https://arxiv.org/pdf/2006.11477.pdf

Wav2vec2.0 contributions

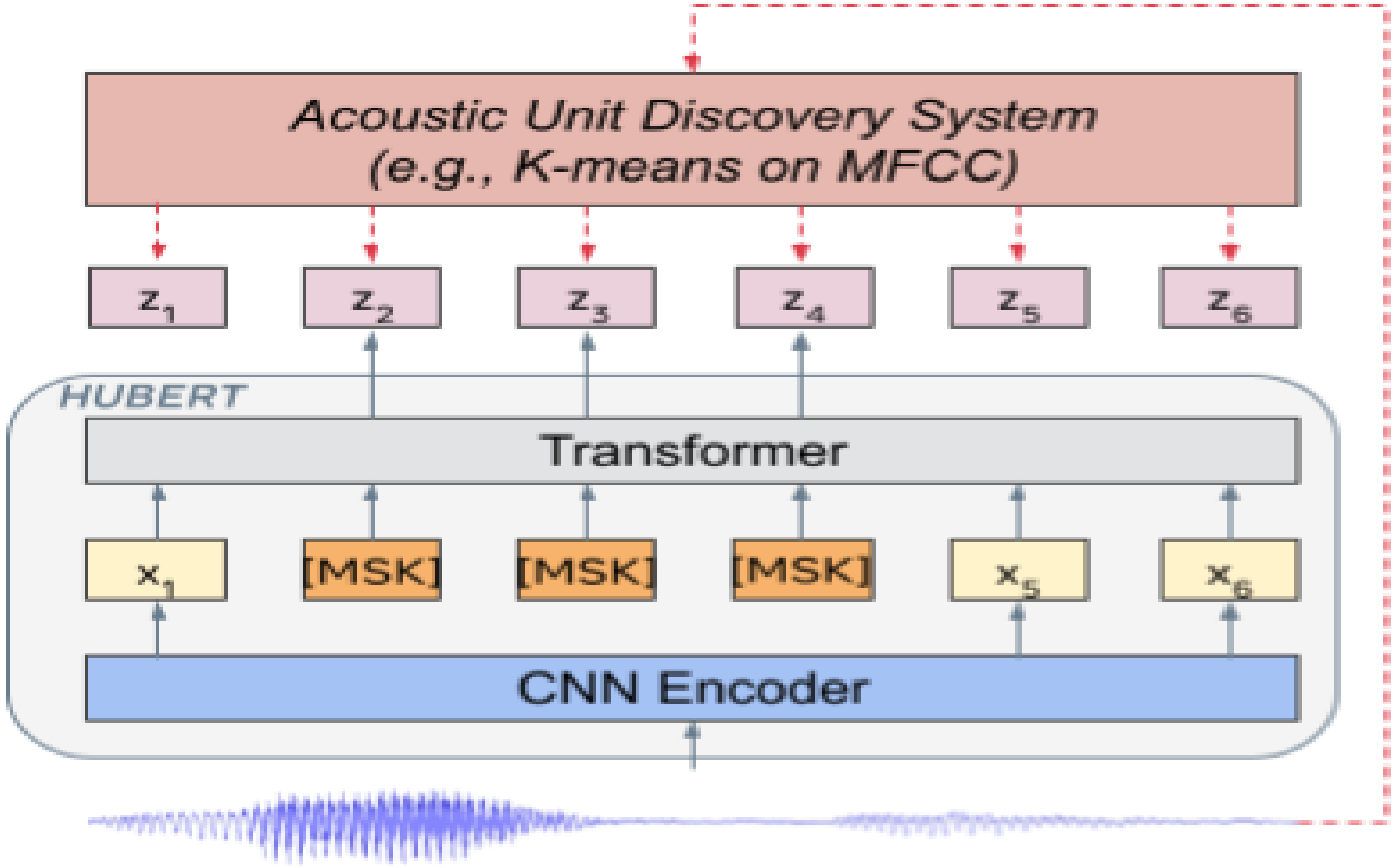
- The first approach to show significant improvements for low-resource ASR.
- Impressive results on multilingual representations.
- Strong performance on a wide range of downstream speech tasks.

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

Wav2vec2.0 contributions

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

HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units



Predictive approaches

https://arxiv.org/pdf/2106.07447.pdf

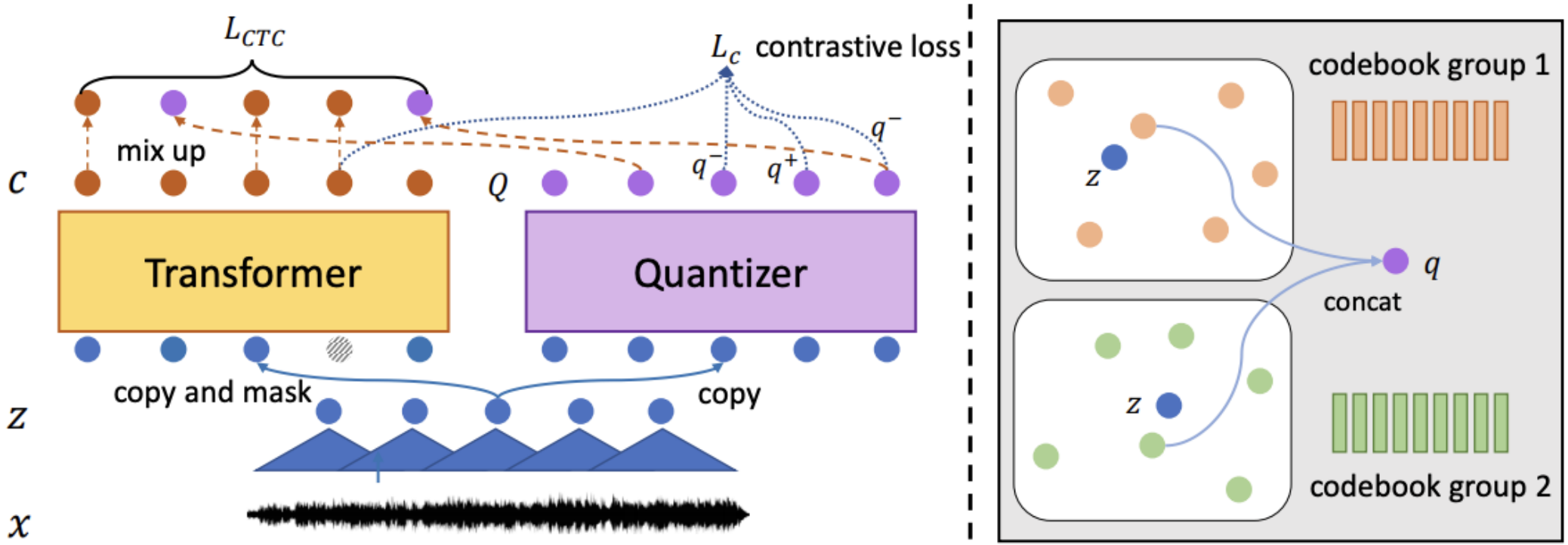
HuBERT contributions

- 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.
- 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.
  - ….

UniSpeech: Unified Speech Representation Learning with Labeled and Unlabeled Data

Chengyi Wang [*1] Yu Wu [2] Yao Qian [2] Kenichi Kumatani [2] Shujie Liu [2] Furu Wei [2] Michael Zeng [2]
Xuedong Huang [2]

UniSpeech



https://arxiv.org/pdf/2101.07597.pdf

SUPERB: Speech processing Universal PERformance Benchmark

Shu-wen Yang[1], *Po-Han Chi*[1*], *Yung-Sung Chuang*[1*], *Cheng-I Jeff Lai*[2*], *Kushal Lakhotia*[3*], *Yist Y. Lin*[1*], *Andy T. Liu*[1*], *Jiatong Shi*[4*], *Xuankai Chang*[6], *Guan-Ting Lin*[1], *Tzu-Hsien Huang*[1], *Wei-Cheng Tseng*[1], *Ko-tik Lee*[1], *Da-Rong Liu*[1], *Zili Huang*[4], *Shuyan Dong*[5†], *Shang-Wen Li*[5†], *Shinji Watanabe*[6], *Abdelrahman Mohamed*[3], *Hung-yi Lee*[1]

[1]National Taiwan University, Taiwan
[2]Massachusetts Institute of Technology, USA
[3]Facebook AI Research, USA
[4]Johns Hopkins University, USA
[5]Amazon AI, USA
[6]Carnegie Mellon University, USA

leo19941227@gmail.com, kushall@fb.com, clai24@mit.edu, jshi34@jhu.edu,
shangwel@amazon.com, shinjiw@ieee.org, hungyilee@ntu.edu.tw

https://arxiv.org/pdf/2105.01051.pdf

SUPERB: Speech processing Universal PERformance Benchmark

Method	Network	#Params	Stride	Input	Corpus	Pretraining	Official Github
FBANK	-	0	10ms	waveform	-	-	-
PASE+ [16]	SincNet, 7-Conv, 1-QRNN	7.83M	10ms	waveform	LS 50 hr	multi-task	santi-pdp / pase
APC [7]	3-GRU	4.11M	10ms	FBANK	LS 360 hr	F-G	iamyuanchung / APC
VQ-APC [32]	3-GRU	4.63M	10ms	FBANK	LS 360 hr	F-G + VQ	iamyuanchung / VQ-APC
NPC [33]	4-Conv, 4-Masked Conv	19.38M	10ms	FBANK	LS 360 hr	M-G + VQ	Alexander-H-Liu / NPC
Mockingjay [8]	12-Trans	85.12M	10ms	FBANK	LS 360 hr	time M-G	s3prl / s3prl
TERA [9]	3-Trans	21.33M	10ms	FBANK	LS 960 hr	time/freq M-G	s3prl / s3prl
DeCoAR 2.0 [10]	12-Trans	89.84M	10ms	FBANK	LS 960 hr	time M-G + VQ	awslabs / speech-representations
modified CPC [34]	5-Conv, 1-LSTM	1.84M	10ms	waveform	LL 60k hr	F-C	facebookresearch / CPC_audio
wav2vec [12]	19-Conv	32.54M	10ms	waveform	LS 960 hr	F-C	pytorch / fairseq
vq-wav2vec [13]	20-Conv	34.15M	10ms	waveform	LS 960 hr	F-C + VQ	pytorch / fairseq
wav2vec 2.0 Base [14]	7-Conv 12-Trans	95.04M	20ms	waveform	LS 960 hr	M-C + VQ	pytorch / fairseq
wav2vec 2.0 Large [14]	7-Conv 24-Trans	317.38M	20ms	waveform	LL 60k hr	M-C + VQ	pytorch / fairseq
HuBERT Base [35]	7-Conv 12-Trans	94.68M	20ms	waveform	LS 960 hr	M-P + VQ	pytorch / fairseq
HuBERT Large [35]	7-Conv 24-Trans	316.61M	20ms	waveform	LL 60k hr	M-P + VQ	pytorch / fairseq

