Speech representation learning and the emergence of Textless NLP research

Abdelrahman Mohamed

Meta, Fundamental AI Research (FAIR)

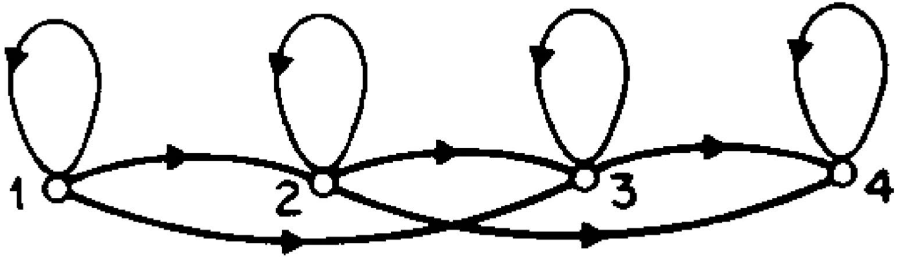
Outline

- 3 waves of speech representation learning
- SOTA Speech SSL methods
- Generative Spoken LMs
- Privacy-preserving Speech SSL

3 waves of speech representation learning

1st Wave: Clustering and Mixture Models

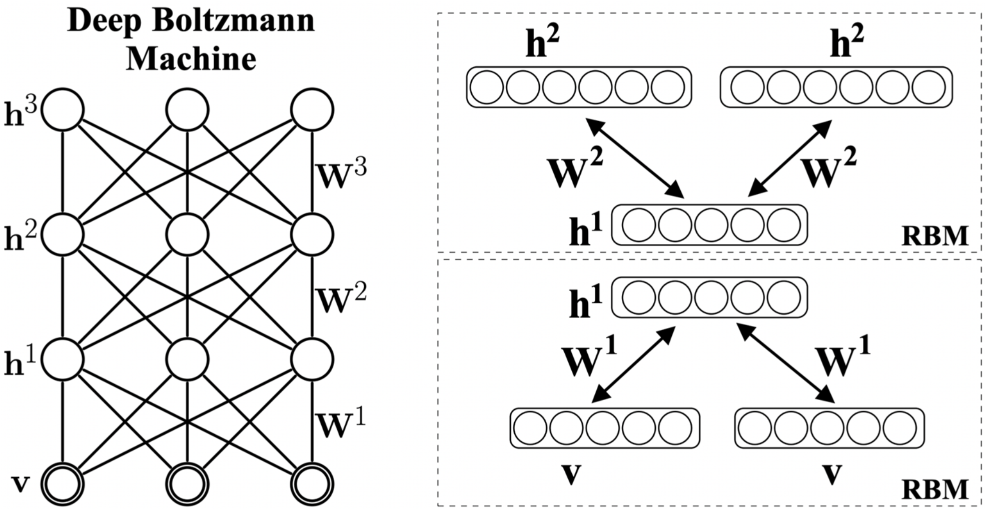
- Early works used simple clustering methods
- Gaussian Mixture Models / Hidden Markov models (GMMs/HMMs)
- Extracting features from generative models



From Rabiner 1989

2nd Wave: Stacked Neural Models

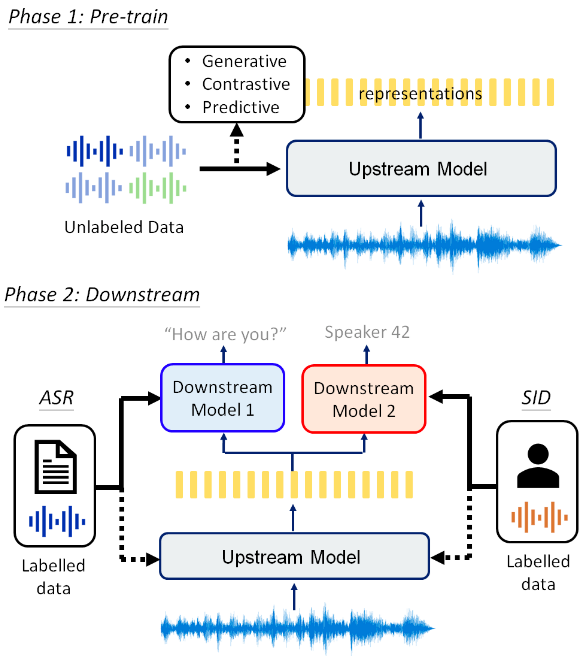
- Neural models have higher capacity for modeling inputs.
- Techniques include:
    - Restricted Boltzmann Machines (RBM)
    - Denoising AutoEncoders (DAE)
    - Noise Contrastive Estimation (NCE)
    - Sparse coding.
- Higher capacity networks achieved by building ‘deep’ networks with multiple layers of representations.



From Salakhutdinov and Hinton 2009

3rd Wave: Learning Through Pre-text Task Optimization

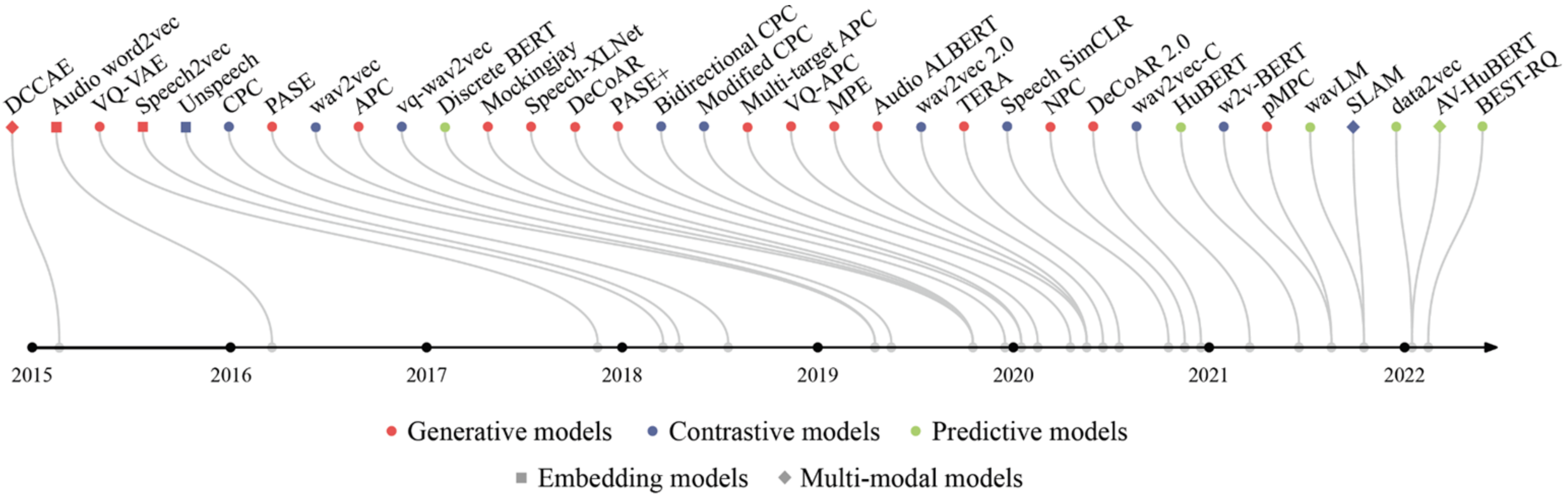
- Learn networks to map the input to desired representations by solving a *pre-text task*, with the following characteristics:
  - All layers are trained end-to-end to optimize a single pre-text task
  - Deep networks with many layers are used
  - The representation model is evaluated on many tasks
- The third wave looks at designing a pre-text task, which allows the model to efficiently use knowledge from unlabeled data.
  - Generate an object from partial information
  - Use previous tokens in the sentence to predict the next token
  - Contrastive learning



From Mohamed et. al 2022

Self-Supervised Speech Representations

Speech representation learning methods



Speech representation learning methods

Contrastive approaches

Speech representation learning methods

Contrastive approaches

Predictive approaches

Speech representation learning methods

Contrastive approaches

Predictive approaches

Generative approaches

wav2vec 2.0

wav2vec 2.0

- The first approach to show significant improvements for low-resource ASR.

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0

- The first approach to show significant improvements for low-resource ASR.
- Impressive results on multilingual representations.

