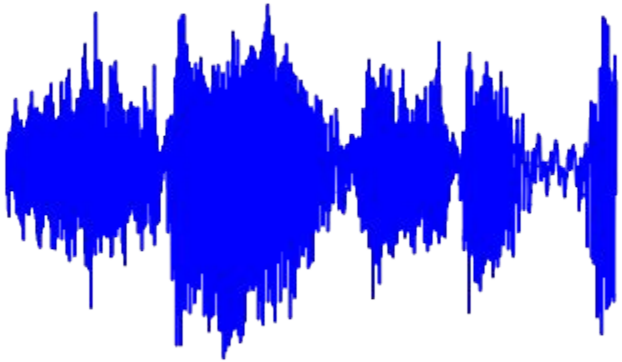
Variable-rate hierarchical CPC leads to acoustic unit discovery in speech

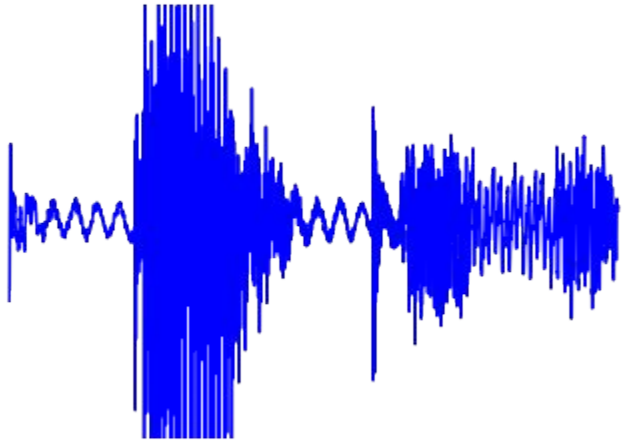
36th Conference on Neural Information Processing Systems (NeurIPS 2022)

Santiago Cuervo[1, 2],
Adrian Łańcucki[3], Ricard Marxer[2], Paweł Rychlikowski[1], Jan Chorowski[4],

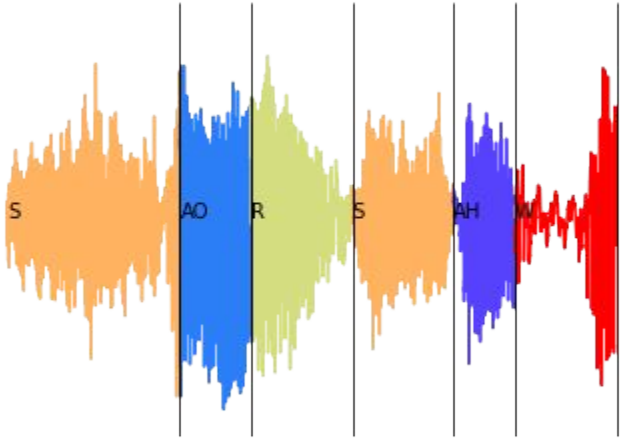
[1] *University of Wroclaw, Poland*
[2] *Université de Toulon, Aix Marseille Univ, CNRS, LIS, France*
[3] *NVIDIA, Poland*
[4] *Pathway, France*

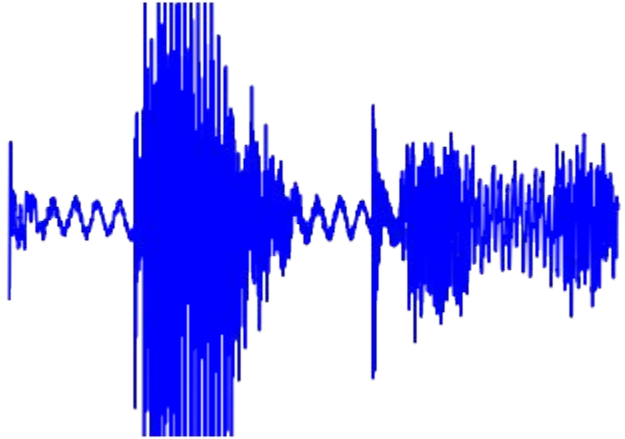
Our setup: unsupervised acoustic unit discovery



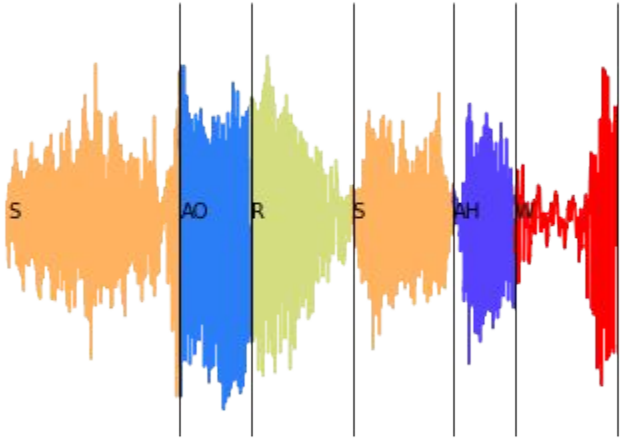


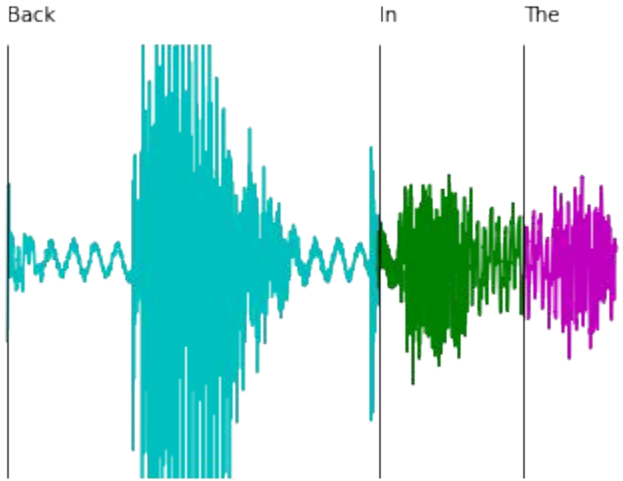
Our setup: unsupervised acoustic unit discovery