SUPERB: Speech processing Universal PERformance Benchmark

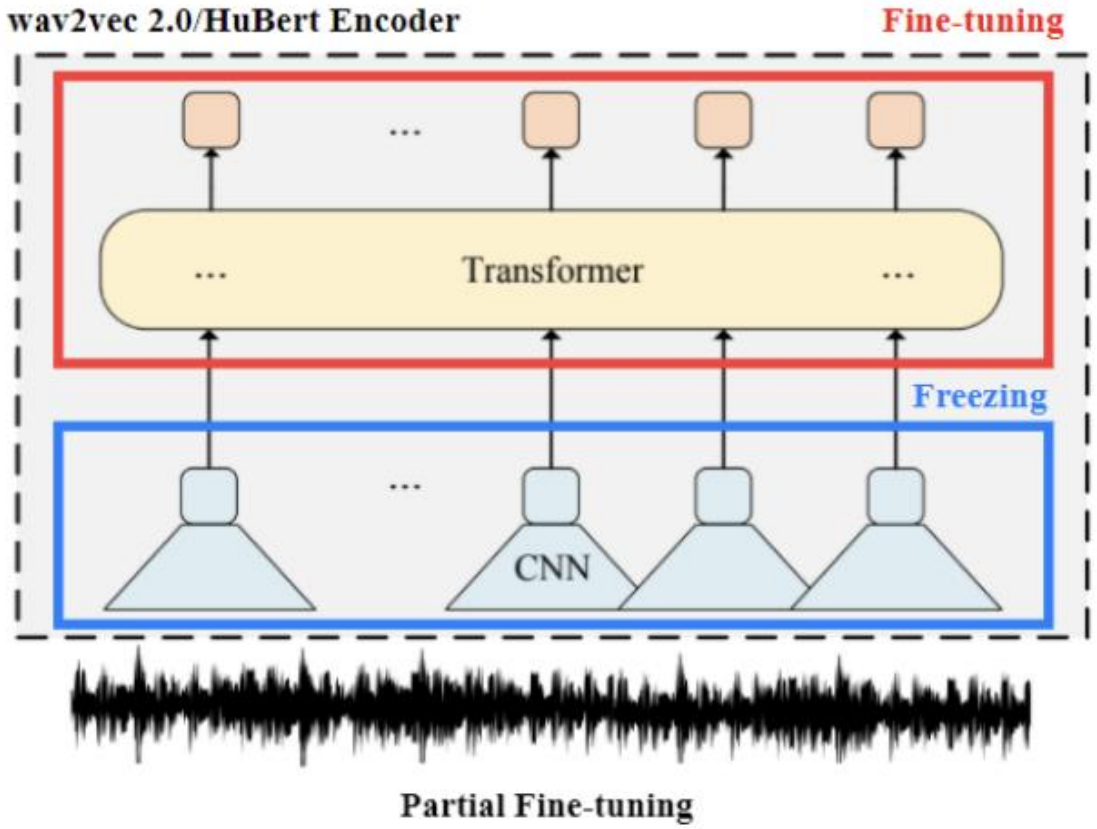
	PR	KS	IC	SID	ER	ASR (WER)		QbE	SF		ASV	SD
	PER ↓	Acc ↑	Acc ↑	Acc ↑	Acc ↑	w/o ↓	w/ LM ↓	MTWV ↑	F1 ↑	CER ↓	EER ↓	DER ↓
FBANK	82.01	8.63	9.10	8.5E-4	35.39	23.18	15.21	0.0058	69.64	52.94	9.56	10.05
PASE+ [16]	58.87	82.54	29.82	37.99	57.86	25.11	16.62	0.0072	62.14	60.17	11.61	8.68
APC [7]	41.98	91.01	74.69	60.42	59.33	21.28	14.74	0.0310	70.46	50.89	8.56	10.53
VQ-APC [32]	41.08	91.11	74.48	60.15	59.66	21.20	15.21	0.0251	68.53	52.91	8.72	10.45
NPC [33]	43.81	88.96	69.44	55.92	59.08	20.20	13.91	0.0246	72.79	48.44	9.4	9.34
Mockingjay [8]	70.19	83.67	34.33	32.29	50.28	22.82	15.48	6.6E-04	61.59	58.89	11.66	10.54
TERA [9]	49.17	89.48	58.42	57.57	56.27	18.17	12.16	0.0013	67.50	54.17	15.89	9.96
DeCoAR 2.0 [10]	14.93	94.48	90.80	74.42	62.47	13.02	9.07	0.0406	83.28	34.73	7.16	6.59
modified CPC [34]	42.54	91.88	64.09	39.63	60.96	20.18	13.53	0.0326	71.19	49.91	12.86	10.38
wav2vec [12]	31.58	95.59	84.92	56.56	59.79	15.86	11.00	0.0485	76.37	43.71	7.99	9.9
vq-wav2vec [13]	33.48	93.38	85.68	38.80	58.24	17.71	12.80	0.0410	77.68	41.54	10.38	9.93
wav2vec 2.0 Base [14]	5.74	96.23	92.35	75.18	63.43	6.43	4.79	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	87.11	27.31	5.65	**5.62**
HuBERT Base [35]	5.41	96.30	98.34	81.42	64.92	6.42	4.79	**0.0736**	88.53	25.20	**5.11**	5.88
HuBERT Large [35]	**3.53**	95.29	**98.76**	**90.33**	**67.62**	**3.62**	**2.94**	0.0353	**89.81**	**21.76**	5.98	5.75

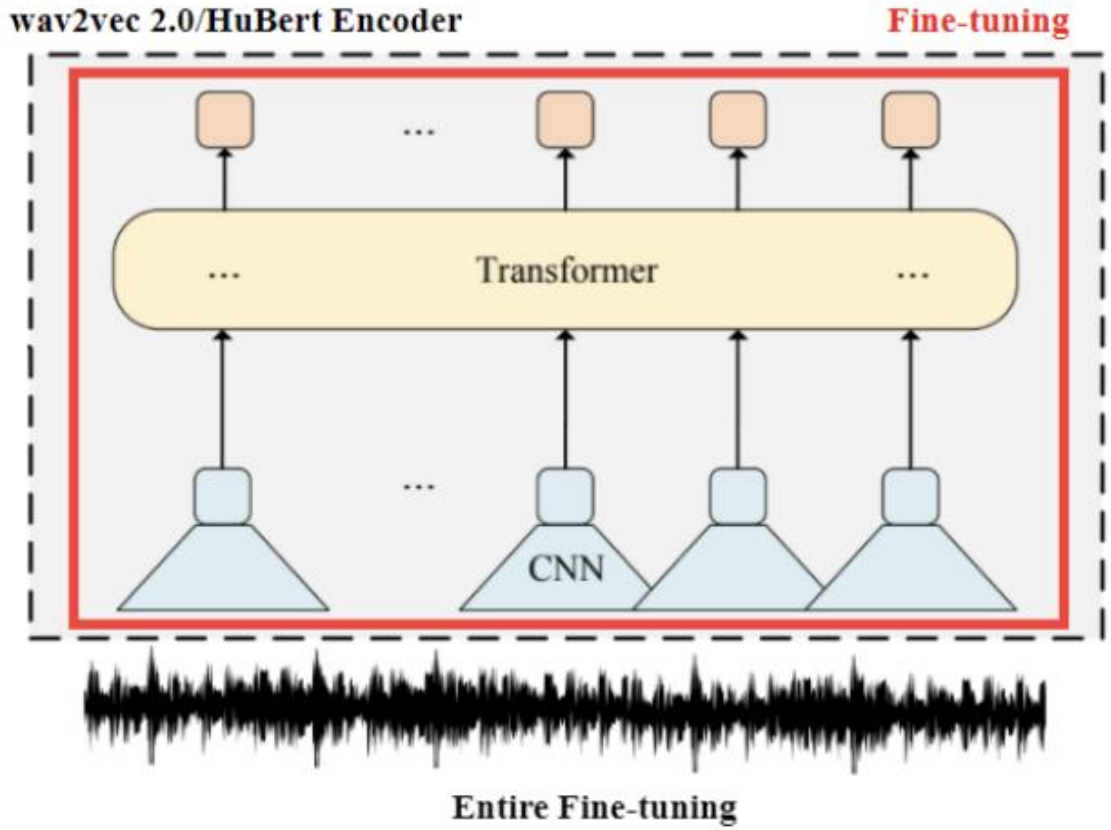
A Fine-tuned Wav2vec2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding

Yingzhi Wang[1], *Abdelmoumene Boumadane*[1], *Abdelwahab Heba*[1]

[1]Zaion Lab, Zaion, Paris, France

ywang@zaion.ai, aboumadane@zaion.ai, aheba@zaion.ai





https://arxiv.org/pdf/2111.02735.pdf

Model	SER (WA%)	SV (EER%)
EF-w2v-base	75.90	2.77
PF-w2v-base	77.02	3.15
EF-w2v-base-960h	73.64	4.46
PF-w2v-base-960h	73.84	4.38
EF-w2v-large	77.00	3.42
PF-w2v-large	77.47	3.85
EF-w2v-large-960h	73.00	4.27
PF-w2v-large-960h	76.75	4.47
EF-hbt-base	76.53	2.84
PF-hbt-base	76.60	3.13
EF-hbt-large	78.52	2.86
PF-hbt-large	**79.58**	3.21
EF-hbt-large-960h	78.78	**2.36**
PF-hbt-large-960h	78.96	2.38
Frozen-w2v-base[15]	63.43	6.02
Frozen-w2v-large[15]	65.64	5.65
Frozen-hbt-base[15]	64.92	5.11
Frozen-hbt-large[15]	67.62	5.98
Attention Pooling[32]	71.75	-
Siamese Capsule[33]	-	3.14

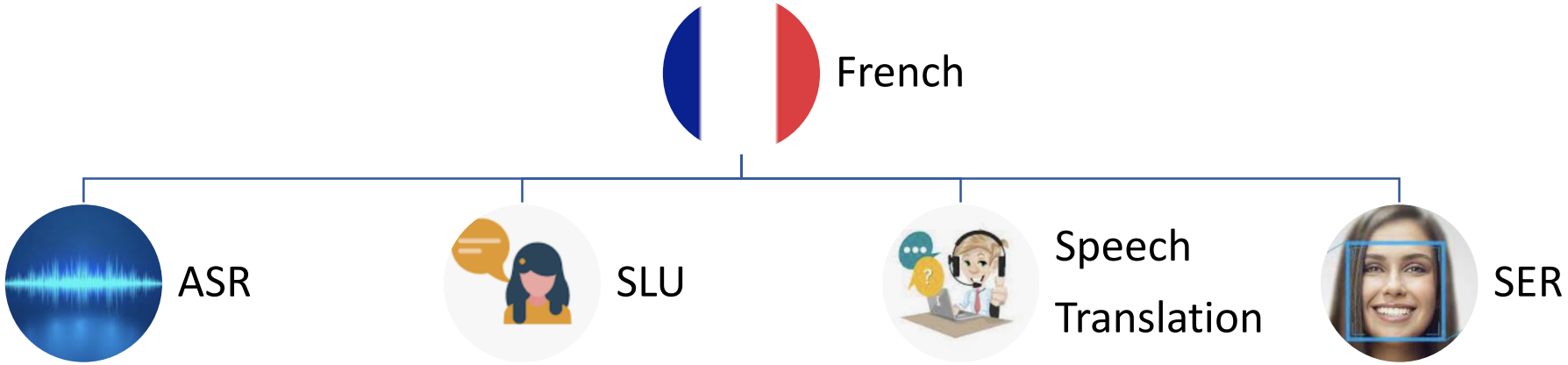
Model	IC (ACC%)	SF (F1%)
EF-w2v-base	87.13	74.32
PF-w2v-base	86.58	74.73
EF-w2v-base-960h	85.89	74.33
PF-w2v-base-960h	86.13	73.78
EF-w2v-large	85.80	72.45
PF-w2v-large	86.29	73.16
EF-w2v-large-960h	86.10	73.39
PF-w2v-large-960h	86.35	74.03
EF-hbt-base	87.44	75.06
PF-hbt-base	87.51	75.32
EF-hbt-large	**89.38**	78.43
PF-hbt-large	89.22	**78.92**
EF-hbt-large-960h	88.71	78.89
PF-hbt-large-960h	88.32	78.17
Frozen-w2v-base	47.15	37.66
Frozen-w2v-large	3.88	3.85
Frozen-hbt-base	68.74	57.06
Frozen-hbt-large	74.42	60.07
CTI[34]	86.92	74.66

