Recent Advances in Speech Representation Learning

Abdelrahman Mohamed

facebook
Artificial Intelligence Research

North Star for Speech Representation Learning (SRL)

- **Pre-training**: Reduce labeling cost for a wide range of scenarios.
- **Inclusive:** Serves the needs of everyone with written and spoken-only languages and dialects (with lexical differences).
- **Capture natural interactions / conversations**: Content, style, emotion, hesitation.
- **Learning like a baby**: listening, talking, and interacting.

Recent related work:

- **Benchmarks:**

“Libri-Light: A Benchmark for ASR with Limited or No Supervision”

“SUPERB: Speech processing Universal PERformance Benchmark”

- **Weakly- and Semi-supervised Learning for speech recognition:**

“Training ASR models by Generation of Contextual Information”

“Large scale weakly and semi-supervised learning for low-resource video ASR”

“Contrastive Semi-supervised Learning for ASR”

- **Self-supervised speech representation learning:**

“Effectiveness of self-supervised pre-training for speech recognition”

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

“Unsupervised Cross-lingual Representation Learning for Speech Recognition”

“HuBERT: how much can a bad teacher benefit ASR pre-training?”

- **Language generation:**

“Generative Spoken Language Modeling from Raw Audio”

“Speech Resynthesis from Discrete Disentangled Self-Supervised Representations”

Positive outcomes:

- **High and low-resource written languages**: Impressive results across a wide range of languages and scenarios.
- **Near-zero supervision for ASR:**
  1. Below 10% WER using 10min of labeled data on public benchmarks
  2. Production-quality WER under challenging conditions using 100h of

labels – even with 10h and 1h of labeled data in many cases.

- **Generative Spoken Language Modeling:**
  1. Audio-only language modeling and speech generation
  2. Competitive subjective scores to character-based system
  3. Ultra-lightweight speech codec

Challenges:

- **Requires large volumes of audio-only resources**: Way more than what a human encounters before conversational understanding.
- **Large computational resources**: High bar of entry to this research area.
- **Huge dependence on textual resources**:

    1. Hard expansion to spoken-only dialects and languages.
    2. It limits modeling non-lexical signals in conversations, e.g., hesitation, laughter, interruptions.

Recent related work:

- **Benchmarks:**

"Libri-Light: A Benchmark for ASR with Limited or No Supervision"

"SUPERB: Speech processing Universal PERformance Benchmark"

- **Weakly- and Semi-supervised Learning for speech recognition:**

"Training ASR models by Generation of Contextual Information"

"Large scale weakly and semi-supervised learning for low-resource video ASR"

"Contrastive Semi-supervised Learning for ASR"

- **Self-supervised speech representation learning:**

"Effectiveness of self-supervised pre-training for speech recognition"

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

"Unsupervised Cross-lingual Representation Learning for Speech Recognition"

"HuBERT: how much can a bad teacher benefit ASR pre-training?"

- **Language generation:**

"Generative Spoken Language Modeling from Raw Audio"

"Speech Resynthesis from Discrete Disentangled Self-Supervised Representations"

Recent related work:

- **Benchmarks:**

"Libri-Light: A Benchmark for ASR with Limited or No Supervision"

"SUPERB: Speech processing Universal PERformance Benchmark"

- **Weakly- and Semi-supervised Learning for speech recognition:**

"Training ASR models by Generation of Contextual Information"

"Large scale weakly and semi-supervised learning for low-resource video ASR"

"Contrastive Semi-supervised Learning for ASR"

- **Self-supervised speech representation learning:**

"Effectiveness of self-supervised pre-training for speech recognition"

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

"Unsupervised Cross-lingual Representation Learning for Speech Recognition"

"HuBERT: how much can a bad teacher benefit ASR pre-training?"

- **Language generation:**

"Generative Spoken Language Modeling from Raw Audio"

"Speech Resynthesis from Discrete Disentangled Self-Supervised Representations"

Outline

- Contrastive Semi-supervised Learning (CSL) for ASR
- HuBERT: how much can a bad teacher benefit ASR pre-training?
- Generative Spoken Language Modeling (GSLM)

“Contrastive Semi-supervised Learning (CSL) for ASR”

Alex Xiao, Christian Fuegen, Abdelrahman Mohamed

CSL – Motivations:

- Pseudo-labeling (PL) is the most adopted pre-training method for ASR.
- PL performance greatly suffers from degrading teacher quality in low-resource setups and under domain transfer.
- Contrastive self-supervised pre-training approaches are gaining momentum; however, positive and negative data sampling is tricky for speech.

CSL – Approach:

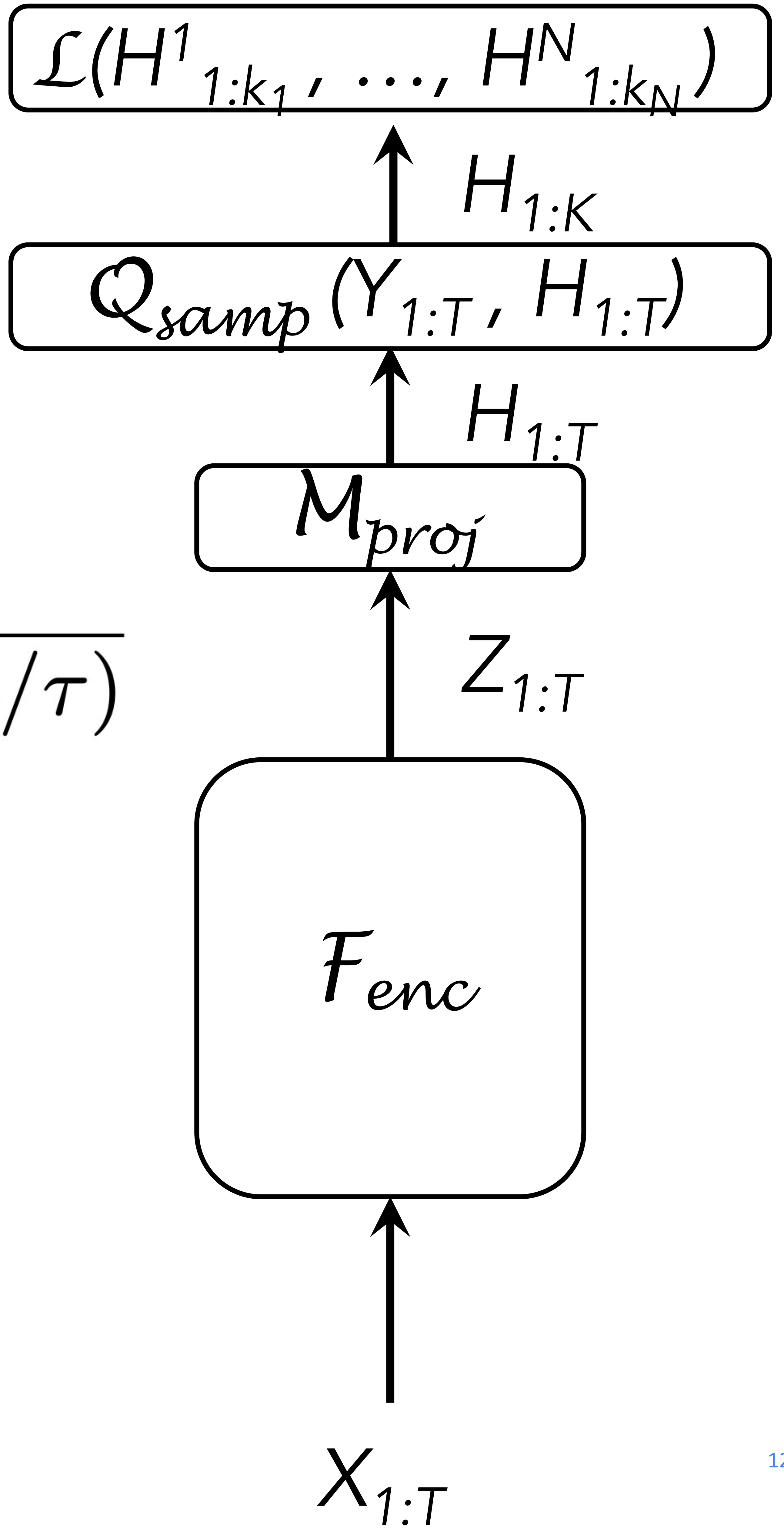
- CSL uses a contrastive loss in the semi-supervised learning setup.
- By utilizing a supervised teacher, CSL bypasses the challenge of positive and negative sample selection of self-supervised methods.
- CSL is resilient to errors in teacher-generated targets since it relies on the relative distance between labels.

CSL – Details:

- Positive and negative samples are audio segments.
- Positive samples are segments with the same label.

$$\mathcal{L} = \frac{1}{S}\sum_{i=1}^{S}\frac{1}{|P(i)|}\sum_{h_p \in P(i)} -\log \frac{\exp(h_i \cdot h_p/\tau)}{\sum_{h \in N(i)\cup\{h_p\}} \exp(h_i \cdot h/\tau)}$$

- CSL requires at least two positive instances of each label in the mini-batch.
- Label-Aware Batching (LAB) is used to boost the sampling of rare sounds.



CSL – Experimental setup:

- We use de-identified public FB videos in British English and Italian.
- 10hr and 1hr of labeled data are used for teacher training and final fine-tuning.
- 75,000hr of unlabeled audio is used for pre-training in each language.
- For reference, we report the performance of a fully supervised system for British English (650hr) and Italian (3,700hr).

CSL – Results:

- Word Error Rates (WER) are 8% lower for CSL relative to PL for the 10hr case.
- WERR is up to 17% under domain transfer and 19% for the ultra low-resource 1hr case.
- Both CSL and PL benefit from iterative labeling, while CSL is still more than 6.5% better than PL after three generations.

		British English	**Italian**
	Supervised Baseline		
A1	Full supervised data	23.1 (650hr)	11.9 (3,700hr)
A2	10hr of labels	50.7	31.8
A3	1hr of labels	80.5	-
	Pre-training using 10hr of teacher supervision		
B1	Pseudo-Labeling (PL)	32.0	17.2
B2	CSL	**29.4**	**16.0**
	Pre-training using 1hr of teacher supervision		
C1	Pseudo-Labeling (PL)	53.1	-
C2	CSL	**42.8**	-
C3	CSL (Gen2)	**32.3**	-

	PL	**CSL**	**WERR**
British English Videos	32.0	29.4	8.1
General English Videos	37.2	32.8	11.8
Message Dictation	21.6	17.8	17.3
Long-form Conversation	26.0	22.0	15.4

“HuBERT: how much can a bad teacher benefit ASR pre-training?*”

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, Abdelrahman Mohamed

*A longer journal/arxiv version will be out soon

HuBERT – Motivations:

- Big success of self-supervised learning for CV, NLP, and Speech.
- Positive and negative data sampling is tricky in contrastive self-supervised methods for speech.
- Speech has some unique challenges:

1. The instance classification does not hold with multiple sounds in each input.
2. During pre-training, there is no prior lexicon of discrete sound units.
3. Sound units are of variable length, and their boundaries are not known.