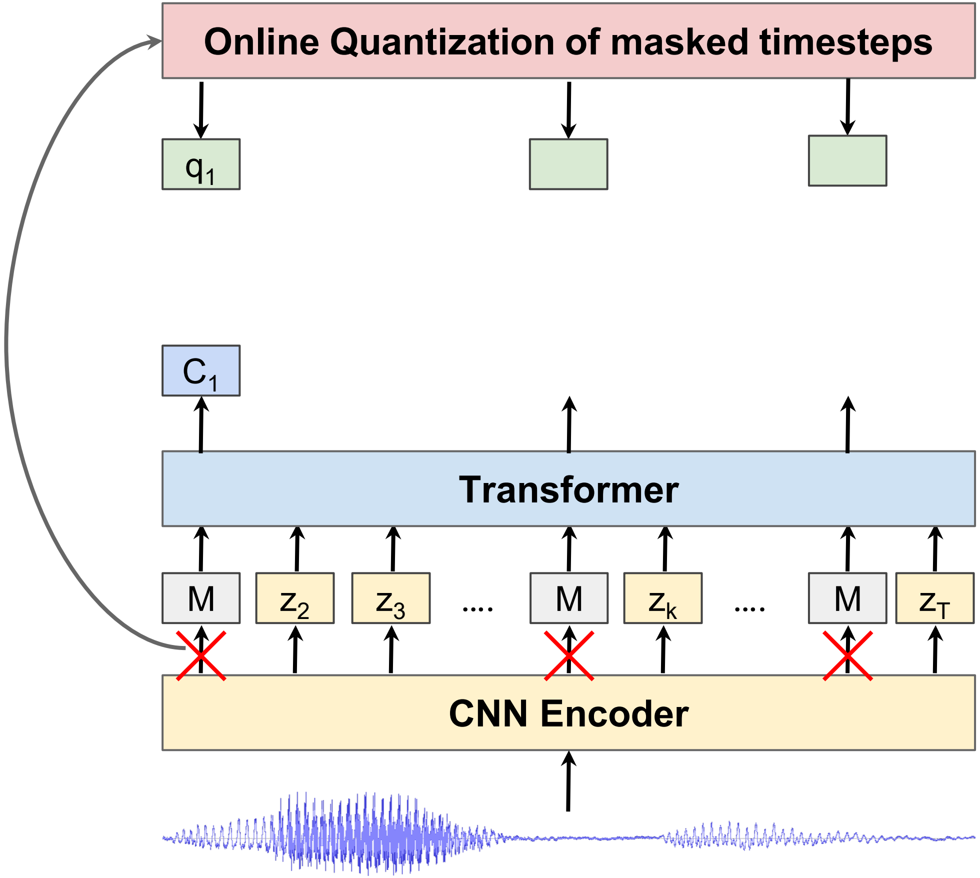
Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0

- 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.

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

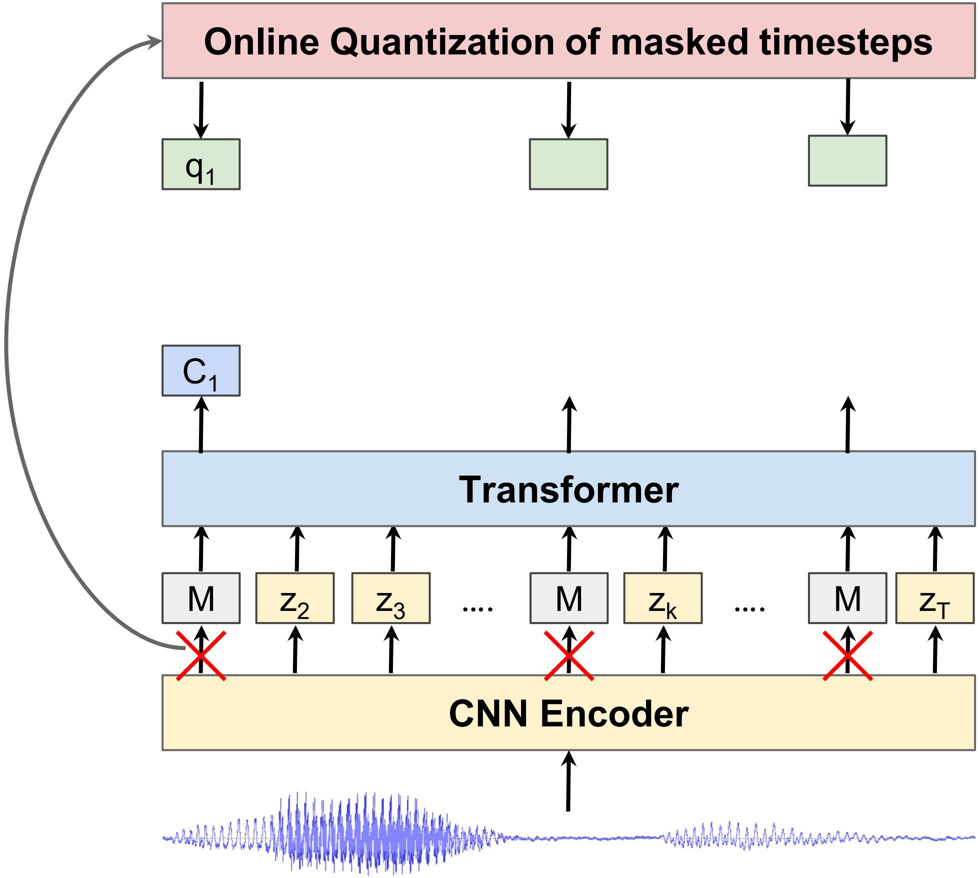
wav2vec 2.0: The pretext task



Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: The pretext task

- The goal is to maximize the similarity between the learned contextual representation and the quantized input features at the same position.

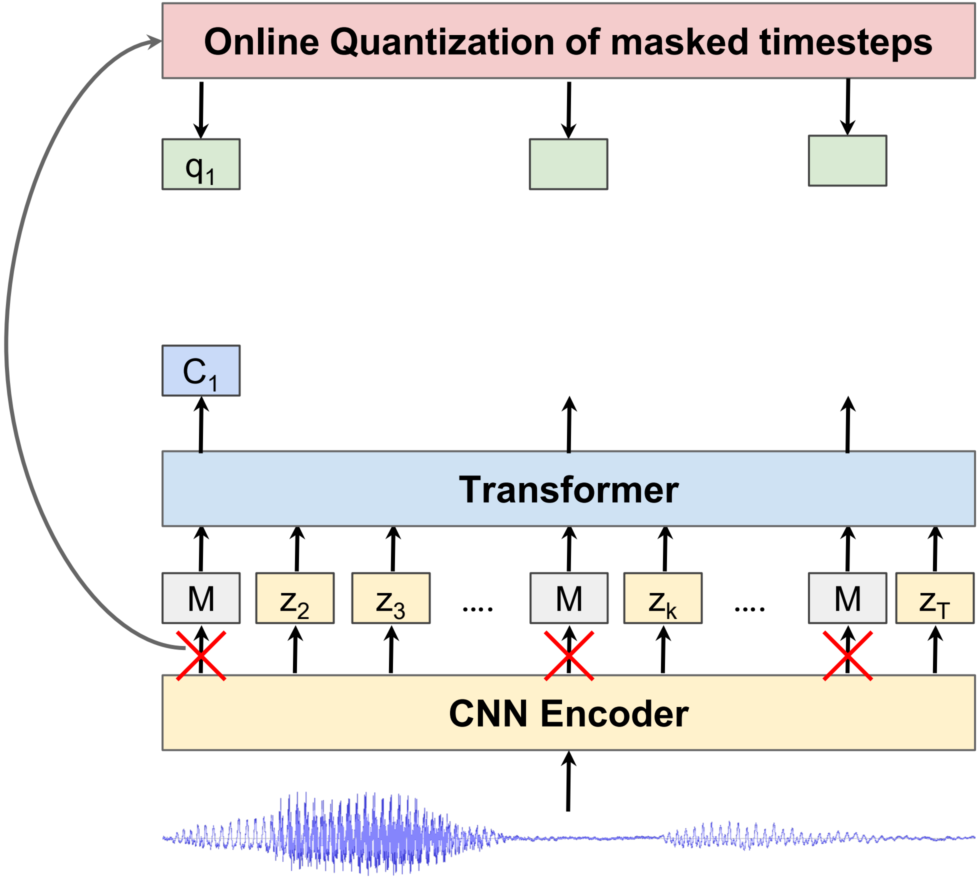


Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: The pretext task

- The goal is to maximize the similarity between the learned contextual representation and the quantized input features at the same position.

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



Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Results

- The first approach to get into single-digit WER on Librispeech test-other using only 10 mins of labels.

Model	Unlabeled data	LM	dev clean	dev other	test clean	test 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

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Results

- It is the first self-supervised approach to produce competitive results compared to semi-supervised learning approaches.

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

wav2vec 2.0: Results

- It is th
comp

Model	Unlabeled data	LM	dev clean	dev other	test clean	test other
Supervised						
CTC Transf [51]	-	CLM+Transf.	2.20	4.94	2.47	5.45
S2S Transf. [51]	-	CLM+Transf.	2.10	4.79	2.33	5.17
Transf. Transducer [60]	-	Transf.	-	-	2.0	4.6
ContextNet [17]	-	LSTM	1.9	3.9	1.9	4.1
Conformer [15]	-	LSTM	2.1	4.3	1.9	3.9
Semi-supervised						
CTC Transf. + PL [51]	LV-60k	CLM+Transf.	2.10	4.79	2.33	4.54
S2S Transf. + PL [51]	LV-60k	CLM+Transf.	2.00	3.65	2.09	4.11
Iter. pseudo-labeling [58]	LV-60k	4-gram+Transf.	1.85	3.26	2.10	4.01
Noisy student [42]	LV-60k	LSTM	1.6	3.4	1.7	3.4
This work						
LARGE - from scratch	-	Transf.	1.7	4.3	2.1	4.6
BASE	LS-960	Transf.	1.8	4.7	2.1	4.8
LARGE	LS-960	Transf.	1.7	3.9	2.0	4.1
	LV-60k	Transf.	1.6	3.0	1.8	3.3

Baevski et al, 2020 “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”

Hidden Unit BERT (HuBERT)

HuBERT

- A simple method to apply BERT style representation learning for speech.

Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT

- 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.

Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

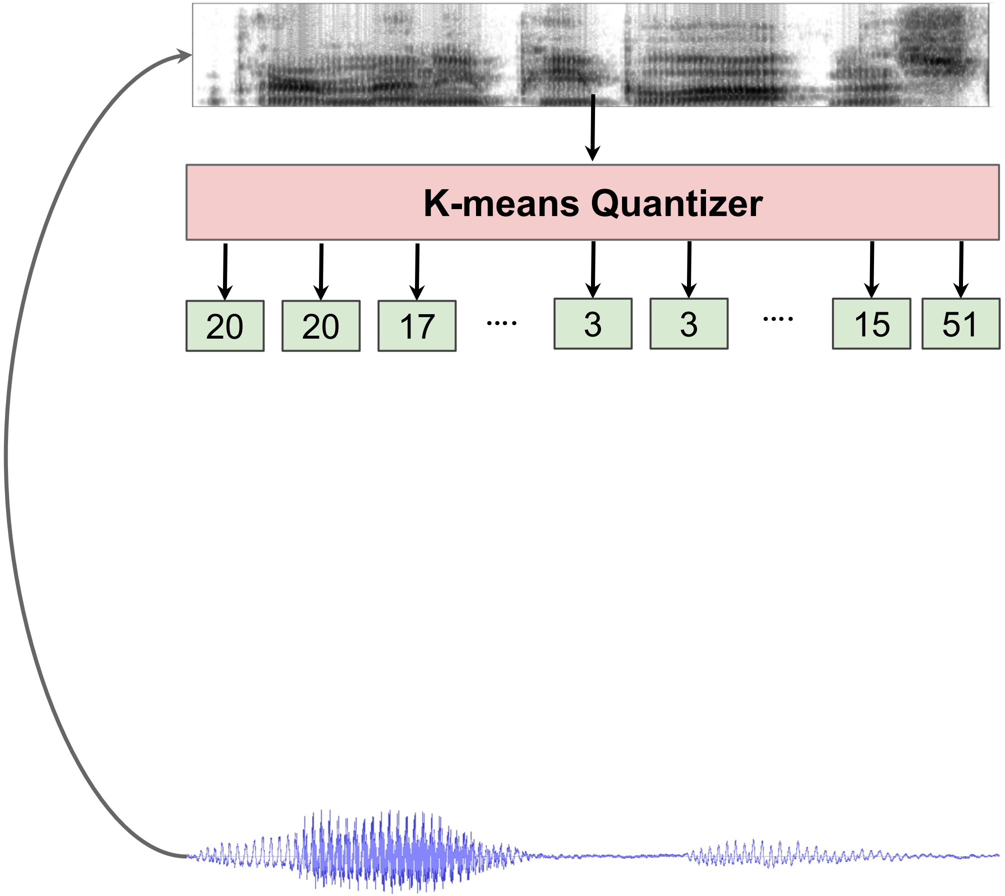
HuBERT

- 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.
- With its high-quality discrete units, HuBERT facilitated Textless NLP research.

Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task

- The K-means quantizer produces frame-level labels.

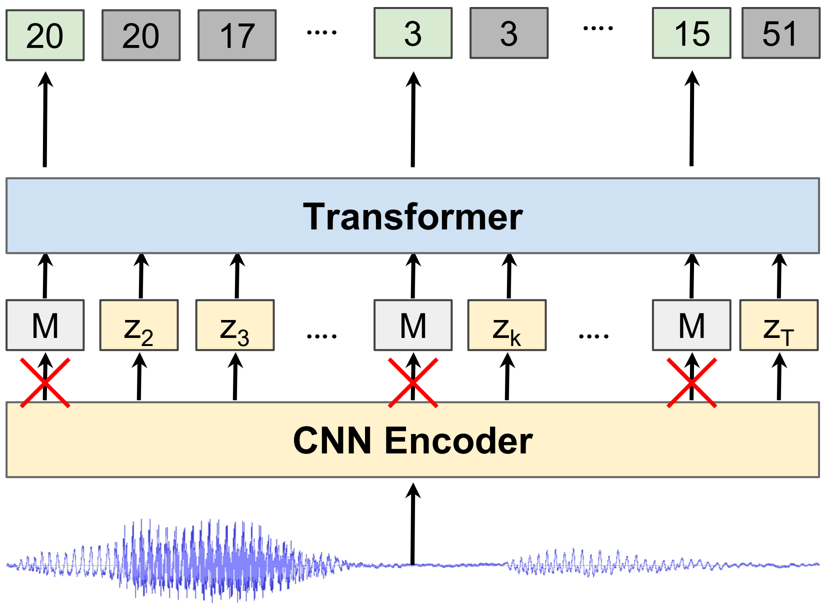


Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task

- Although the frame labels are imperfect, their consistency is more important!
- The model is trained using masked prediction:

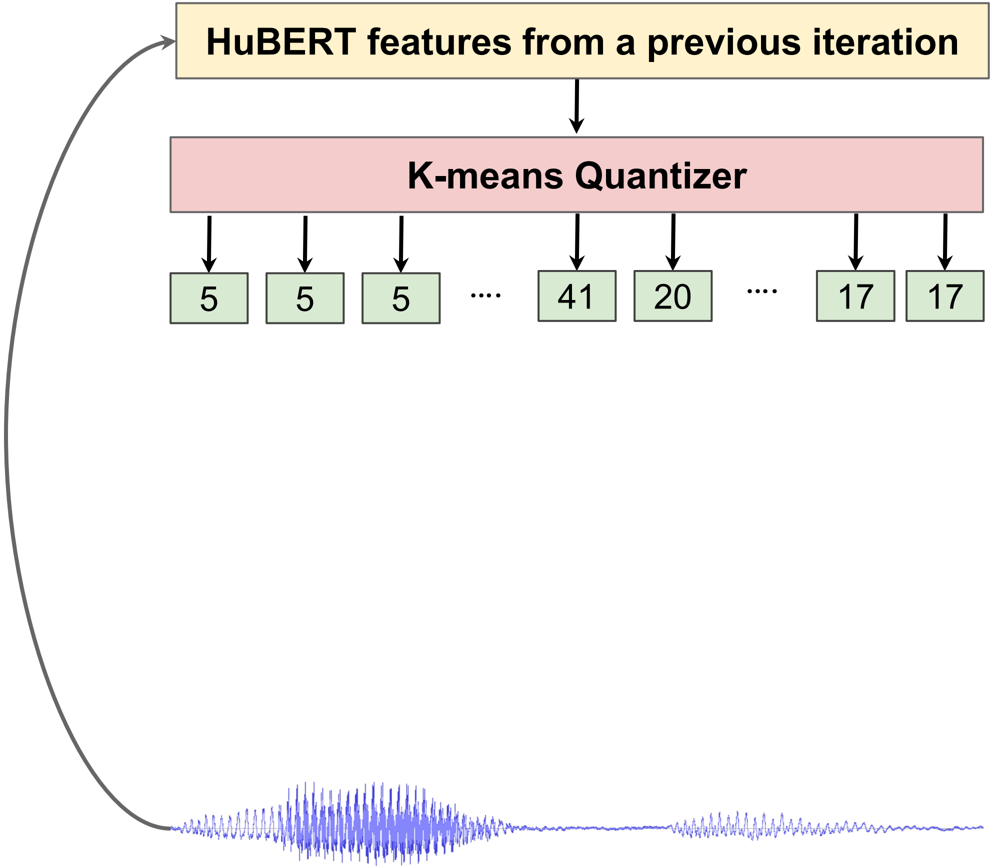
$$\mathcal{L}_m = \sum_{t \in M} -\log p(y_t \mid X)$$



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task

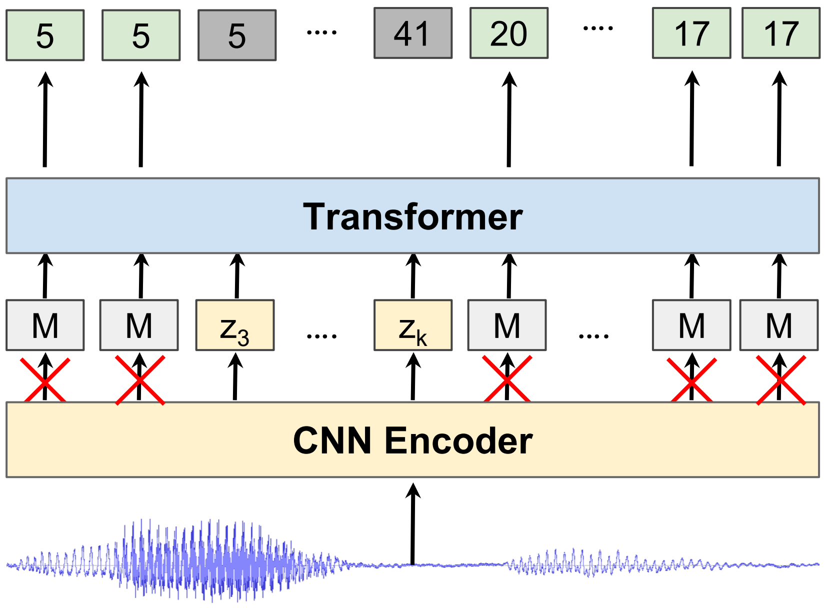
- Then the process can be repeated using learned HuBERT features from a previous iteration.