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, 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

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



https://arxiv.org/pdf/2104.11462.pdf

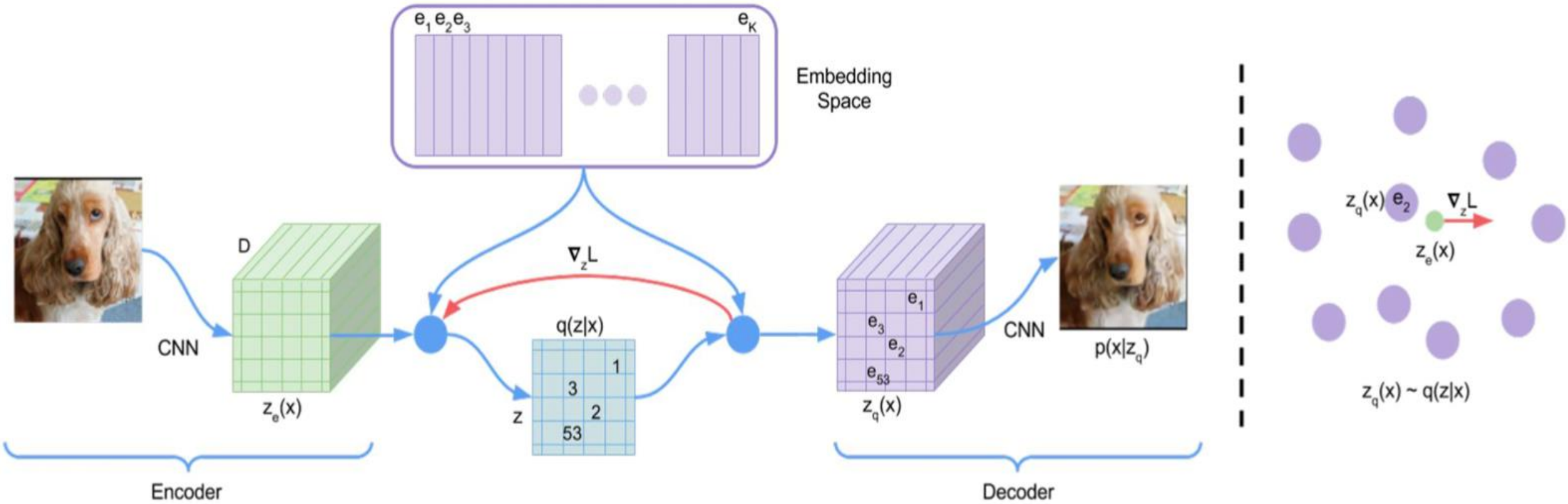
Generative approaches

Neural Discrete Representation Learning

Aaron van den Oord
DeepMind
avdnoord@google.com

Oriol Vinyals
DeepMind
vinyals@google.com

Koray Kavukcuoglu
DeepMind
korayk@google.com

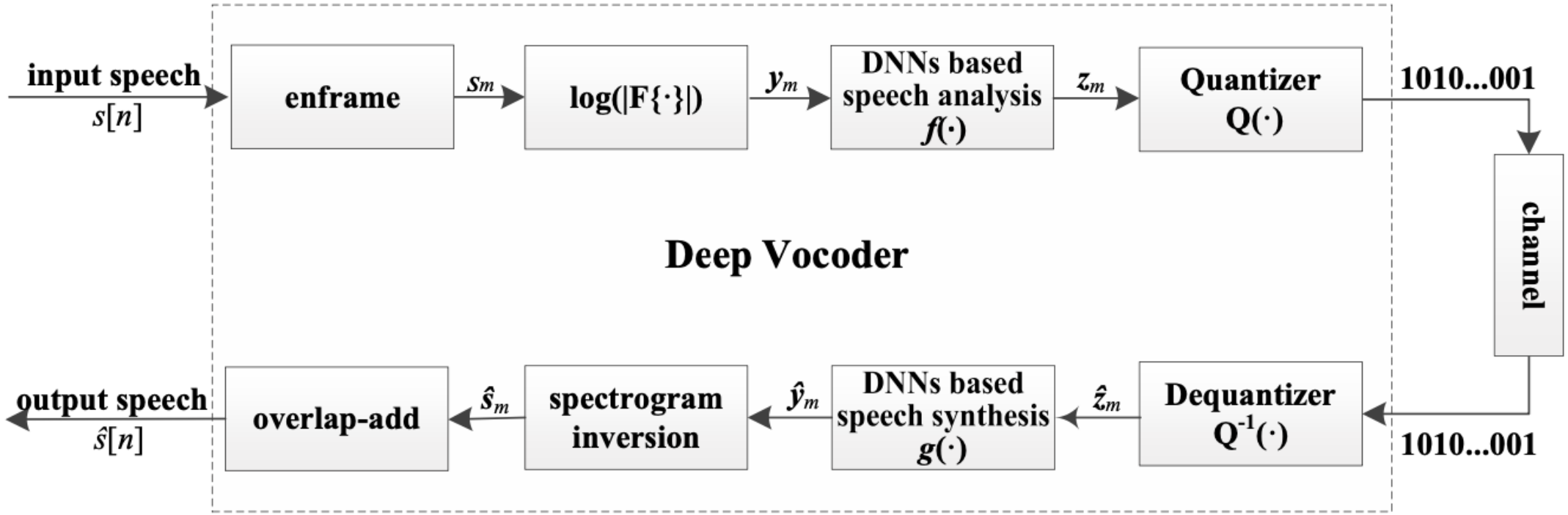


https://arxiv.org/pdf/1711.00937.pdf

DEEP VOCODER: LOW BIT RATE COMPRESSION OF SPEECH WITH DEEP AUTOENCODER

Gang Min [1], *Changqing Zhang* [1], *Xiongwei Zhang* [2], *Wei Tan* [1]

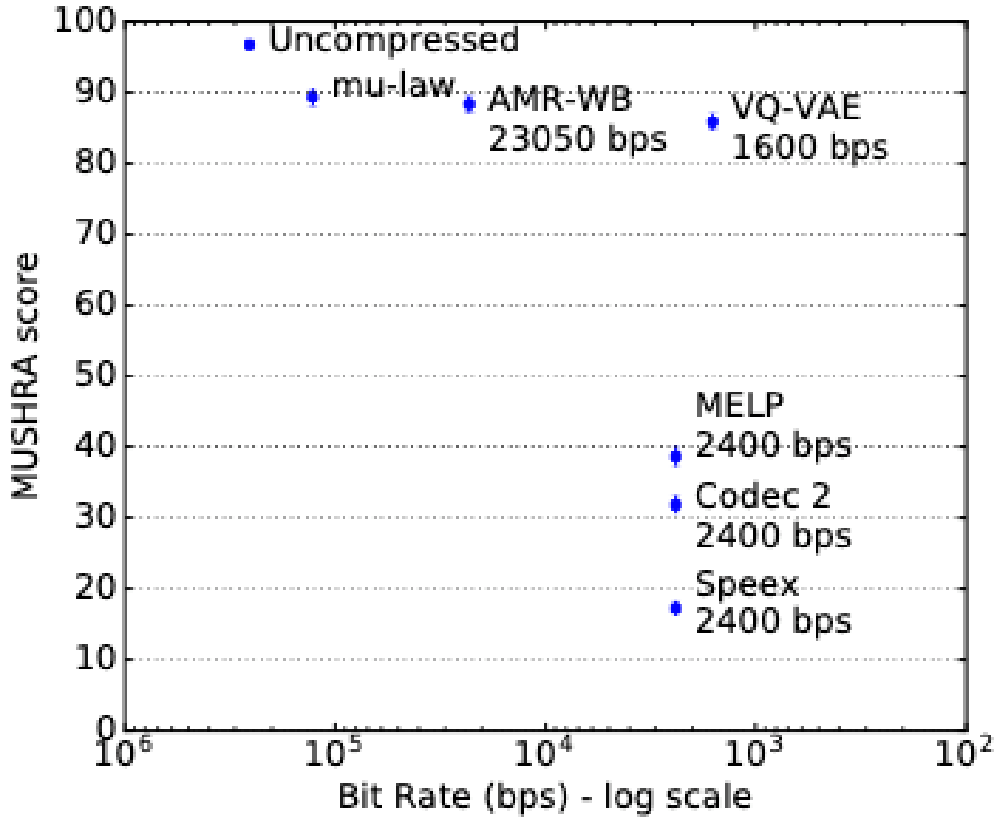
[1] Institute of Information and Communication, National University of Defense Technology, Xi'an, China
[2] Army Engineering University of PLA, Nanjing, China
{mgxaty, xwzhang}@gmail.com, {zhangcq1108, dzxxlab}@163.com