HuBERT – Approach:

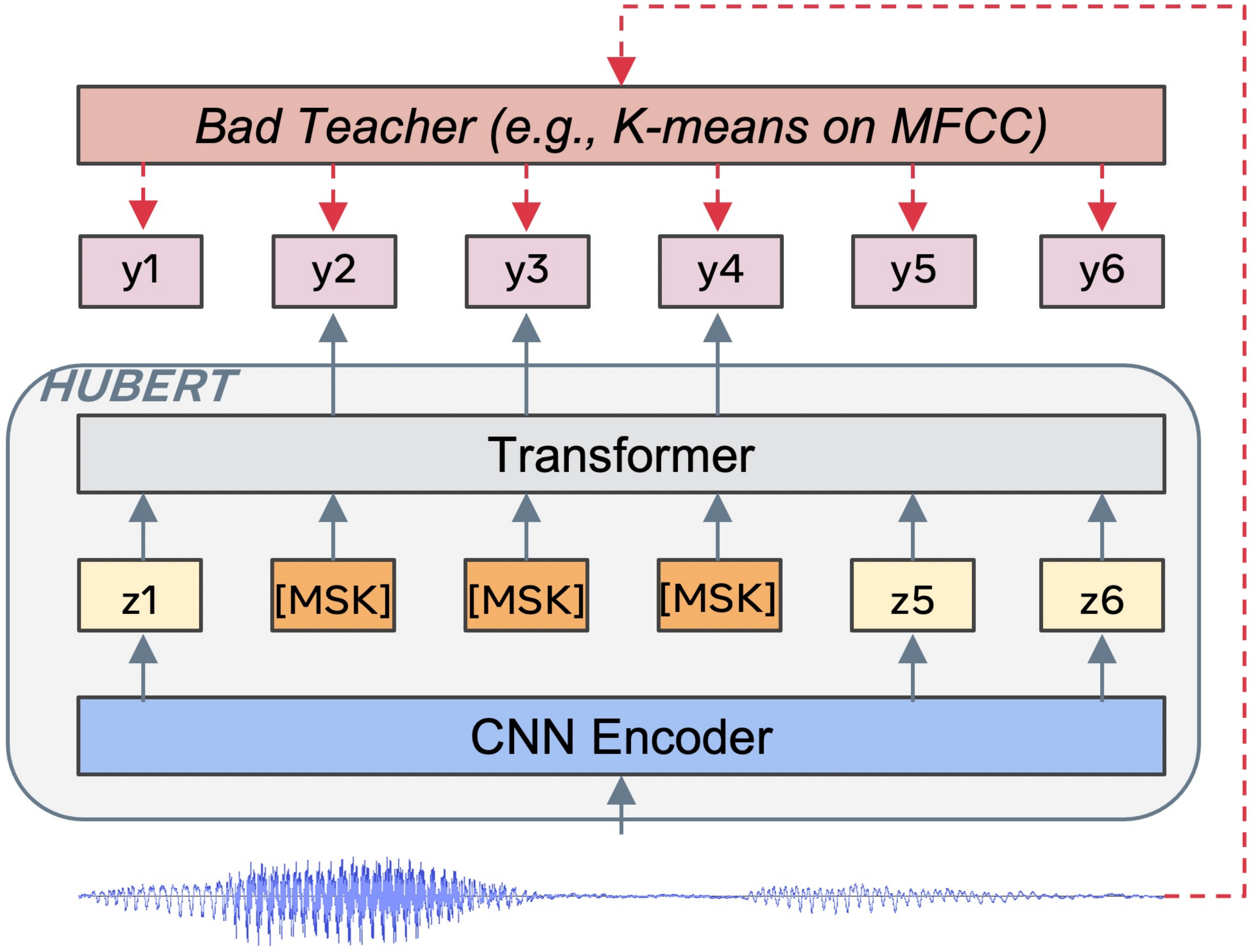
- **HuBERT** = **H**idden **U**nit **BERT**
- We apply the masked prediction loss given continuous inputs with targets generated from an offline k-means clustering.
- Even if the k-means model represents a lousy teacher, its consistency is more important than its quality.
- Intuitively, the HuBERT model learns both acoustic and language models from continuous inputs to minimize the masked prediction loss.

HuBERT – Details:

- Small codebook sizes, e.g. 100, 500.
- The loss is only applied over masked regions.

$$L_m(\theta; X, M, Y) = \sum_{t \in M} \log p(y_t \mid \tilde{X}, t)$$

- The learned latent features can be quantized for another learning iteration.
- GMMs or HMMs may replace k-means for better initial labels.