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: The pretext task

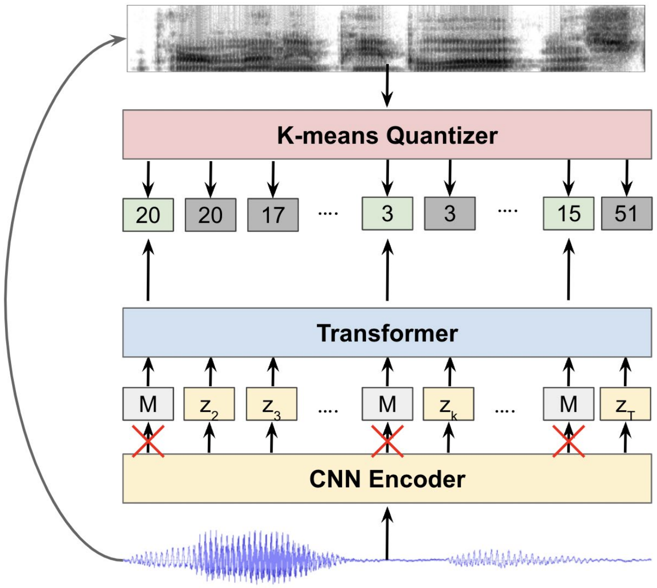
- Then the process can be repeated using learned HuBERT features from a previous iteration.



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: Implementation details

- A small codebook size, e.g., 50, 100, is used for the initial training iteration to focus on phonetic differences rather than speaker and style.
- For the subsequent two iterations, layers 6 and 9 of the base architecture (12 layers) are used for the clustering steps. They found empirically to contain higher quality features over many speech tasks.



Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

HuBERT: Results

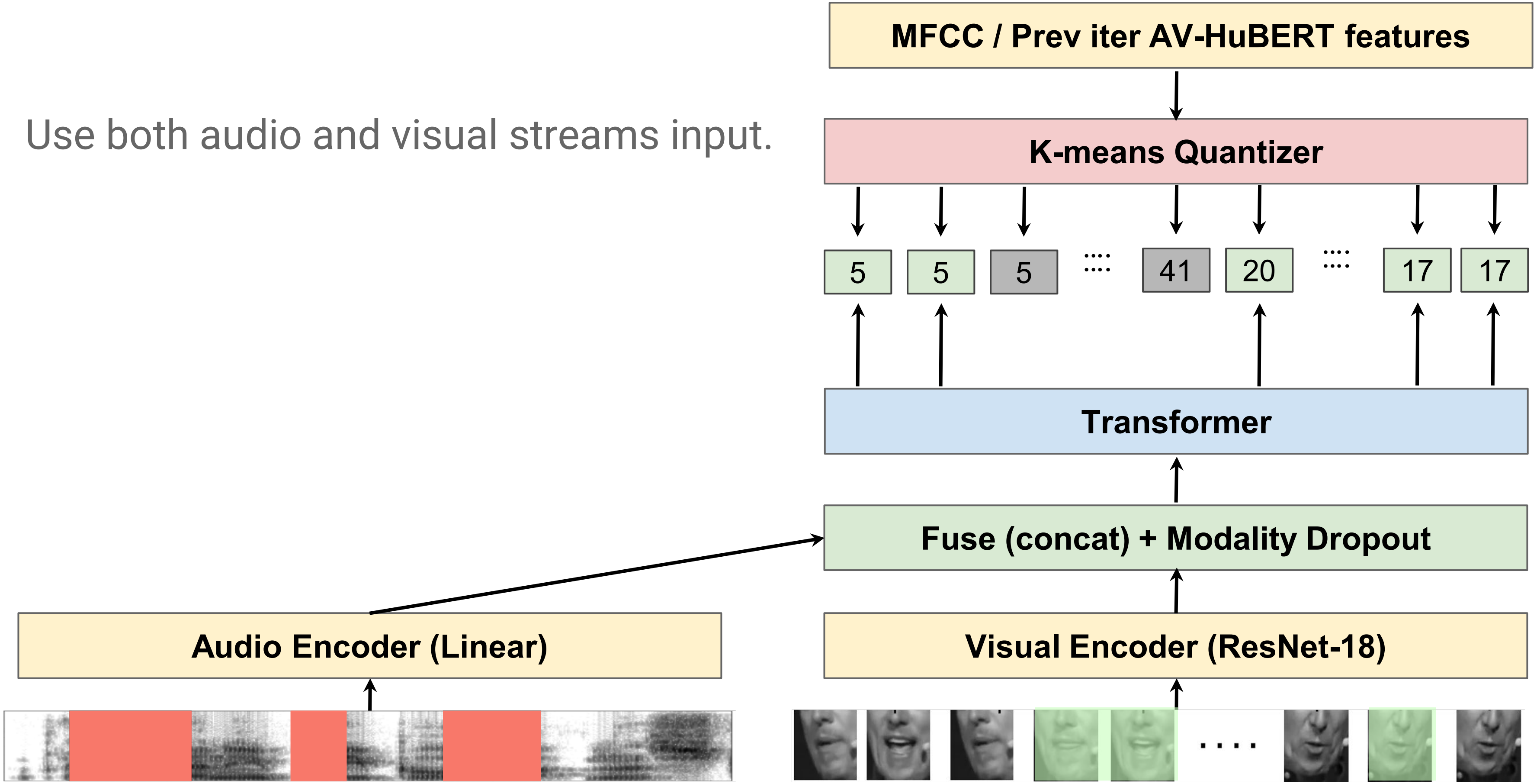
- Matched or beat the SOTA on ASR.
- The best representations for multiple downstream tasks (time of submission).
- The basis for WavLM, the current best system from Microsoft.

	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]	58.95	82.54	29.82	37.99	57.86	24.92	16.61	0.0072	62.14	60.17	11.61	8.68
APC [7]	42.21	91.01	74.69	60.42	59.33	21.61	15.09	0.0310	70.46	50.89	8.56	10.53
VQ-APC [32]	41.49	91.11	74.48	60.15	59.66	21.72	15.37	0.0251	68.53	52.91	8.72	10.45
NPC [33]	43.69	88.96	69.44	55.92	59.08	20.94	14.69	0.0246	72.79	48.44	9.4	9.34
Mockingjay [8]	70.84	83.67	34.33	32.29	50.28	23.72	15.94	6.6E-04	61.59	58.89	11.66	10.54
TERA [9]	49.17	89.48	57.90	57.57	56.27	18.45	12.44	0.0013	67.50	54.17	15.89	9.96
modified CPC [34]	42.54	91.88	64.09	39.63	60.96	20.02	13.57	0.0326	71.19	49.91	12.86	10.38
wav2vec [12]	32.24	95.59	84.92	56.56	59.79	16.40	11.30	0.0485	76.37	43.71	7.99	9.9
vq-wav2vec [13]	34.24	93.38	85.68	38.80	58.24	18.70	12.69	0.0410	77.68	41.54	10.38	9.93
wav2vec 2.0 Base [14]	5.56	96.23	92.35	75.18	63.43	9.57	6.32	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	86.94	27.80	5.65	**5.62**
HuBERT Base [35]	5.05	96.30	98.34	81.42	64.92	6.74	4.93	**0.0736**	88.53	25.20	**5.11**	5.88
HuBERT Large [35]	**3.28**	95.29	**98.76**	**90.33**	**67.62**	**3.67**	**2.91**	0.0353	**89.81**	**21.76**	5.98	5.75

Hsu et al 2021, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”