https://arxiv.org/pdf/1905.04709.pdf

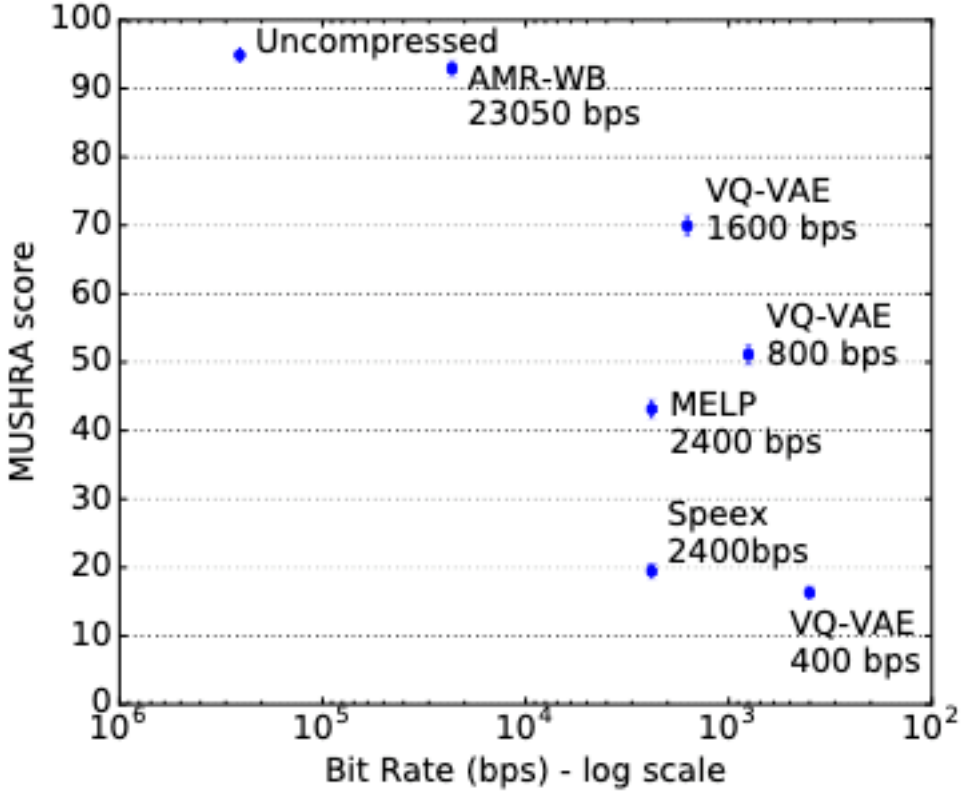
LOW BIT-RATE SPEECH CODING WITH VQ-VAE AND A WAVENET DECODER

single voice present in the train set



https://arxiv.org/pdf/1910.06464.pdf

LibriSpeech data



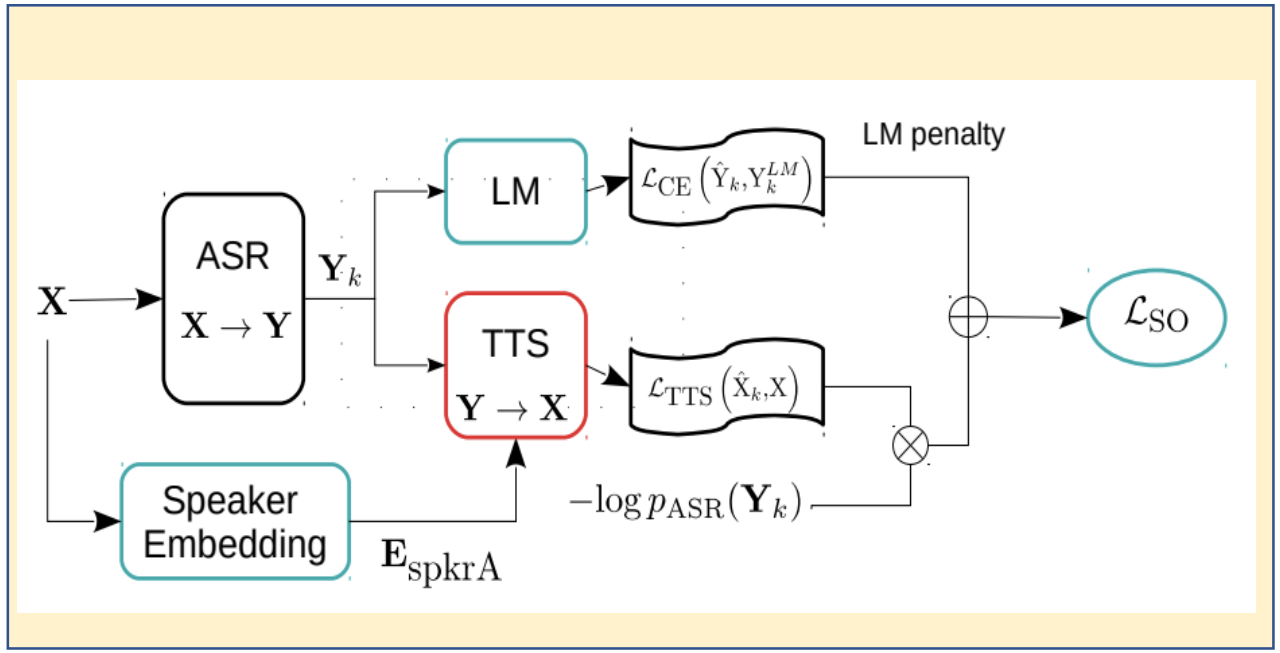
Google

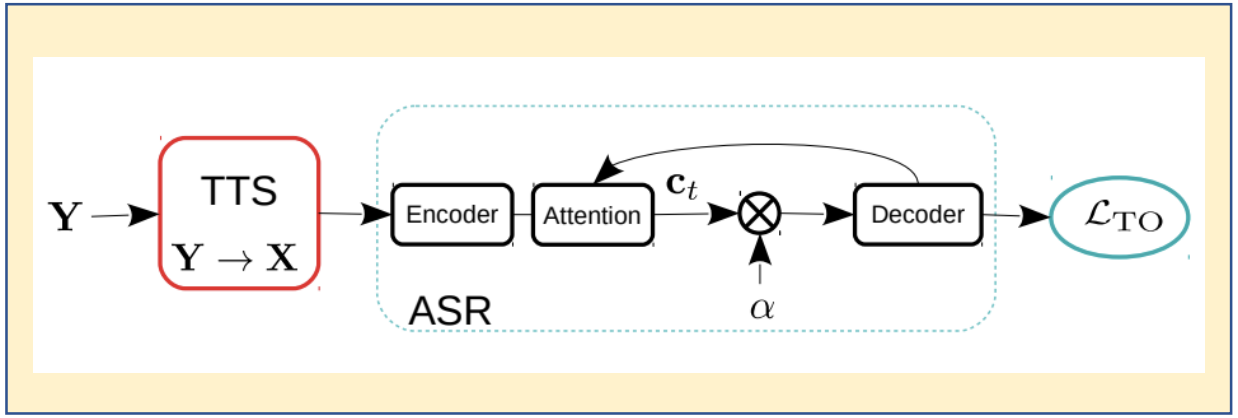
EAT: ENHANCED ASR-TTS FOR SELF-SUPERVISED SPEECH RECOGNITION

Murali Karthick Baskar[ϕ], *Lukáš Burget*[ϕ], *Shinji Watanabe*[†], *Ramon Fernandez Astudillo*[π], *and Jan "Honza" Černocký*[ϕ]

[ϕ]Brno University of Technology, [†] Johns Hopkins University, [π]IBM Research,

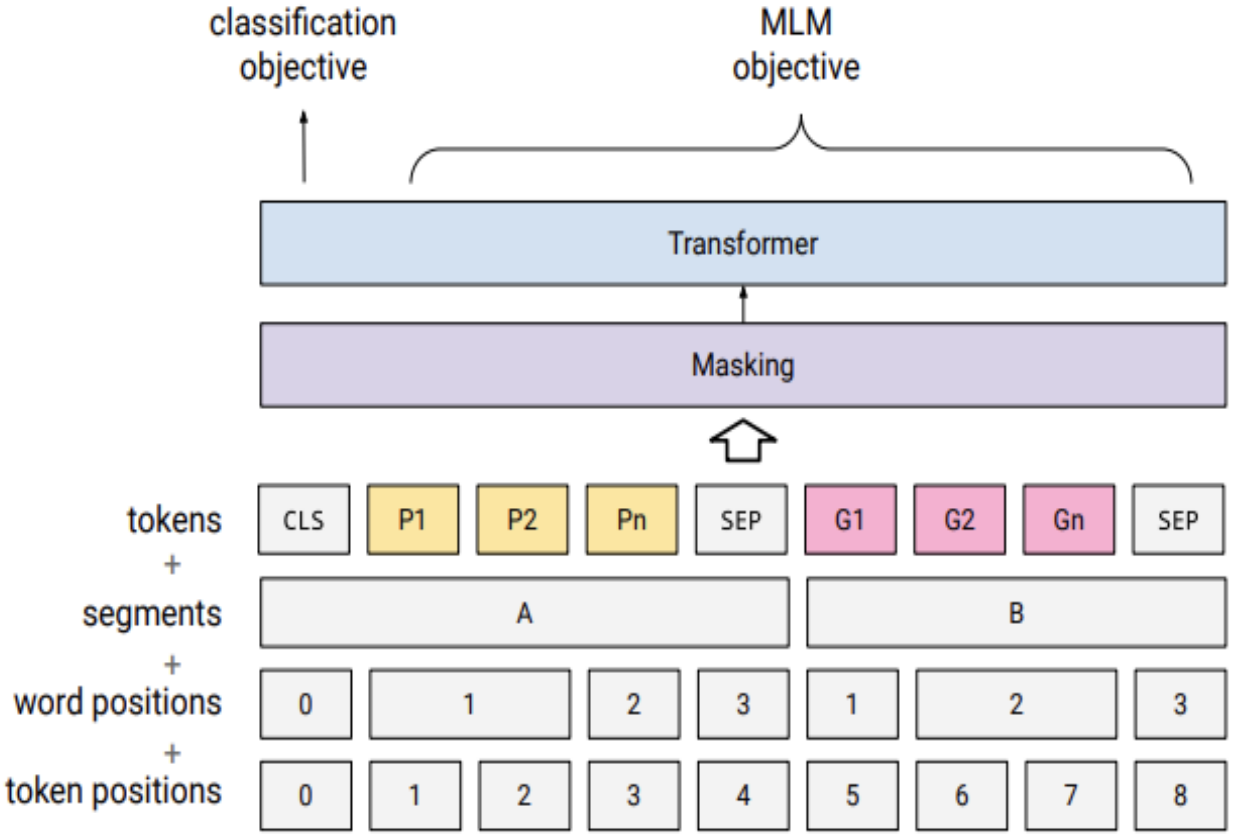
baskar@fit.vutbr.cz



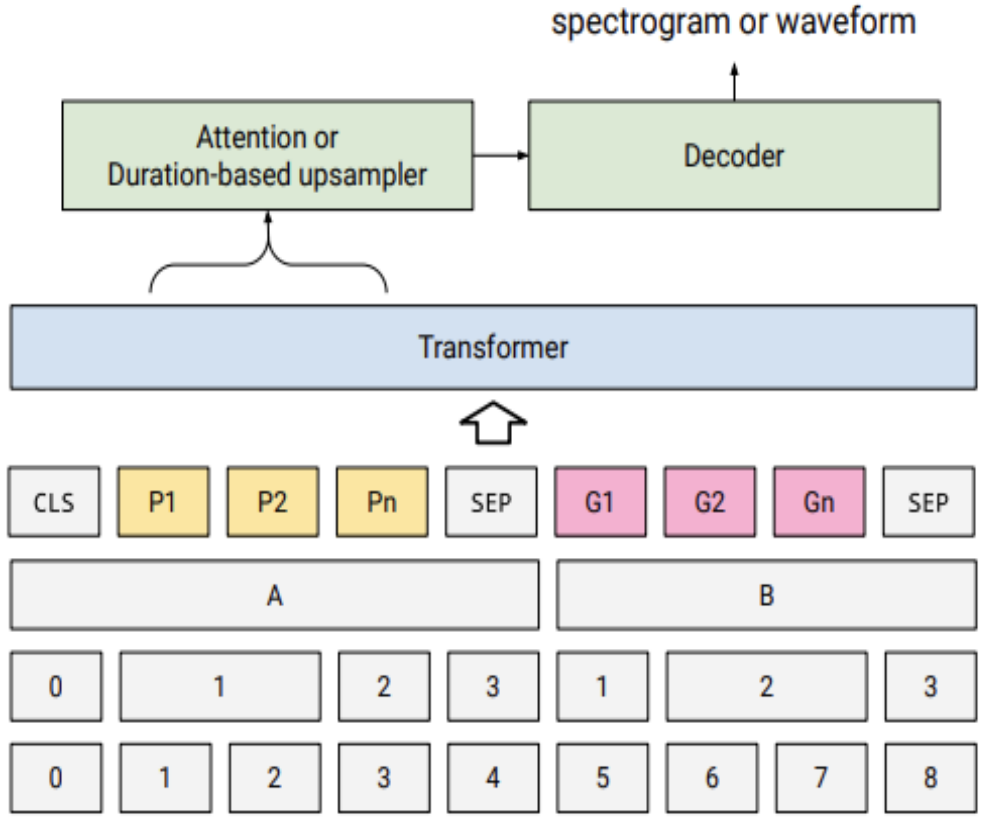


https://arxiv.org/pdf/2104.07474.pdf

PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS



(a) PnG BERT and pre-training objectives.