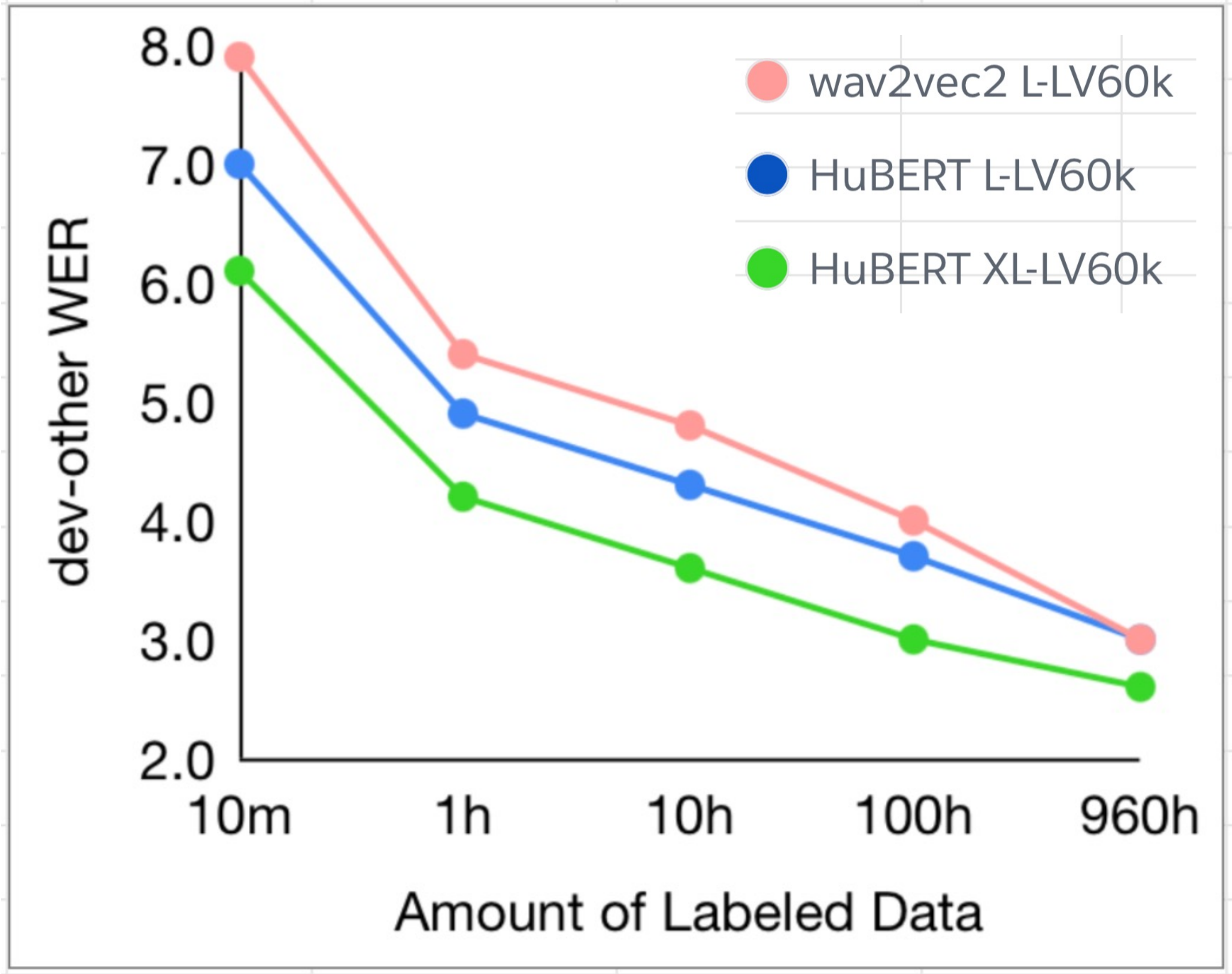


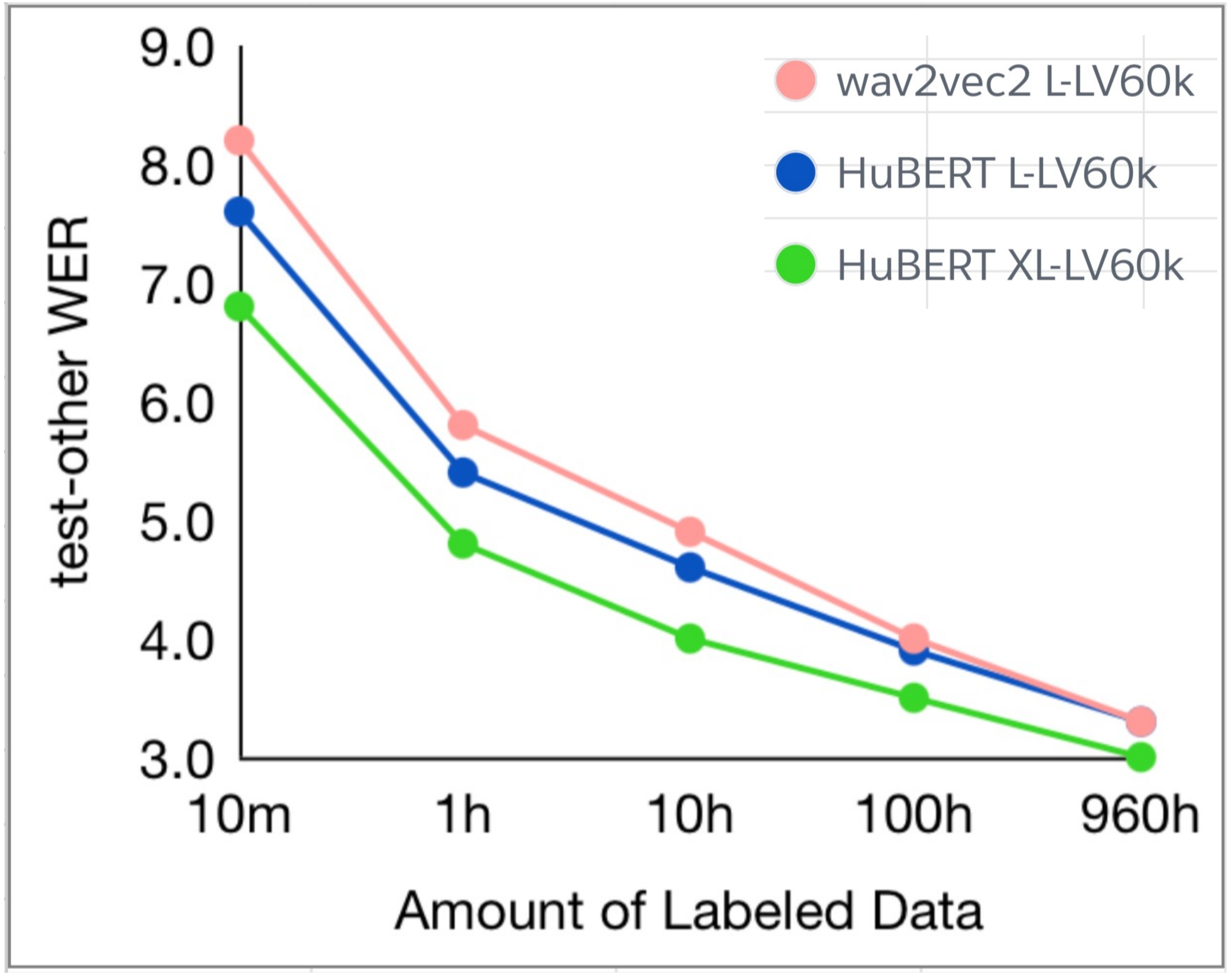
HuBERT – Experimental setup:

- We use the Librispeech 960hr and Libri-light 60,000hr data for pre-training.
- Fine-tuning is done on 10min, 1hr, 10hr, 100hr, or 960hr of labels.
- The official Librispeech language modeling text data is used during decoding.
- A Transformer LM is used for decoding with a sweep over the validation set for best decoder hyper-parameters.