AV-HuBERT: Audio-Visual ASR

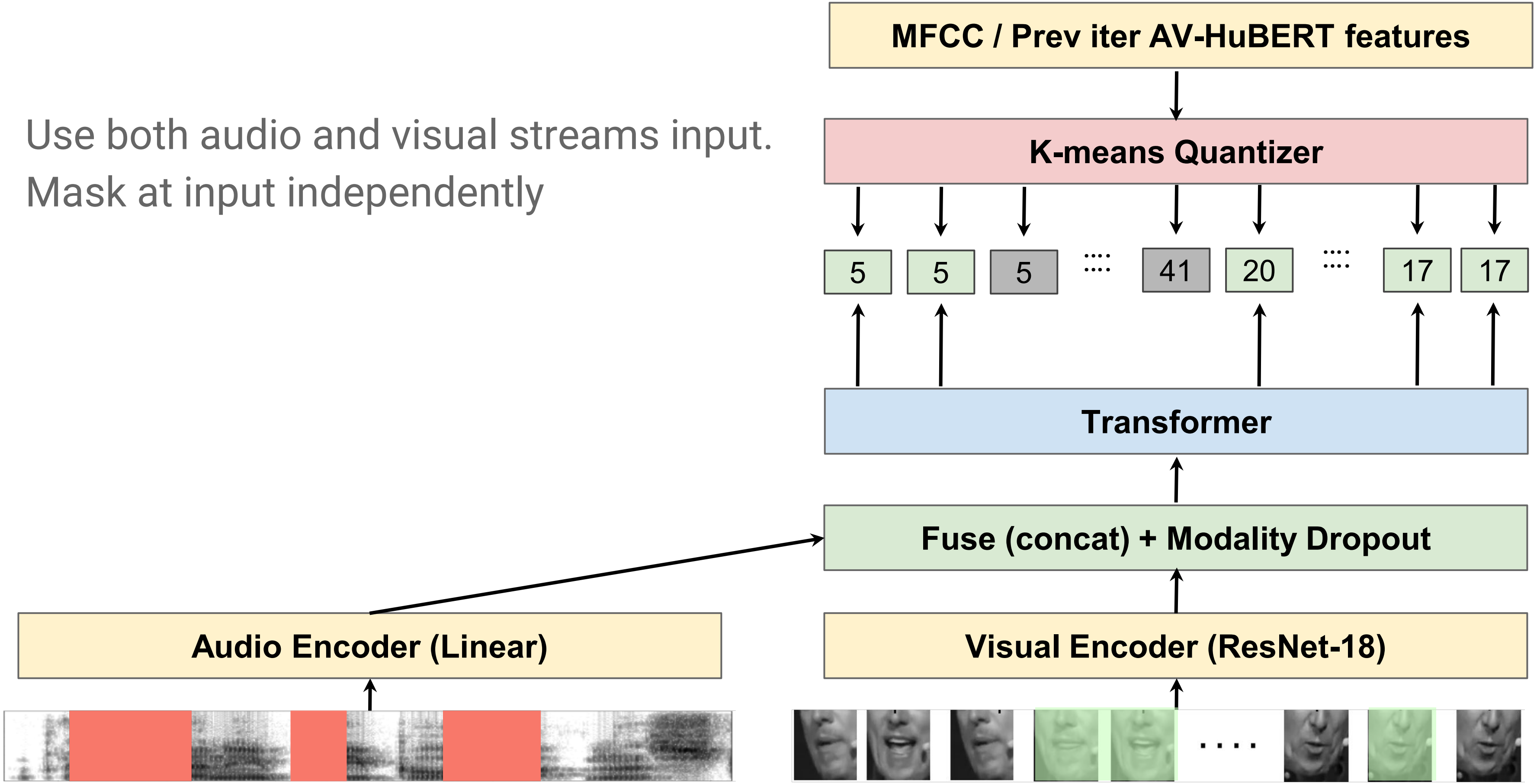
- Use both audio and visual streams input.



Shi et al 2022, "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction"

AV-HuBERT: Audio-Visual ASR

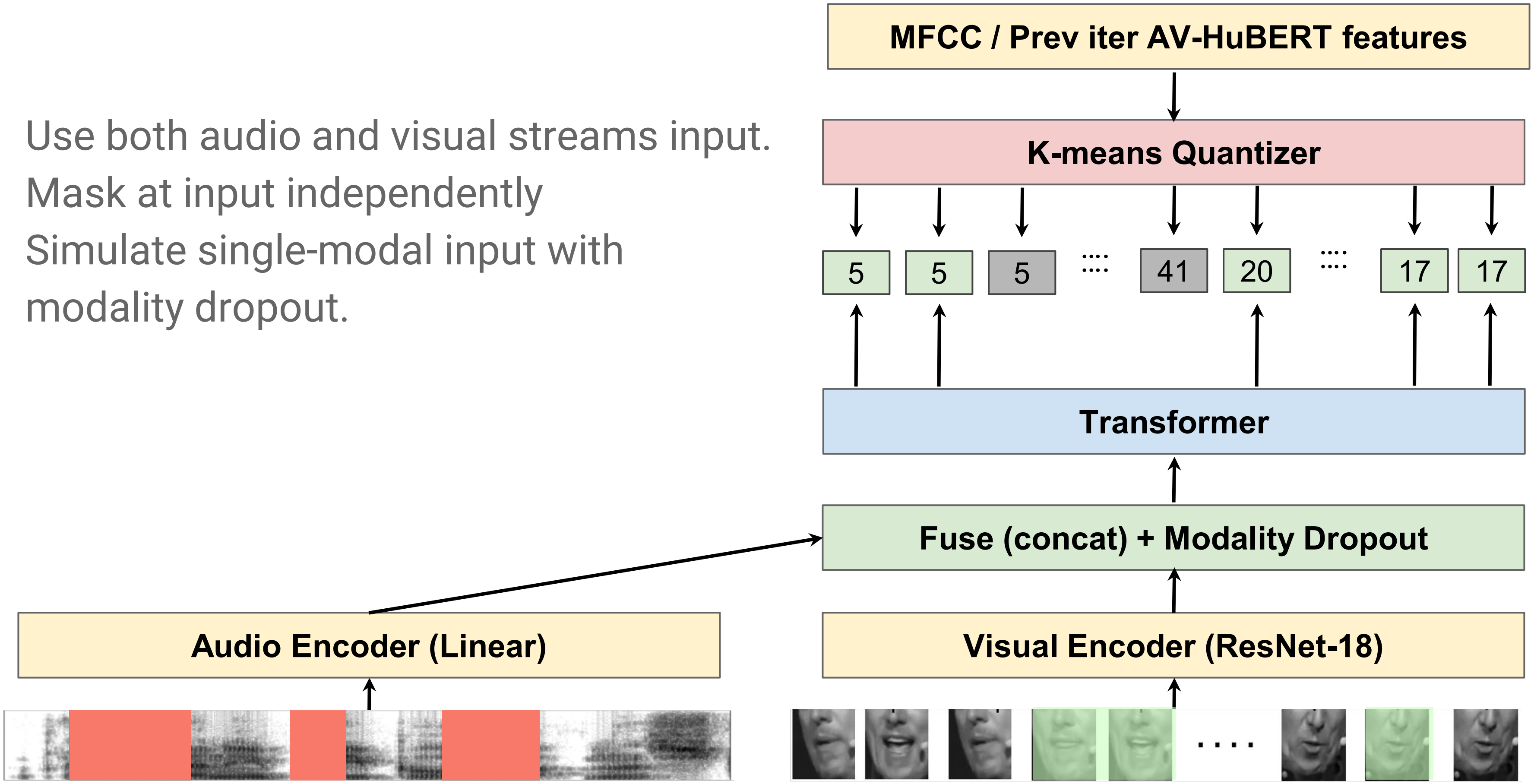
- Use both audio and visual streams input.
- Mask at input independently



Shi et al 2022, "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction"

AV-HuBERT: Audio-Visual ASR

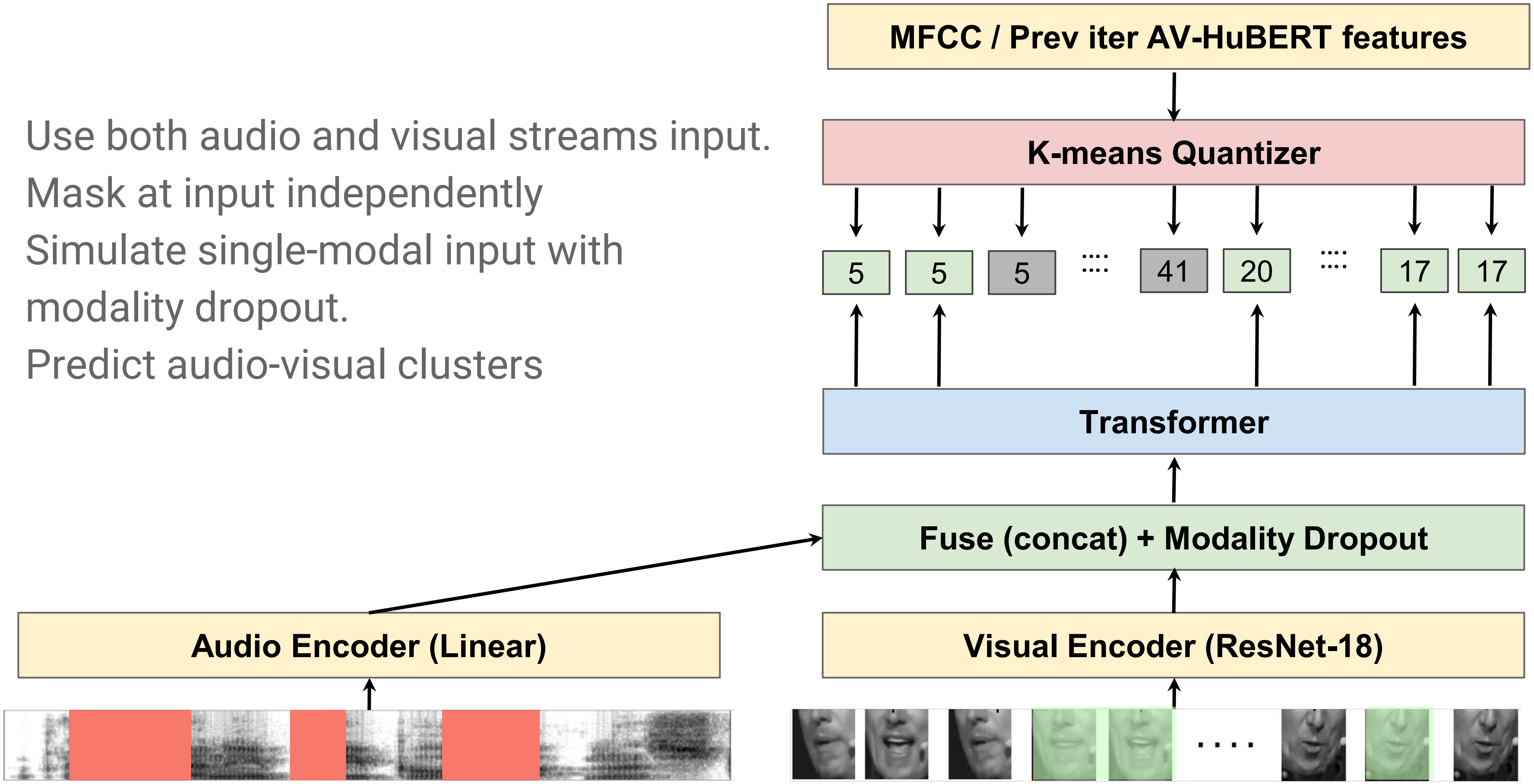
- Use both audio and visual streams input.
- Mask at input independently
- Simulate single-modal input with modality dropout.