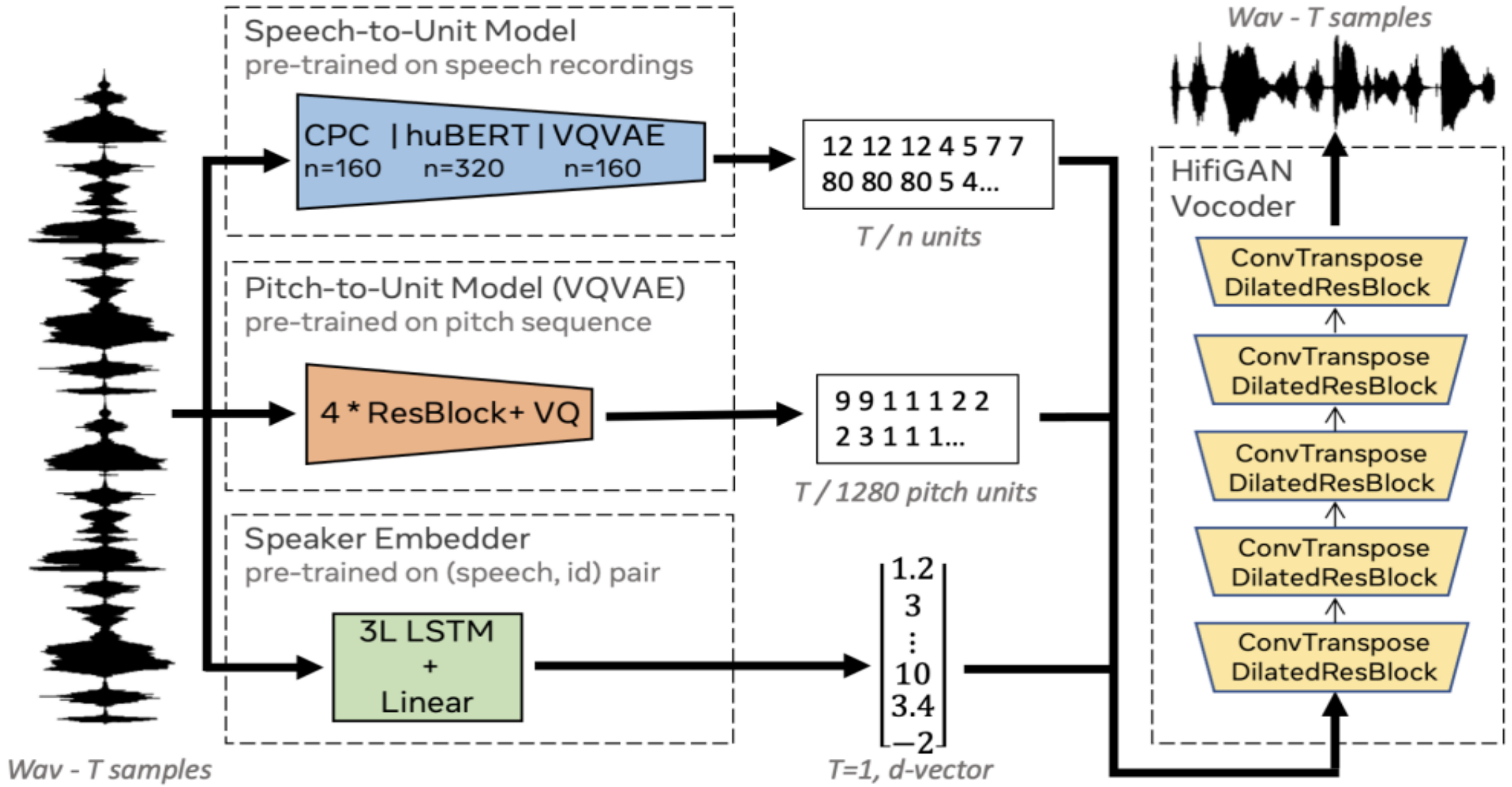


(b) Using PnG BERT as the encoder for a neural TTS model.

https://arxiv.org/pdf/2103.15060.pdf

Google Research

Speech Resynthesis from Discrete Disentangled Self-Supervised Representations



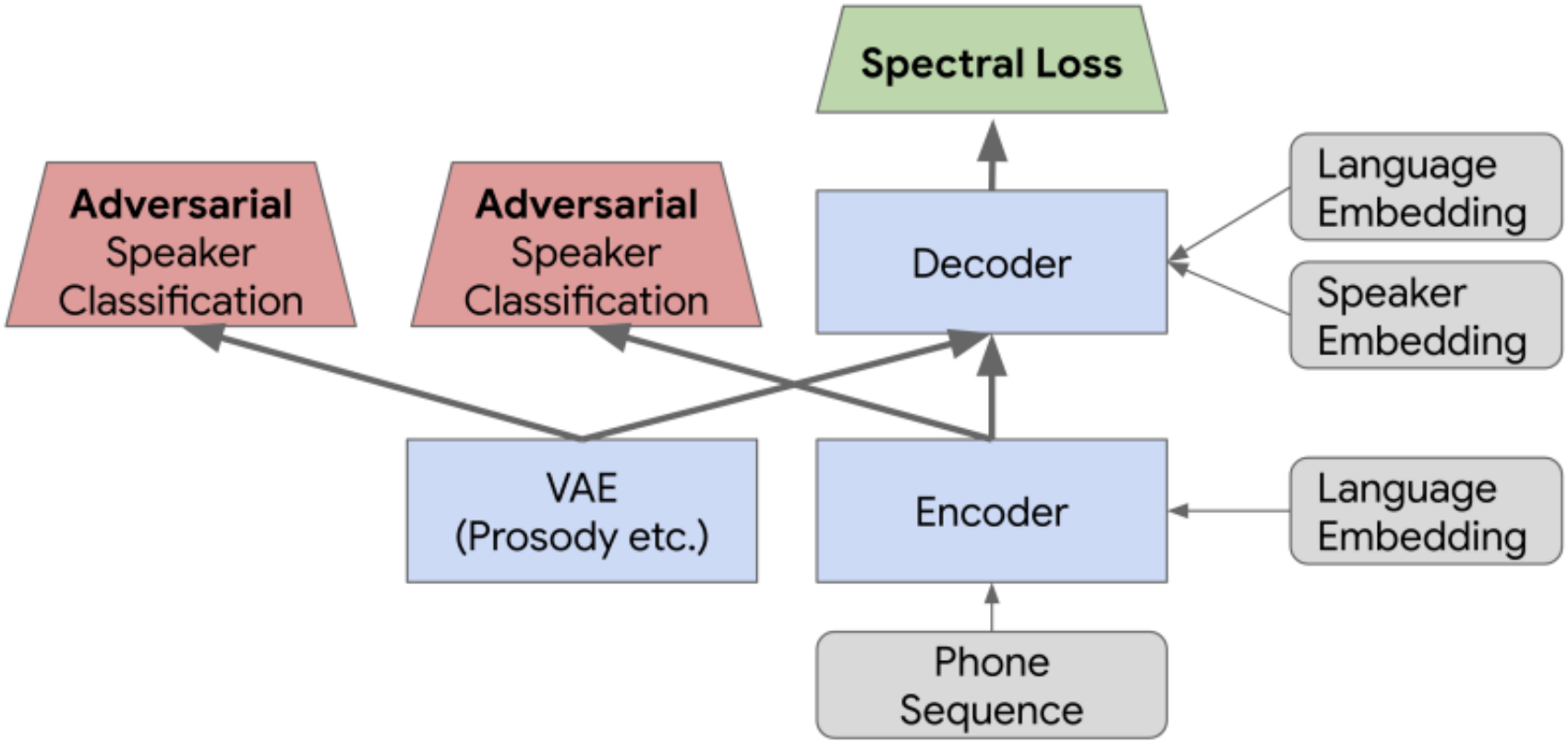
https://arxiv.org/pdf/2104.00355.pdf

Facebook AI Research

INJECTING TEXT IN SELF-SUPERVISED SPEECH PRETRAINING

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Gary Wang, Pedro Moreno

Google, Inc.



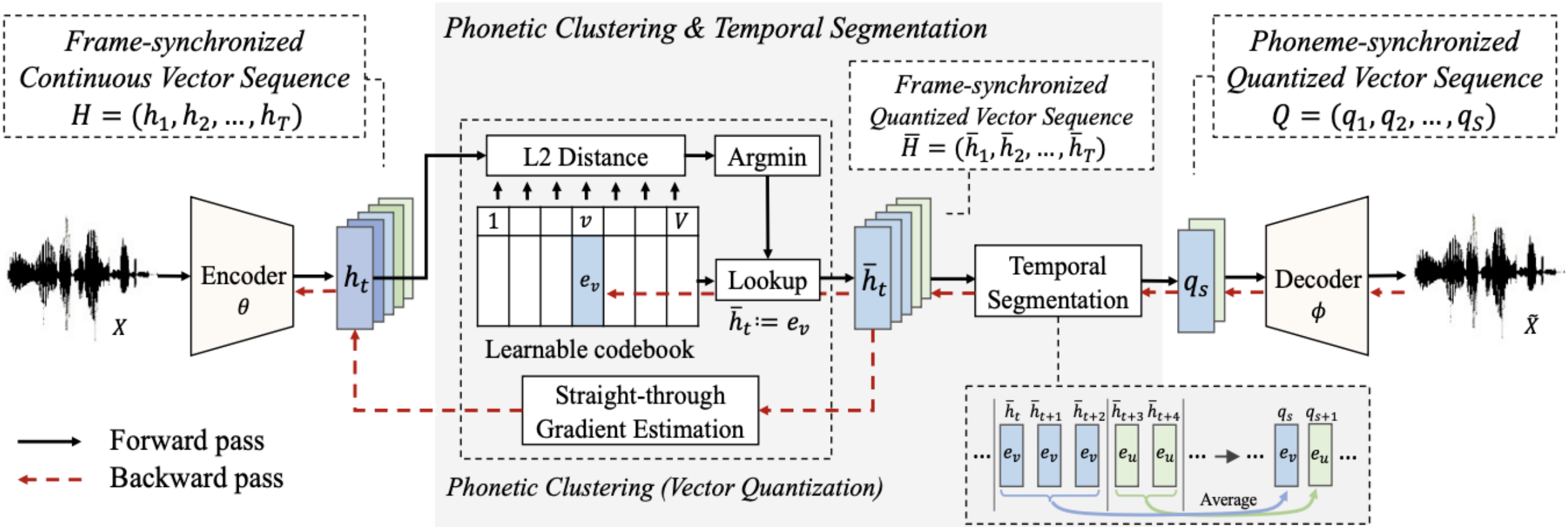
https://arxiv.org/pdf/2108.12226.pdf

TOWARDS UNSUPERVISED SPEECH RECOGNITION AND SYNTHESIS WITH QUANTIZED SPEECH REPRESENTATION LEARNING

Alexander H. Liu[†] *Tao Tu*[†] *Hung-yi Lee* *Lin-shan Lee*

College of Electrical Engineering and Computer Science, National Taiwan University