HuBERT – Results:

- Using at most three clustering steps, HuBERT is as effective or better than Wav2Vec 2.0
- Using a 1B model improves the performance across all sizes of labeled data for the challenging dev/test_other condition (up to 19% and13%).
- Starting from a GMM provides some gains over k-means as well as using multiple teacher labels during pre-training. Both gains are much smaller than an extra clustering iteration.





HuBERT – “SUPERB” Results (soon to be public):

	PR	KS	IC	SID	ER	ASR (WER)		QbE	SF		SV	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]	58.88	82.37	30.29	35.84	57.64	24.92	16.61	7.0E-4	60.41	62.77	10.91	8.52
APC [7]	41.85	91.04	74.64	59.79	58.84	21.61	15.09	0.0268	71.26	50.76	8.81	10.72
VQ-APC [32]	42.86	90.52	70.52	49.57	58.31	21.72	15.37	0.0205	69.62	52.21	9.29	10.49
NPC [33]	52.67	88.54	64.04	50.77	59.55	20.94	14.69	0.0220	67.43	54.63	10.28	9.59
Mockingjay [8]	80.01	82.67	28.87	34.50	45.72	23.72	15.94	3.1E-10	60.83	61.15	23.22	11.24
TERA [9]	47.53	88.09	48.8	58.67	54.76	18.45	12.44	8.7E-5	63.28	57.91	16.49	9.54
modified CPC [34]	41.66	92.02	65.01	42.29	59.28	20.02	13.57	0.0061	74.18	46.66	9.67	11.00
wav2vec [12]	32.39	94.09	78.91	44.88	58.17	16.40	11.30	0.0307	77.52	41.75	9.83	10.79
vq-wav2vec [13]	53.49	92.28	59.4	39.04	55.89	18.70	12.69	0.0302	70.57	50.16	9.50	9.93
wav2vec 2.0 Base [14]	28.37	92.31	58.34	45.62	56.93	9.57	6.32	8.8E-4	79.94	37.81	9.69	7.48
HuBERT Base [35]	6.85	**95.98**	95.94	64.84	62.94	6.74	4.93	**0.0759**	86.24	28.52	**7.22**	6.76
HuBERT Large [35]	**3.72**	93.15	**98.37**	**66.40**	**64.93**	**3.67**	**2.91**	0.0360	**88.68**	**23.05**	7.70	**6.23**

“Generative Spoken Language Modeling (GSLM) from Raw Audio”

Kushal Lakhotia, Evgeny Kharitonov*, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, Emmanuel Dupoux*

GSLM – Motivations:

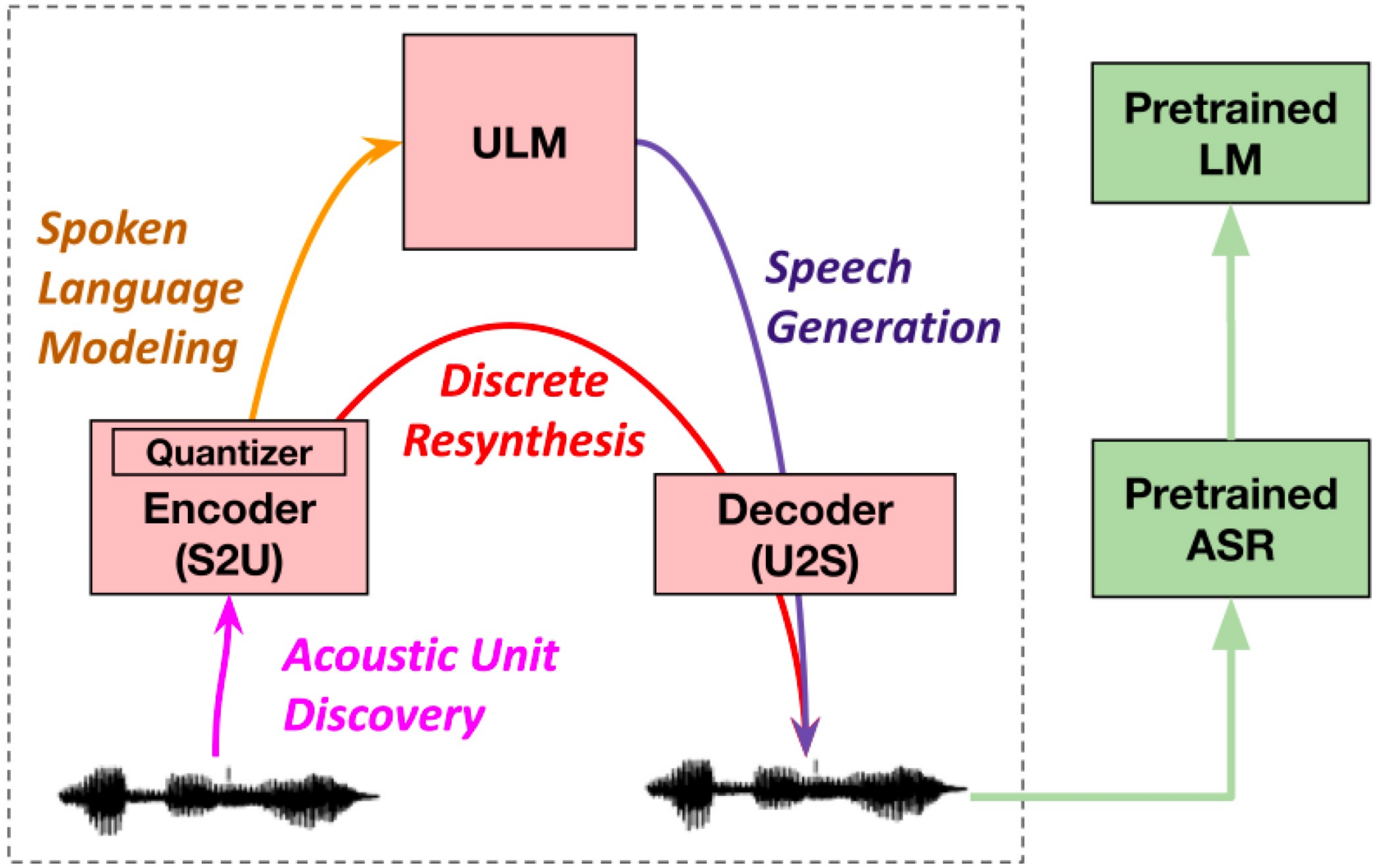
- Babies learn their first language through spoken interaction (without text).
- The big success of self-supervised representation learning for few- and zero-shot downstream scenarios.
- Speech processing methods are heavily dependent on textual resources – leaving out spoken-only dialects and languages, e.g., Swiss Germain, Igbo, and many dialects of Arabic.
- Limited work on modeling natural spoken cues while learning representations, e.g. hesitation, laughter, interruptions.