Shi et al 2022, “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction”

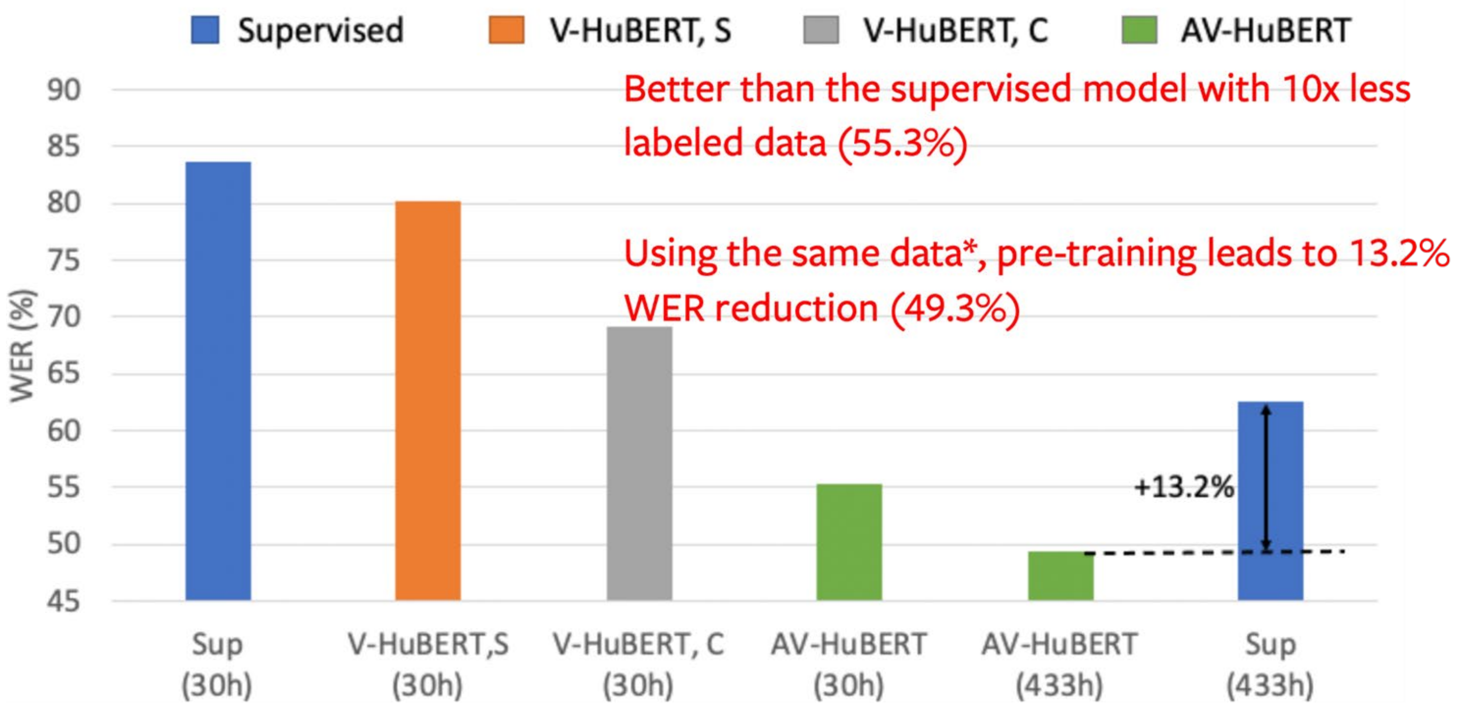
AV-HuBERT: Audio-Visual ASR

- Use both audio and visual streams input.
- Mask at input independently
- Simulate single-modal input with modality dropout.
- Predict audio-visual clusters



Shi et al 2022, “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction”

AV-HuBERT: Results



Shi et al 2022, “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction”

AV-HuBERT: Insights

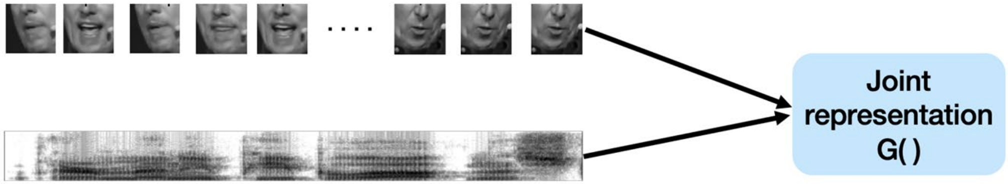
- Going directly to predict text labels from visual input is NOT effective

Shi et al 2022, "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction"

AV-HuBERT: Insights

- Going directly to predict text labels from visual input is NOT effective

- Constraining the network into a joint audio-visual space first leads to much more effective representations.



Shi et al 2022, “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction”

Textless NLP

Getting closer to humans

Supervised ASR

- Paired Text-audio
- Lexicon

Getting closer to humans

Supervised ASR

- Paired Text-audio
- Lexicon

Unsupervised ASR

- Unpaired Text-audio
- Lexicon

Getting closer to humans

Supervised ASR

- Paired Text-audio
- Lexicon

Unsupervised ASR

- Unpaired Text-audio
- Lexicon

Textless NLP

- Just audio!

Textless NLP: Motivations

- Babies learn their first language through spoken interaction (without text).

Textless NLP: Motivations

- Babies learn their first language through spoken interaction (without text).
- Speech processing methods leave out spoken-only dialects and languages, e.g., Swiss Germain, Igbo, and Egyptian Arabic.

Textless NLP: Motivations

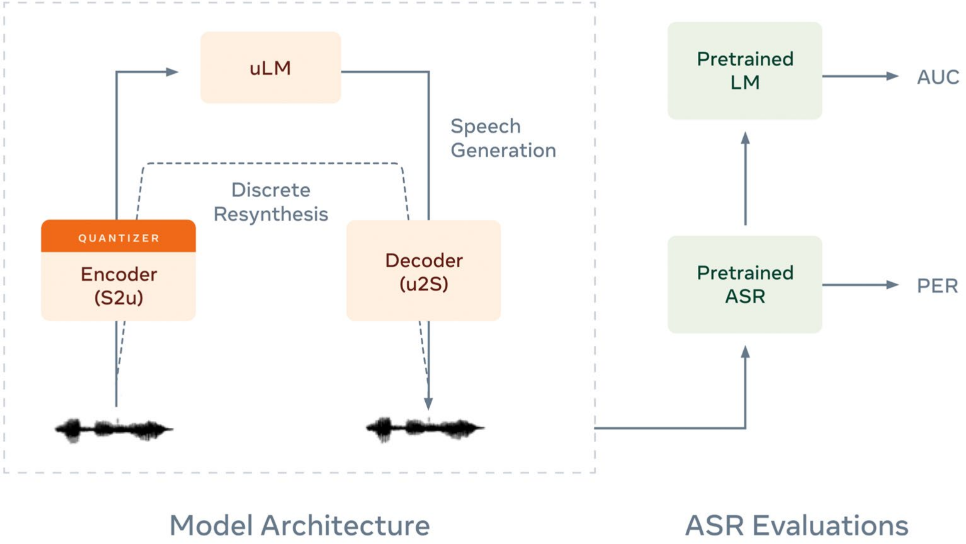
- Babies learn their first language through spoken interaction (without text).
- Speech processing methods leave out spoken-only dialects and languages, e.g., Swiss Germain, Igbo, and Egyptian Arabic.
- Limited work on modeling natural spoken cues while learning representations, e.g. hesitation, laughter, interruptions.

Textless NLP: Applications

- Generative Spoken Language Modeling (GSLM)
- Expressive speech modeling and generation.
- Speech resynthesis, compression.
- Spoken Dialogue Modeling
- Speaker Conversion
- Emotion Conversion
- Speech-to-speech translation
- ....

Textless NLP: GSLM

- GSLM learns jointly the acoustic and linguistic characteristics of a language from raw audio only.