{r07922013, r07922022, hungyilee, lslee}@ntu.edu.tw



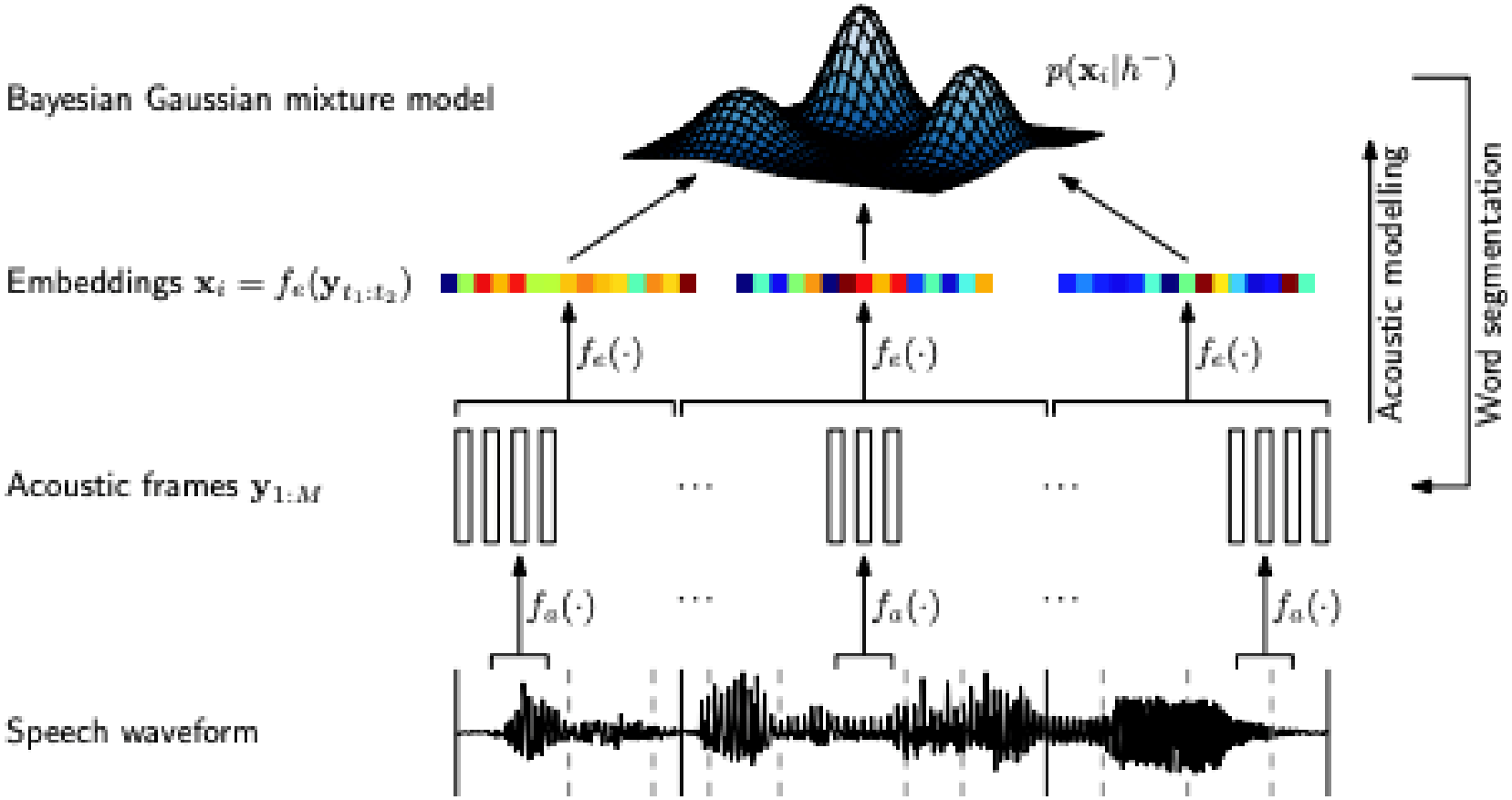
https://arxiv.org/pdf/1910.12729.pdf

A segmental framework for fully-unsupervised large-vocabulary speech recognition

Herman Kamper[1], Aren Jansen[2], Sharon Goldwater[1]

[1]School of Informatics, University of Edinburgh and [2]Google, Inc.

kamperh@gmail.com, arenjansen@google.com, sgwater@inf.ed.ac.uk

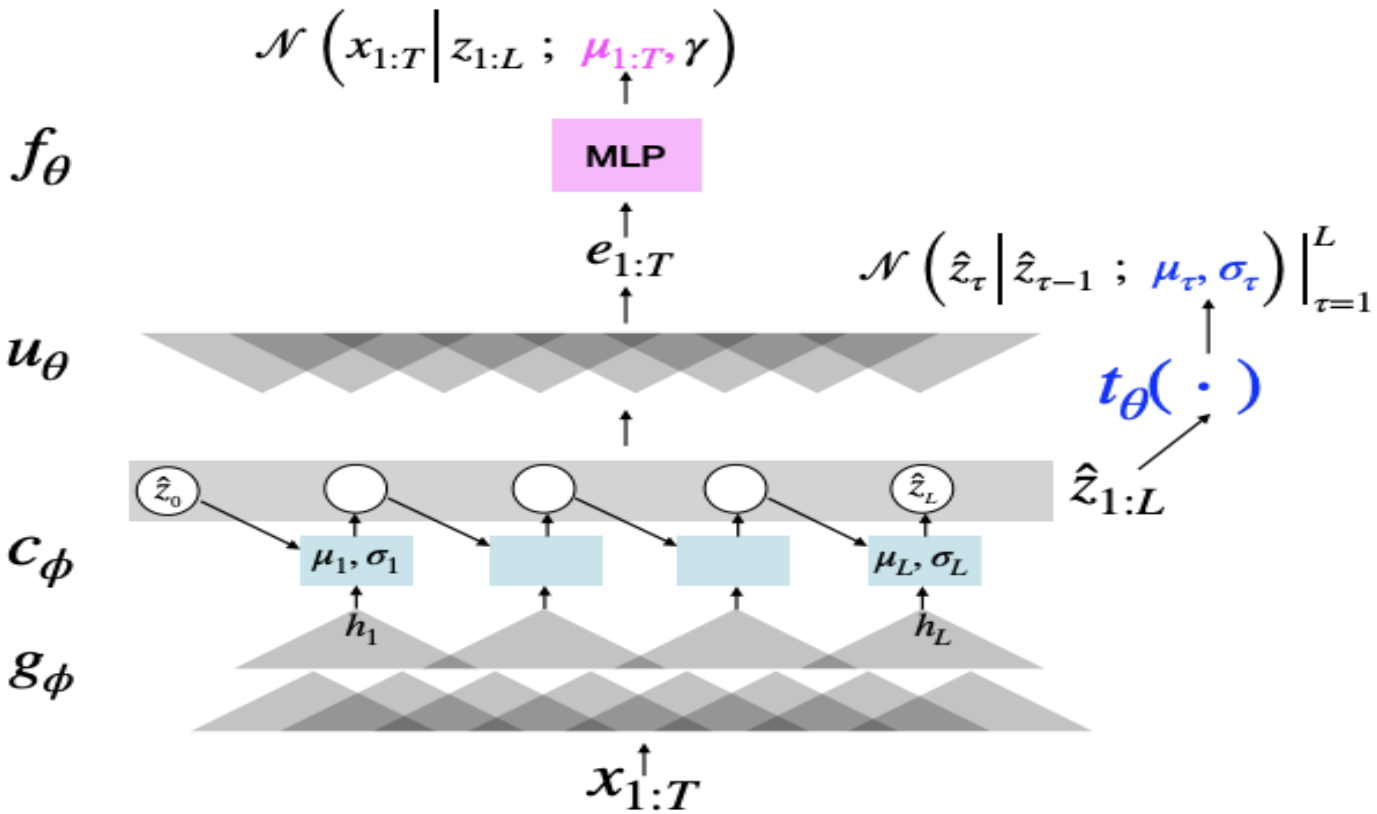


https://arxiv.org/pdf/1606.06950.pdf

A Convolutional Deep Markov Model for Unsupervised Speech Representation Learning

Sameer Khurana[1], *Antoine Laurent*[2], *Wei-Ning Hsu*[1], *Jan Chorowski*[3], *Adrian Lancucki*[4], *Ricard Marxer*[5], *James Glass*[1]

[1]MIT - CSAIL, Cambridge, USA
[2]LIUM - Le Mans University, France
[3]University of Wrocaw, Poland [4]NVIDIA Corporation, Poland [5]LIS - University of Toulon, France
skhurana@mit.edu, antoine.laurent@univ-lemans.fr



https://arxiv.org/pdf/2006.02547.pdf

SPEECH REPRESENTATION LEARNING THROUGH SELF-SUPERVISED PRETRAINING AND MULTI-TASK FINETUNING

Yi-Chen Chen[1], *Shu-wen Yang*[1], *Cheng-Kuang Lee*[2], *Simon See*[2], *Hung-yi Lee*[1]

[1]National Taiwan University, Taiwan [2]NVIDIA AI Technology Center, NVIDIA
[1]{f06942069,r08944041,hungyilee}@ntu.edu.tw [2]{ckl,ssee}@nvidia.com

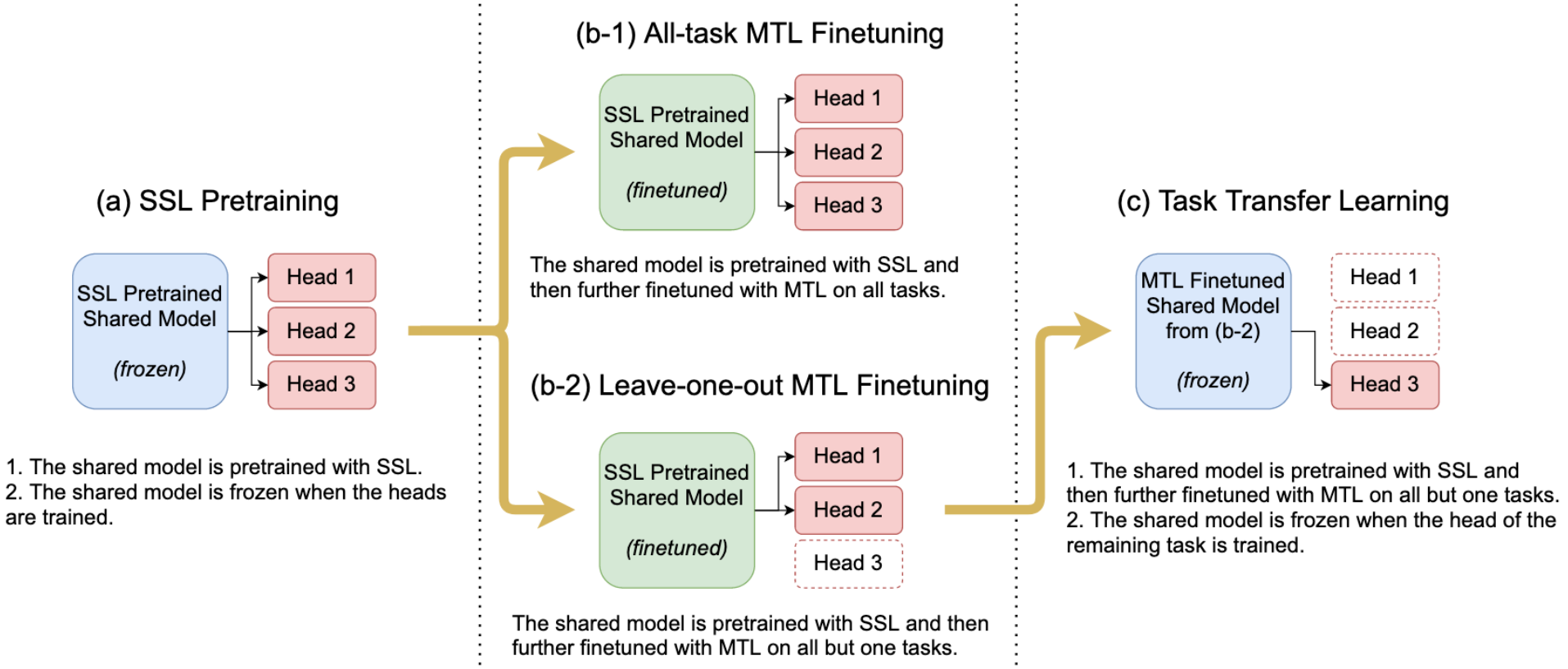


Fig. 1: Four different training scenarios. Scenarios (b-1) and (b-2) require the pretrained shared model from (a). Scenario (c) requires the finetuned shared model from (b-2). The parameters of the shared model in scenarios (a) and (c) are frozen.