GSLM – Approach:

- GSLM learns jointly the acoustic and linguistic characteristics of a language from raw audio only (without text or labels).

- GSLM =

unsupervised speech Encoder (S2U),
+ k-means to get pseudo-text
+ Unit Language Model (ULM)
+ speech synthesizer trained on latent units (U2S).

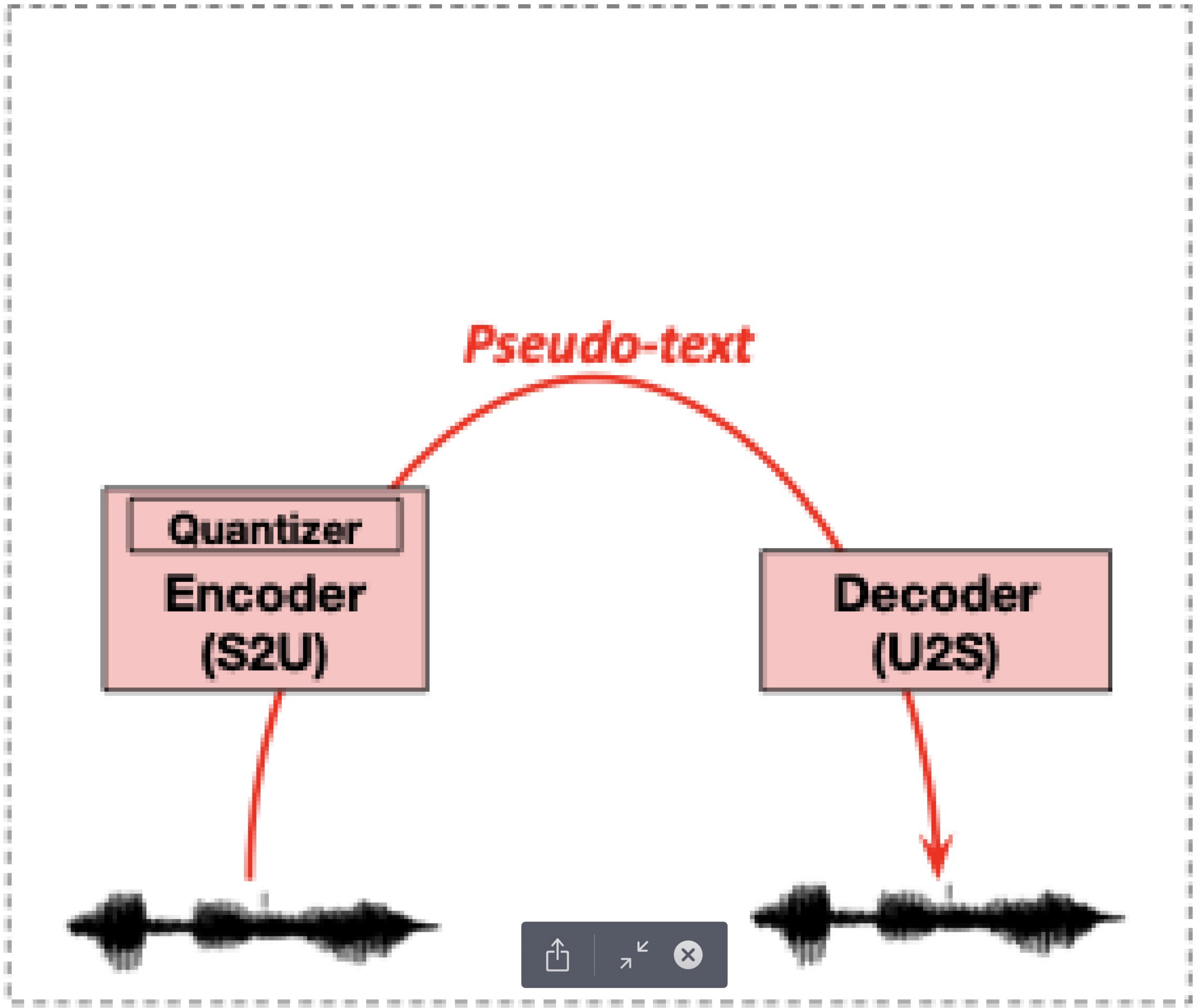


Model architecture and tasks

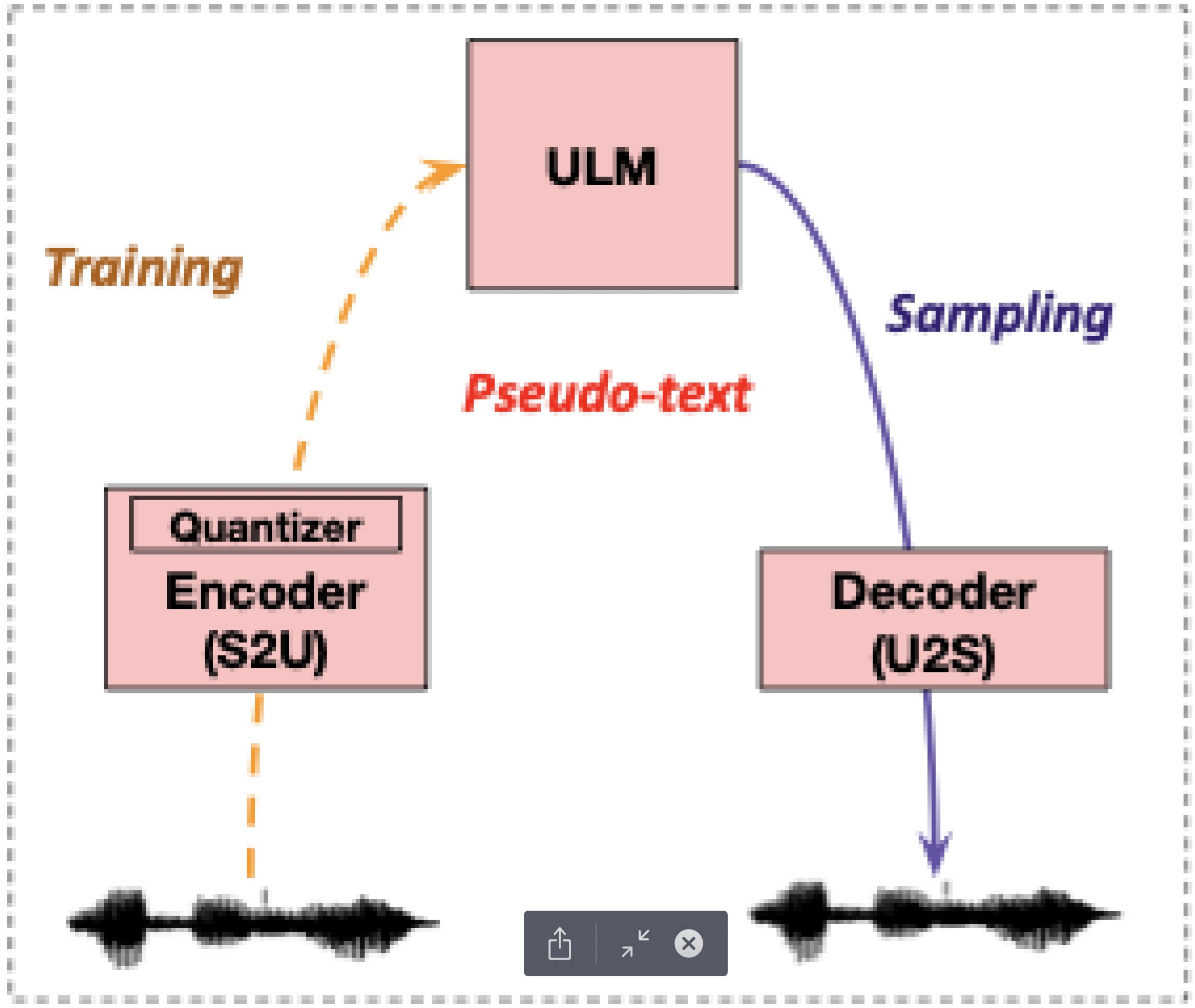
ASR evaluation

GSLM – Three tasks:

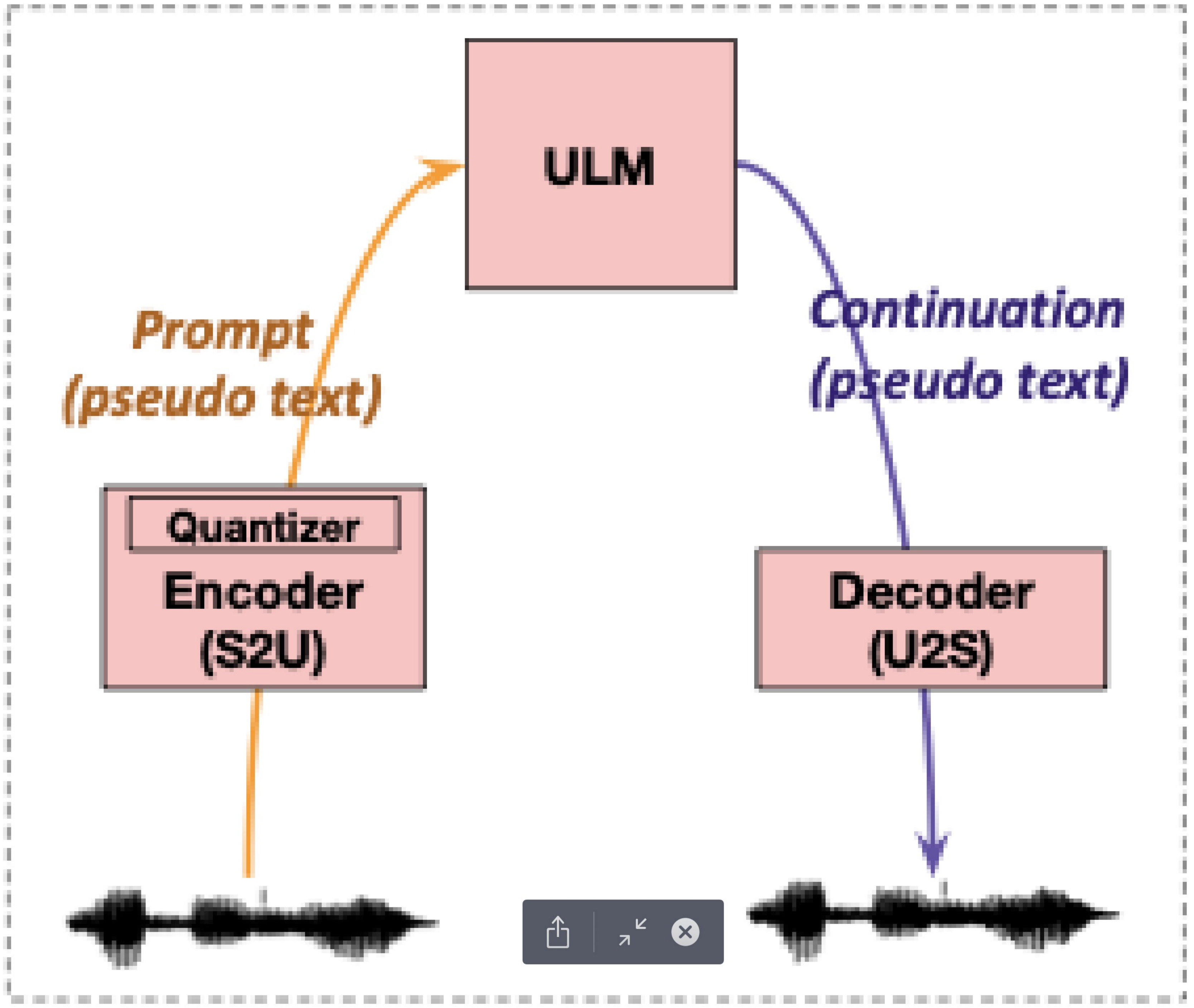
Discrete speech resynthesis



Unconditional speech generation



Conditional speech generation



GSLM – Evaluation metrics (1):

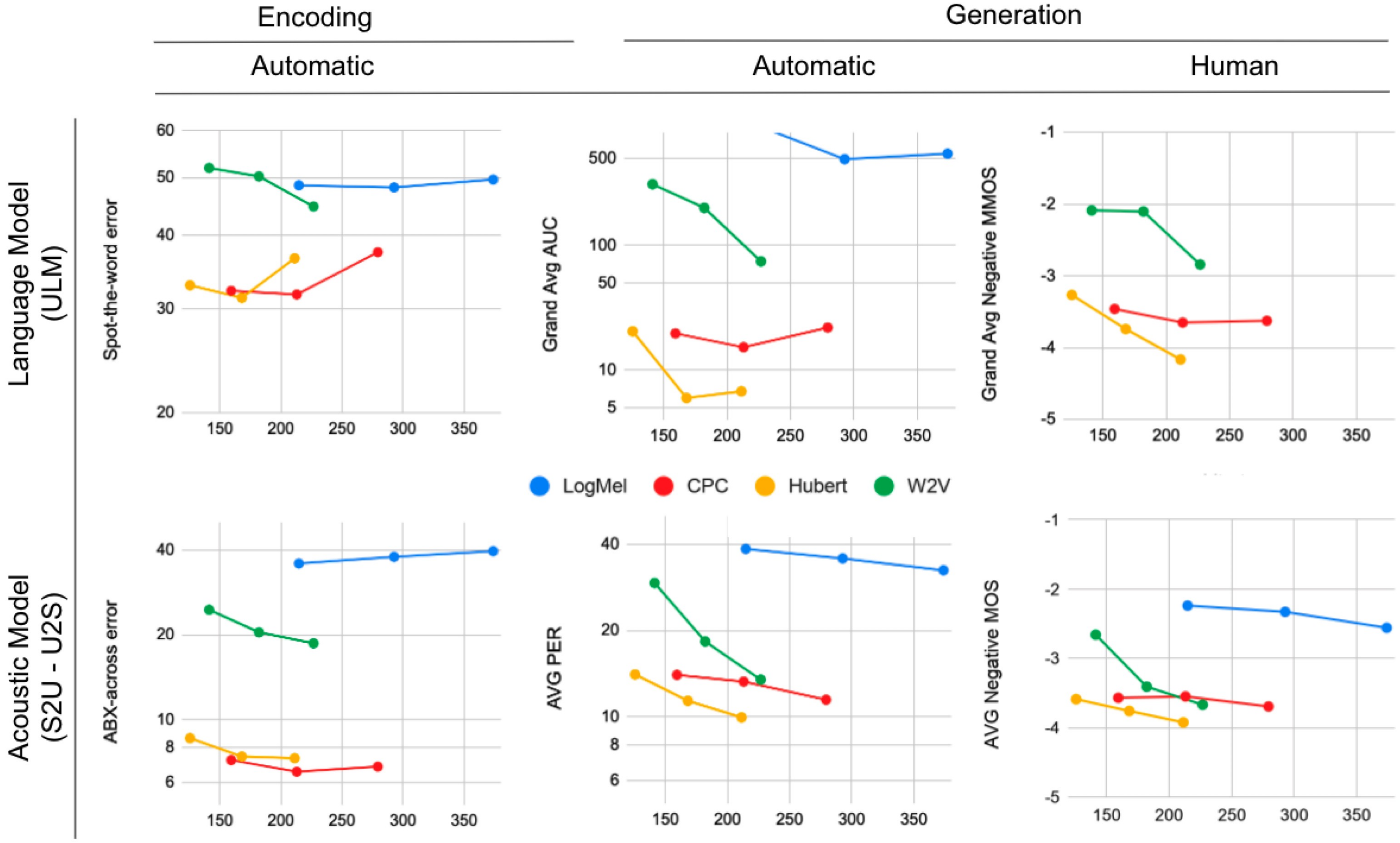
- <u>We need the evaluation metrics to be</u>:
  1. Independent of the learned discrete unit.
  2. Evaluate the intelligibility, diversity, and meaningfulness of the generated content.

- <u>ASR-based automatic evaluation</u>:
  + Use an off-the-shelf ASR system to convert the produced audio into text.
  + Evaluate the resulting text using a pretrained LM.

GSLM – Evaluation metrics (2):

- Speech resynthesis intelligibility: PER
- Speech generation quality and diversity: AUC (perplexity and diversity – more details are in the paper).
- Acoustic level: ABX error
- Lexical level: spot-the-word accuracy
- Subjective human evaluation (we use –ive scores; so the lower the better):
  1. Mean Opinion Score (MOS): Measure intelligibility of audio.
  2. Meaningfulness MOS (MMOS): Evaluates grammar and meaning.

GSLM – Results (the lower the better for all metrics):



GSLM – Listen to samples:

Thank you