Lakhotia et al 2021, "Generative Spoken Language Modeling from Raw Audio"

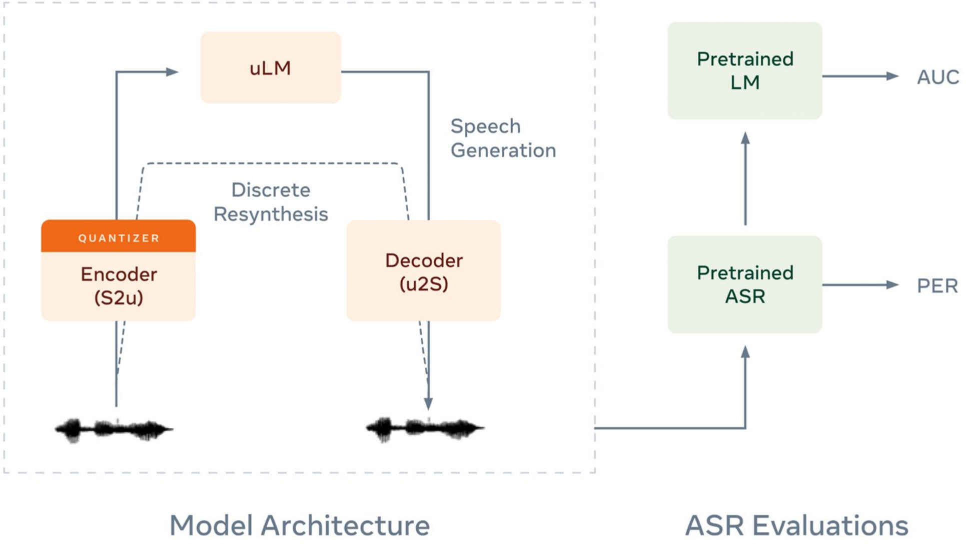
Textless NLP: GSLM

- GSLM learns jointly the acoustic and linguistic characteristics of a language from raw audio only.

- GSLM evaluation metrics should be:

    1. Independent of the learned discrete unit.

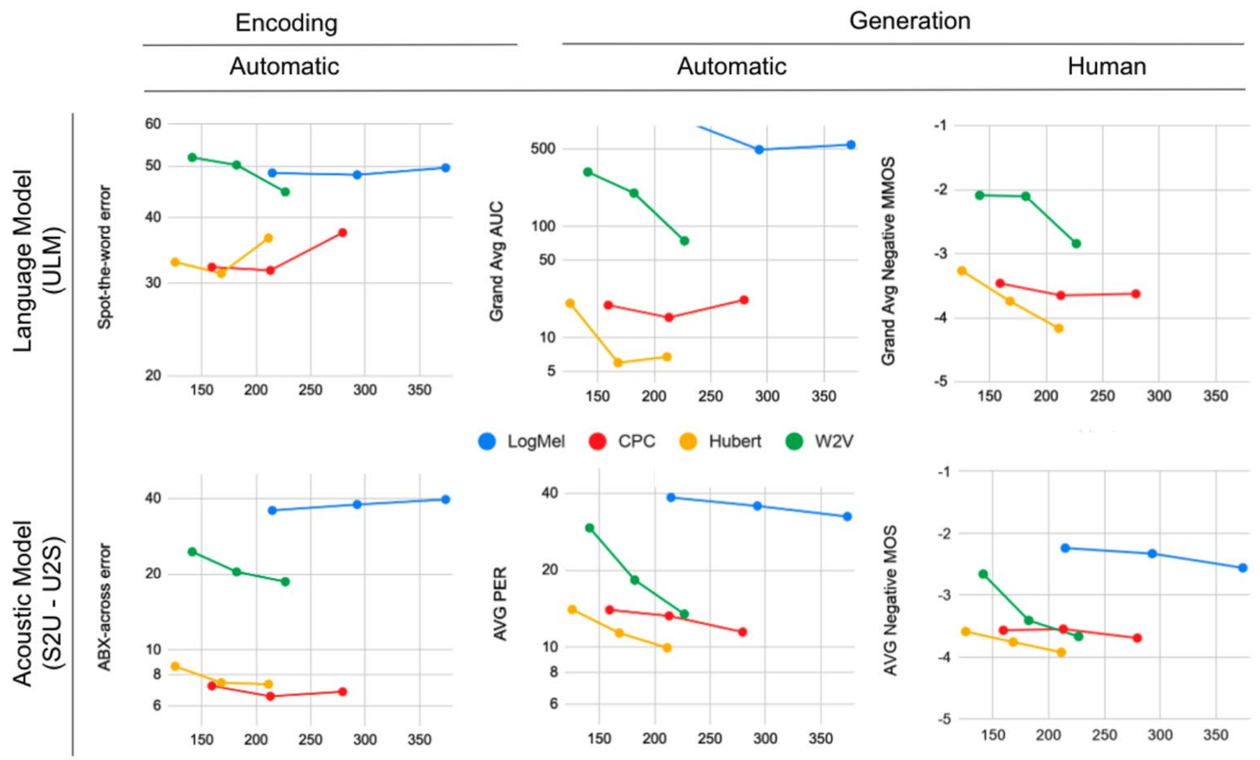
    2. Evaluate the intelligibility, diversity, and meaningfulness of the generated content.



Lakhotia et al 2021, “Generative Spoken Language Modeling from Raw Audio”

Textless NLP: GSLM Results

- Generated content is as good as character-based LM+TTS



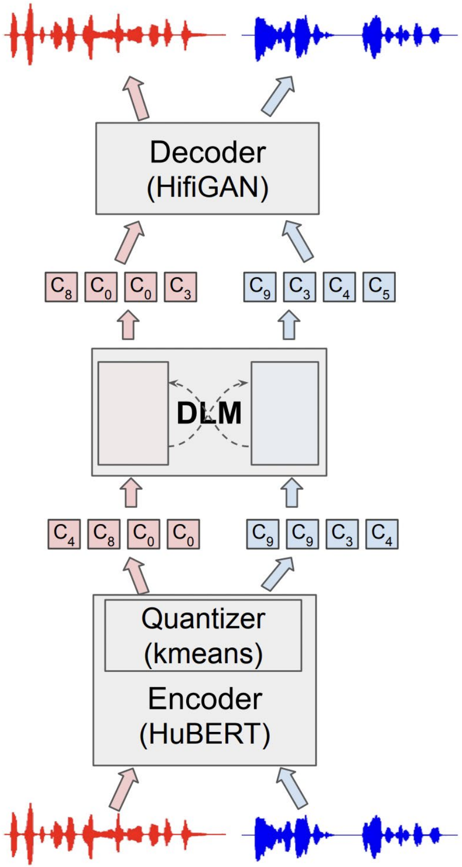
Lakhotia et al 2021, “Generative Spoken Language Modeling from Raw Audio”

Textless NLP: Dialogue Generation

Nguyen et al 2022 "Generative Spoken Dialogue Language Modeling"

Textless NLP: Dialogue Generation

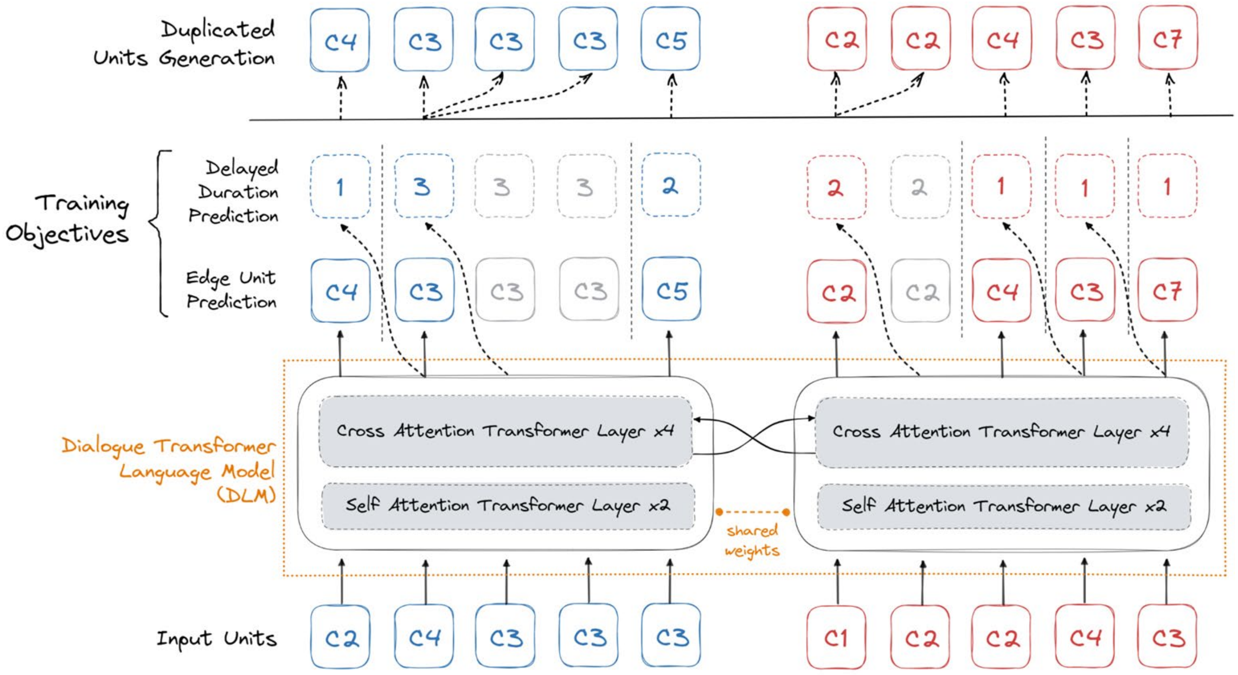
- GSLM was also extended to model and generate multi turn dialogues of Fisher data.