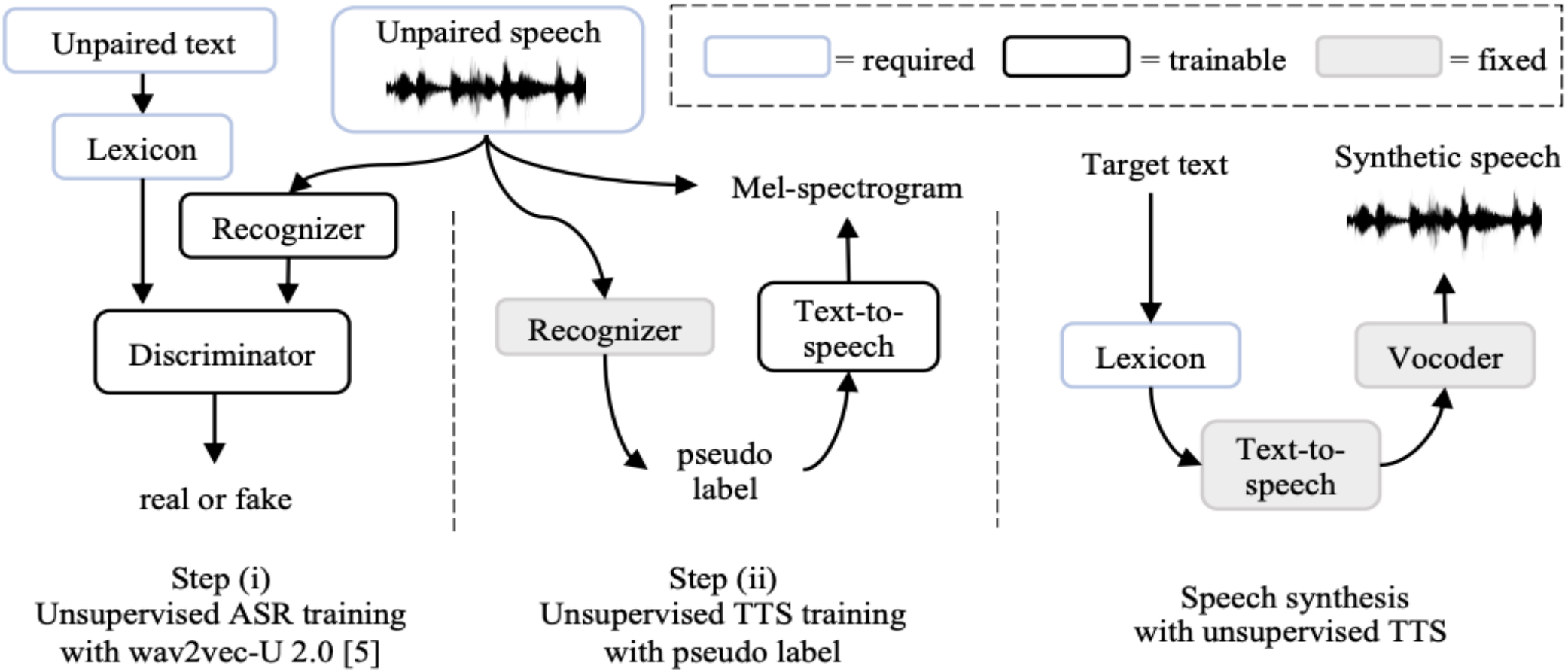
https://arxiv.org/pdf/2110.09930.pdf

Simple and Effective Unsupervised Speech Synthesis

Alexander H. Liu[*1], *Cheng-I Jeff Lai*[*1],
Wei-Ning Hsu[2], *Michael Auli*[2], *Alexei Baevski*[2], *James Glass*[1]

[1]MIT CSAIL [2]Meta AI

{alexhliu,clai24,glass}@mit.edu



https://arxiv.org/pdf/2204.02524.pdf

VQTTS: High-Fidelity Text-to-Speech Synthesis with Self-Supervised VQ Acoustic Feature

*Chenpeng Du, Yiwei Guo, Xie Chen, Kai Yu**

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{duchenpeng, cantabile_kwok, chenxie95, kai.yu}@sjtu.edu.cn

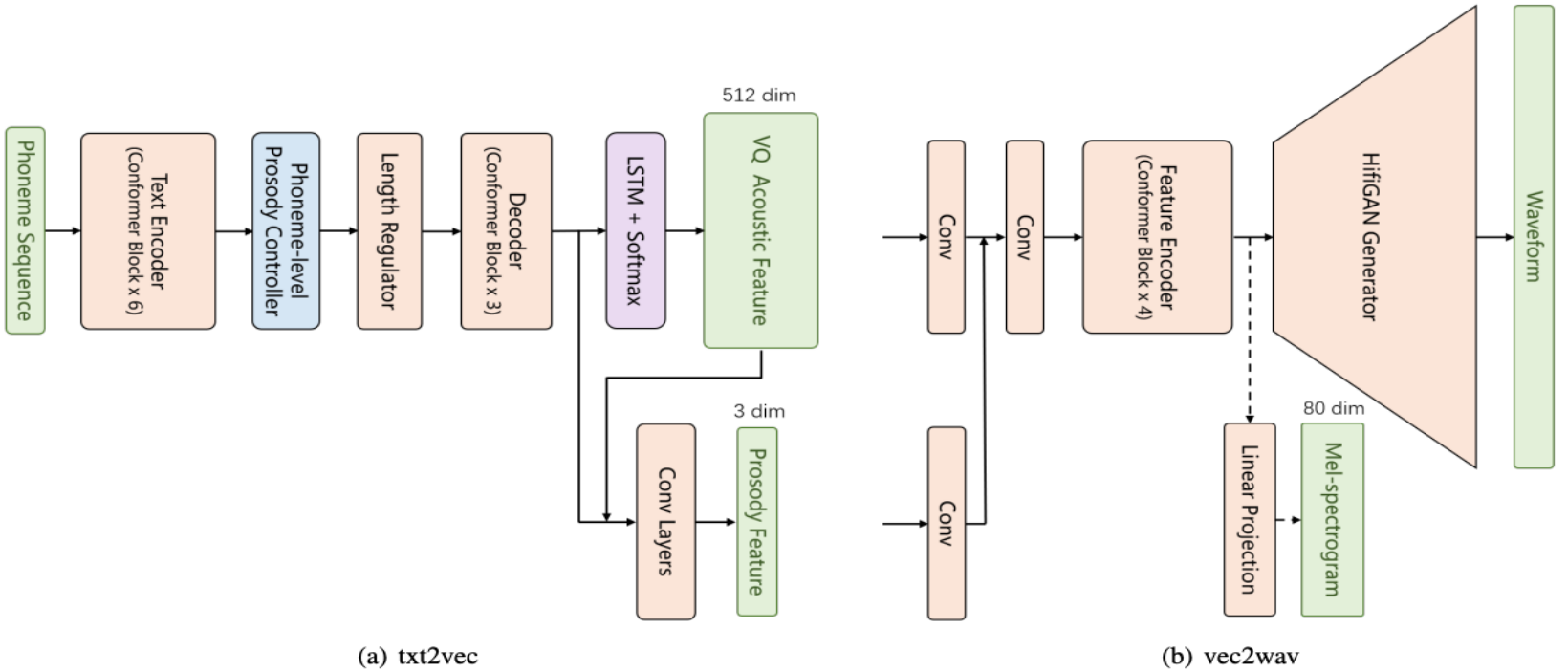


Figure 1: *Model architecture of VQTTS, consisting of txt2vec and vec2wav. The two parts are connected with VQ acoustic feature together with prosody feature.*

https://arxiv.org/pdf/2204.00768.pdf

Multi-modal SSL

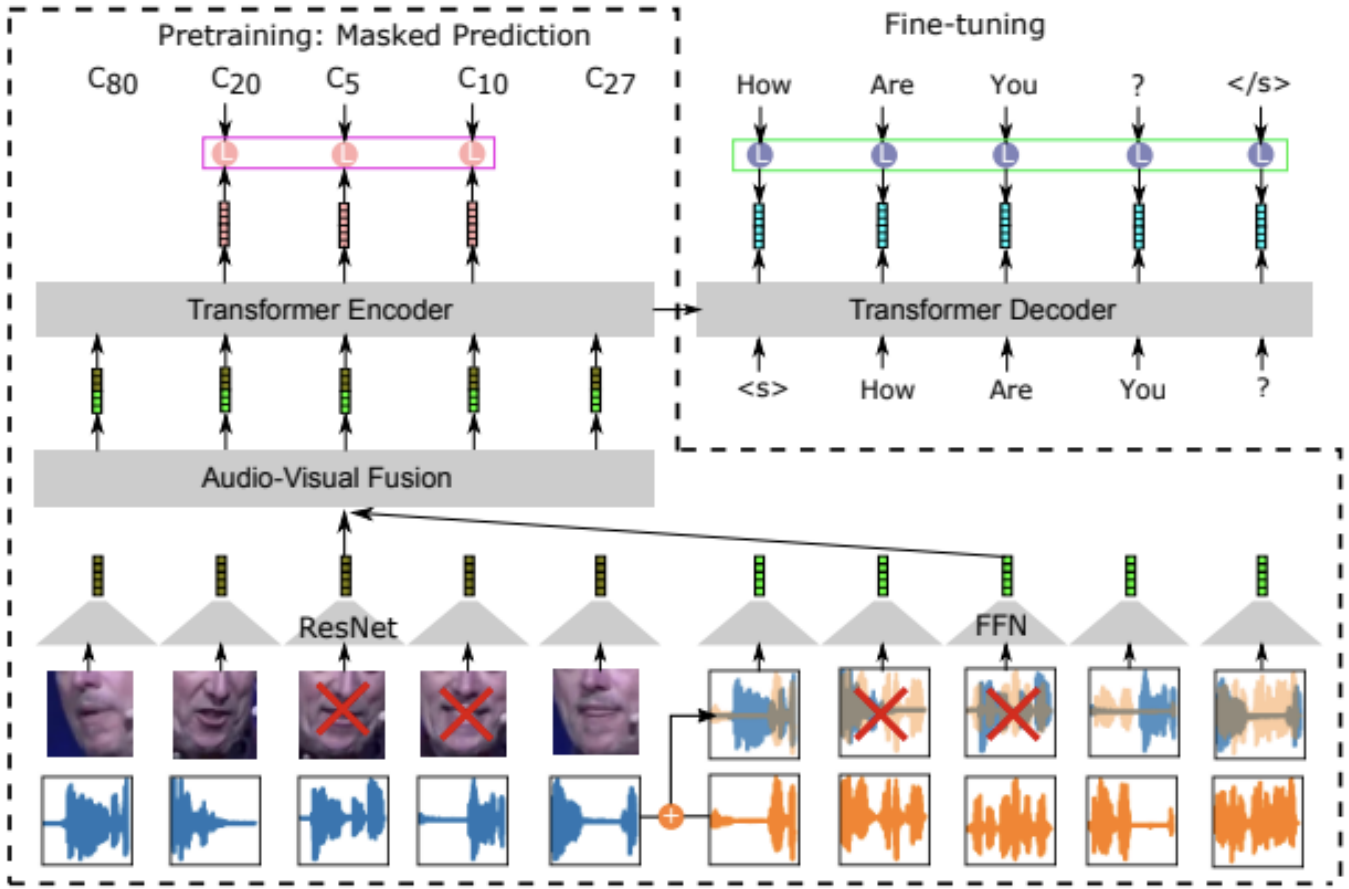
- Seeing a speaker's face
  - Improves the effective SNR by ~22 dB [Sumby & Pollack 1954]
  - Improves understanding of complex or accented speech [Reisberg et al. 1987]
- It's not just the speaker's face…
  - The visual scene informs about the content
- It's not just the visual modality…
  - Physical environment/touch
  - Situational context
  - Listener's internal state

Robust Self-Supervised Audio-Visual Speech Recognition

Bowen Shi[1*], *Wei-Ning Hsu*[2], *Abdelrahman Mohamed*[2]

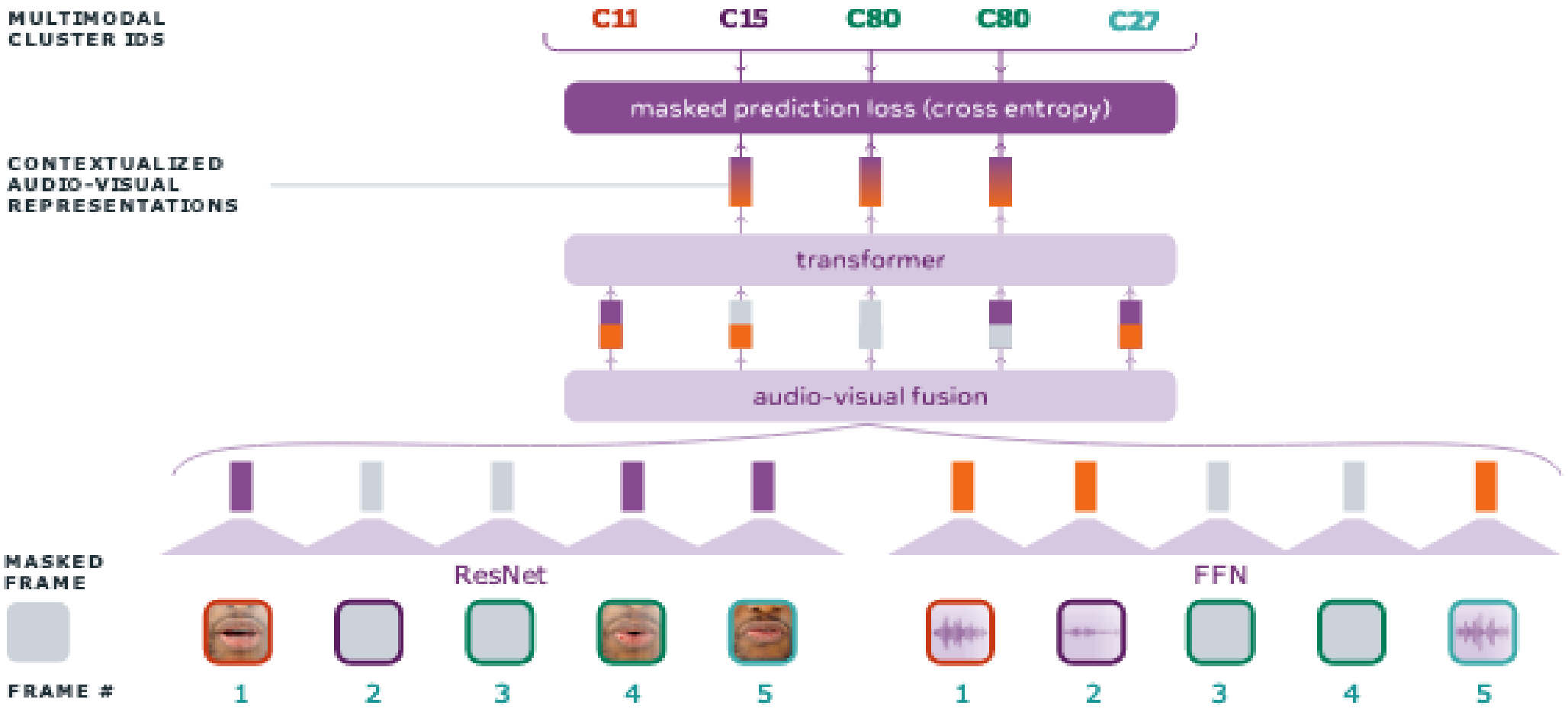
[1]Toyota Technological Institute at Chicago
[2]Meta AI

bshi@ttic.edu, wnhsu@fb.com, abdo@fb.com



https://arxiv.org/pdf/2201.01763.pdf

Learning Audio-Visual Speech Representation By Masked Multimodal Cluster Prediction



Some conclusions

- Self-Supervised Learning (SSL) is a simple solution to train models which are able to solve problems for which one is lacking labeled data.
- SSL is particularly suitable in (un-supervised/semi-supervised) situations for which large amounts of non-annotated data is available.
- SSL produces speech representations suitable for coding, ASR, TTS, speaker recognition, diarization, emotion detection, SLU,…
- SSL should be bundled (fused) in hybrid solutions with existing (traditional, modular) and symbolic AI techniques.

Self-Supervised Speech Processing

Gérard Chollet (CNRS-SAMOVAR)

Yingzhi Wang (Zaion)

gerard.chollet@telecom-sudparis.eu

ywang@zaion.ai

Intelligent Voice® zaion Une école Institut Mines-Télécom TELECOM SudParis cnrs Depuis 80 ans, nos connaissances bâtissent de nouveaux mondes speechmorphing

BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition

Yu Zhang*, Daniel S. Park*, Wei Han*,
James Qin, Anmol Gulati, Joel Shor, Aren Jansen,
Yuanzhong Xu, Yanping Huang, Shibo Wang, Zongwei Zhou,
Bo Li, Min Ma, William Chan, Jiahui Yu, Yongqiang Wang,
Liangliang Cao, Khe Chai Sim, Bhuvana Ramabhadran,
Tara N. Sainath, Françoise Beaufays, Zhifeng Chen, Quoc V. Le, Chung-Cheng Chiu,
Ruoming Pang† and Yonghui Wu

https://arxiv.org/pdf/2109.13226.pdf

UNSUPERVISED TRAINING OF A SPEECH RECOGNIZER USING TV BROADCASTS

Thomas Kemp *Alex Waibel*

Interactive Systems Laboratories, ILKD
University of Karlsruhe
76128 Karlsruhe, Germany

https://www.isca-speech.org/archive/pdfs/icslp_1998/kemp98b_icslp.pdf

UNSUPERVISED TRAINING OF A SPEECH RECOGNIZER: RECENT EXPERIMENTS

Thomas Kemp *Alex Waibel*

Interactive Systems Laboratories, ILKD
University of Karlsruhe
76128 Karlsruhe, Germany

https://isl.anthropomatik.kit.edu/pdf/Kemp1999.pdf

UNSUPERVISED ACOUSTIC MODEL TRAINING USING MULTIPLE SEED ASR SYSTEMS

Horia Cucu, Andi Buzo and Corneliu Burileanu

Speech and Dialogue (SpeeD) Research Laboratory,
University "Politehnica" of Bucharest, Romania
{horia.cucu, andi.buzo, corneliu.burileanu}@upb.ro

http://speed.pub.ro/speed3/wp-content/uploads/2015/07/2014_SLTU_unsupervisedAM.pdf

High level feature extraction for the self-taught learning algorithm

Konstantin Markov[1*] and Tomoko Matsui[2]

https://asmp-eurasipjournals.springeropen.com/track/pdf/10.1186/1687-4722-2013-6.pdf

Gaussian Map based Acoustic Model Adaptation Using Untranscribed Data for Speech Recognition in Severely Adverse Environments

Wooil Kim and John H. L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas, USA
{wikim,John.Hansen}@utdallas.edu, http://crss.utdallas.edu

https://www.isca-speech.org/archive/pdfs/interspeech_2012/kim12f_interspeech.pdf

SEMI-SUPERVISED TRAINING OF DEEP NEURAL NETWORKS

Karel Veselý, Mirko Hannemann, Lukáš Burget

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czech Republic

https://www.fit.vutbr.cz/research/groups/speech/publi/2013/vesely_asru2013_0000267.pdf

SEMI-SUPERVISED TRAINING IN LOW-RESOURCE ASR AND KWS

*Florian Metze**, *Ankur Gandhe**, *Yajie Miao**, *Zaid Sheikh**, *Yun Wang**, *Di Xu**, *Hao Zhang**,
Jungsuk Kim†, *Ian Lane*†, *Won Kyum Lee*†, *Sebastian Stüker*‡, *and Markus Müller*‡

* Language Technologies Institute, † Department of Electrical and Computer Engineering
Carnegie Mellon University; Pittsburgh, PA/ Moffett Field, CA; U.S.A
‡ Karlsruhe Institute of Technology; Karlsruhe; Germany
fmetze@cs.cmu.edu

https://www.cs.cmu.edu/~fmetze/interACT/Publications_files/publications/babel-op1-v2.pdf

SEMI-SUPERVISED SPEECH RECOGNITION VIA GRAPH-BASED TEMPORAL CLASSIFICATION

Niko Moritz, Takaaki Hori, Jonathan Le Roux

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

https://arxiv.org/pdf/2010.15653.pdf

TRAINING LVCSR SYSTEMS ON THOUSANDS OF HOURS OF DATA

G. Evermann, H.Y. Chan, M.J.F. Gales, B. Jia, D. Mrva, P.C. Woodland, K. Yu*

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
Email: ge204@eng.cam.ac.uk

https://www.yumpu.com/en/document/read/19773522/training-lvcsr-systems-on-thousands-of-hours-of-data

An Improved Method for Unsupervised Training of LVCSR Systems

Christian Gollan, Stefan Hahn, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{gollan,hahn,schluter,ney}@cs.rwth-aachen.de

https://www.isca-speech.org/archive/pdfs/interspeech_2007/gollan07_interspeech.pdf

Contextual Semi-Supervised Learning: An Approach To Leverage Air-Surveillance and Untranscribed ATC Data in ASR Systems

Juan Zuluaga-Gomez[1,2], *Iuliia Nigmatulina*[1], *Amrutha Prasad*[1,3], *Petr Motlicek*[1], *Karel Veselý*[3], *Martin Kocour*[3], *Igor Szöke*[4]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
[3]Brno University of Technology, Speech@FIT, IT4I CoE, Brno, Czech Republic
[4]ReplayWell, Brno, Czech Republic

{juan-pablo.zuluaga,iuliia.nigmatulina,aprasad,petr.motlicek}@idiap.ch,
{iveselyk,ikocour}@fit.vutbr.cz, szoke@replaywell.com

https://arxiv.org/pdf/2104.03643.pdf

Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System

Jonas Lööf, Christian Gollan, and Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Dept.
RWTH Aachen University, Aachen, Germany
{loof,gollan,ney}@cs.rwth-aachen.de

https://www.isca-speech.org/archive/pdfs/interspeech_2009/loof09_interspeech.pdf

Modeling and Transforming Speech using Variational Autoencoders

Merlijn Blaauw, Jordi Bonada

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

merlijn.blaauw@upf.edu, jordi.bonada@upf.edu

https://www.isca-speech.org/archive/pdfs/interspeech_2016/blaauw16_interspeech.pdf

SEMI-SUPERVISED ENSEMBLE DNN ACOUSTIC MODEL TRAINING

Sheng Li[1], *Xugang Lu*[2], *Shinsuke Sakai*[1], *Masato Mimura*[1] *and Tatsuya Kawahara*[1]

[1] School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[2] National Institute of Information and Communications Technology, Kyoto, Japan

lisheng@sap.ist.i.kyoto-u.ac.jp

http://sap.ist.i.kyoto-u.ac.jp/EN/bib/intl/LI-ICASSP17.pdf

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016,
9-12 May 2016, Yogyakarta, Indonesia

Variational Inference for Acoustic Unit Discovery

Lucas Ondel*, Lukaš Burget, Jan Černocký

Brno University of Technology, Czech Republic

https://www.researchgate.net/publication/301827883_Variational_Inference_for_Acoustic_Unit_Discovery/fulltext/572ce36f08ae7441518e6772/Variational-Inference-for-Acoustic-Unit-Discovery.pdf?origin=publication_detail