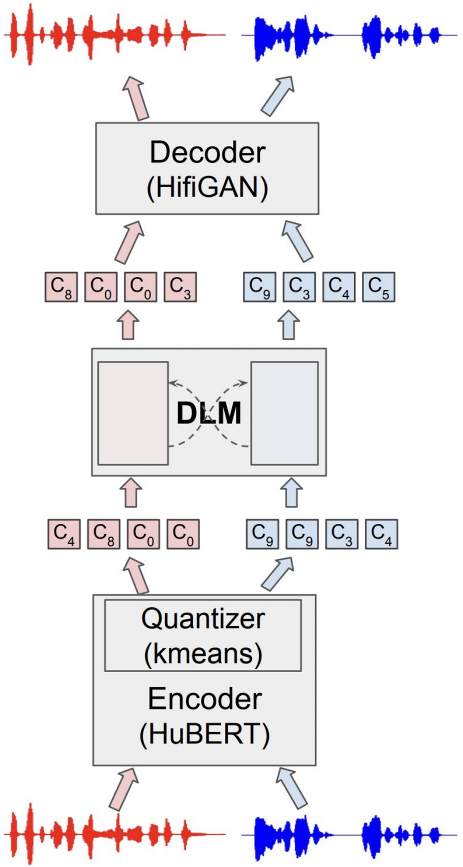


Nguyen et al 2022 "Generative Spoken Dialogue Language Modeling"

Textless NLP: Dialogue Generation

- GSLM was also extended to model and generate multi turn dialogues of Fisher data.

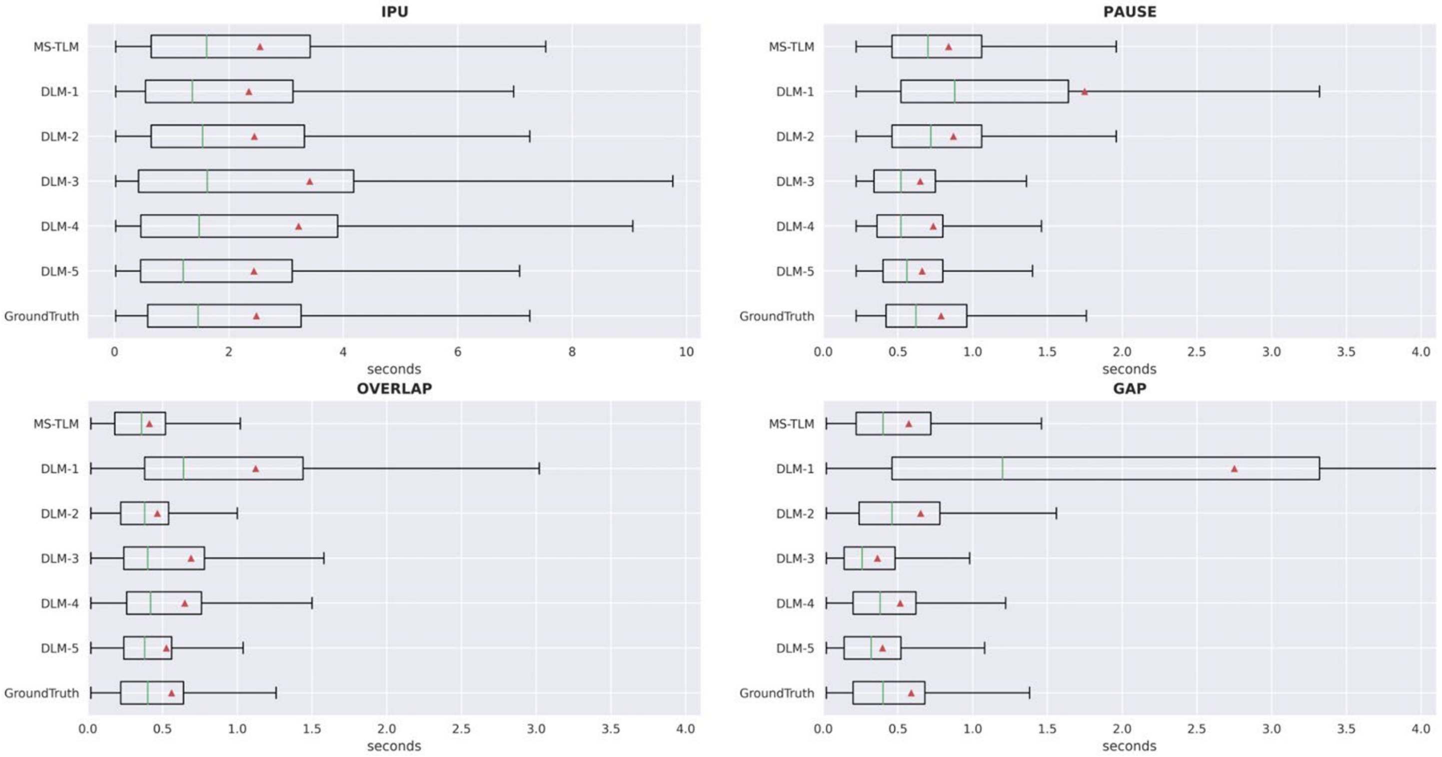




Nguyen et al 2022 "Generative Spoken Dialogue Language Modeling"

Textless NLP: Dialogue Generation Results

- The model learns to mimic the stats of human-human communication



Nguyen et al 2022 "Generative Spoken Dialogue Language Modeling"

Privacy-preserving Speech Representation Learning

Privacy-preservation + lifelong learning

- Can models keep training on device without communication at all with servers?
- Can we build representation models that improves for certain household without degrading for visitors?

LibriContinual for lifelong representation learning

- **LibriContinual** is a new open-source benchmark to test our technology abilities.
- It contains 118 speakers from Librivox.

Subset	#hrs/spkr	#utts/spkr
train-10min	0.17 ± 0.001	114 ± 28
train-30min	0.50 ± 0.001	337 ± 81
train-1hr	1.00 ± 0.001	677 ± 163
train-2hr	2.00 ± 0.001	1356 ± 322
train-5hr	5.00 ± 0.003	3387 ± 806
train-10hr	10.00 ± 0.005	6772 ± 1608
valid	3.13 ± 1.86	2125 ± 1406
test	2.66 ± 1.15	1880 ± 1101

Diwan et al 2022 “Continual Learning For On-Device Speech Recognition Using Disentangled Conformers”

Binary HuBERT for optimized processing

- First step for more optimized training on device.
- Tested two different methods for binarizing HuBERT models

Base Model	Quant	Precision	SUPERB Tasks									Storage (MBs)↓	FLOPs (Gs)↓	QuantOPs (GBits)↓	Runtime (Est. x)↓
			ASR↓	KS↑	SF↑	PR↓	QbE↑	IC↑	ASV↓	SD↓	ER↑				
HuBERT (Base)[3]	–	fp16	6.42	96.59	0.88	5.41	7.36	97.15	5.11	6.20	64.92	189.14	153.14	0.00	1.38
HuBERT (+FastConv[17])	–	fp16	**7.06**	96.62	0.89	6.05	6.91	97.28	5.30	6.32	65.00	**184.42**	110.79	0.00	**1.00**
	SqWQ[2]	w8	9.69	96.88	0.88	7.30	6.19	96.65	5.88	6.52	62.83	99.65	82.24	1898.44	**1.00**
		w4	9.98	96.59	0.88	8.03	5.86	96.26	6.06	6.73	62.79	57.19	82.24	1054.69	0.89
		w2	12.56	94.22	0.86	11.79	5.27	94.02	6.31	7.12	62.38	35.95	82.24	632.81	0.83
		w1	25.37	85.07	0.73	41.77	4.74	64.88	18.23	11.26	54.40	25.34	82.24	421.88	0.80
	BiT-L[1] (Linear Only)	w8a8	7.03	96.85	0.88	6.22	6.36	98.23	5.54	6.36	65.94	99.49	82.29	1898.44	1.00
		w4a4	8.58	96.56	0.88	7.15	6.40	96.10	5.55	6.26	64.12	57.02	82.29	527.34	0.81
		w2a2	10.80	95.88	0.86	8.79	5.62	97.47	5.68	6.55	63.49	35.79	82.29	158.20	0.76
		w1a1	12.23	94.94	0.86	10.49	5.99	96.49	6.55	6.87	63.06	25.17	82.29	52.73	0.75
	BiT-LA[1] (Linear +Attention)	w8a8	7.07	97.21	0.89	6.30	6.40	98.10	5.56	6.24	65.77	99.54	11.82	3868.56	0.63
		w4a4	9.35	96.62	0.88	7.76	6.37	96.92	5.75	6.09	66.58	57.08	11.82	1074.60	0.25
		w2a2	12.68	95.07	0.85	12.56	5.23	95.02	7.40	6.94	63.00	35.84	11.82	322.38	0.15
		w1a1	**15.96**	93.83	0.78	22.96	5.63	93.01	6.83	7.62	61.68	**25.23**	11.82	107.46	**0.12**
DistillHuBERT[5]	–	fp16	13.37	95.98	0.83	16.27	5.11	94.99	8.55	6.19	63.02	46.98	80.34	0.00	0.73

Yeh et al 2022, "Efficient Speech Representation Learning With Low-Bit Quantization"

Questions?