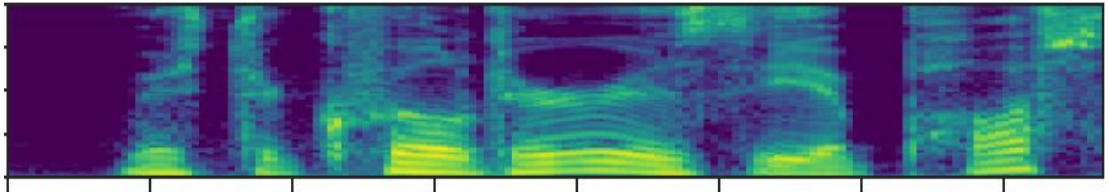


Our setup: unsupervised acoustic unit discovery

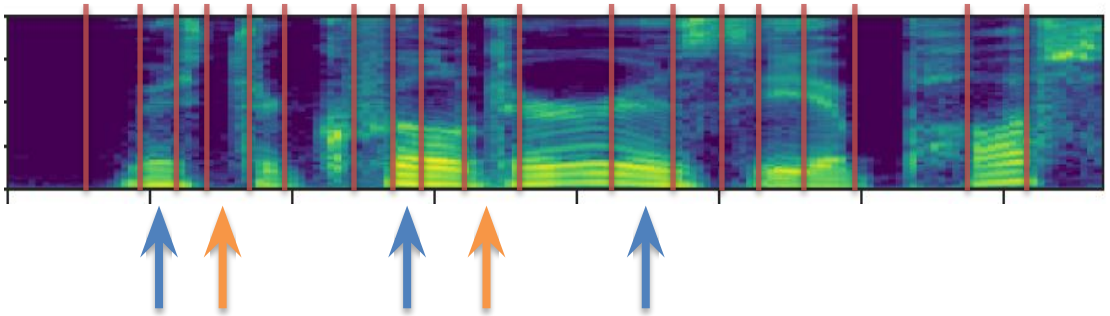




Our goal: unsupervised unit discovery

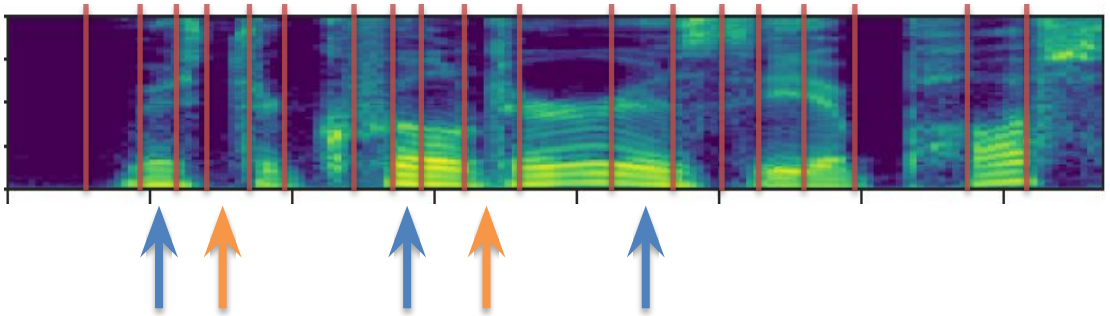


Our goal: unsupervised unit discovery



We want to learn to representations that allow us to segment and cluster speech data in order to discover information bearing units (eg. phonemes, syllables, words, etc.)

Our goal: unsupervised unit discovery

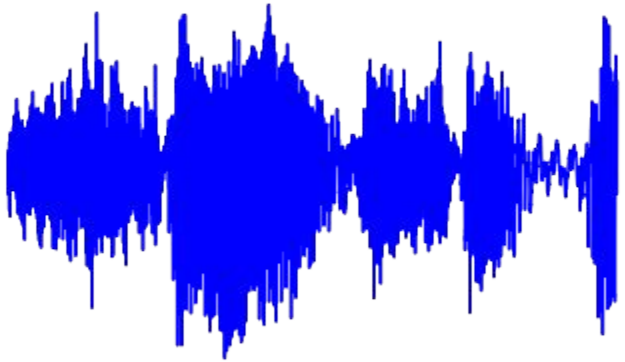


We want to learn to representations that allow us to segment and cluster speech data in order to discover information bearing units (eg. phonemes, syllables, words, etc.)

Motivations:

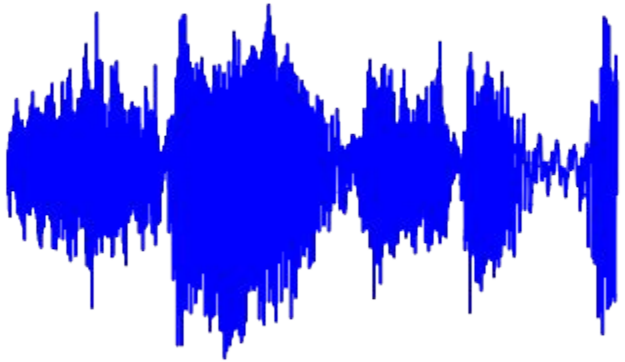
- Discrete units enable text-less NLP
- Reduced sampling rates and therefore reduced computational costs
- How can discrete units emerge from continuous perceptual data? It might provide hints on language acquisition.

Insight: speech is a hierarchical signal



Insight: speech is a hierarchical signal

Acoustic

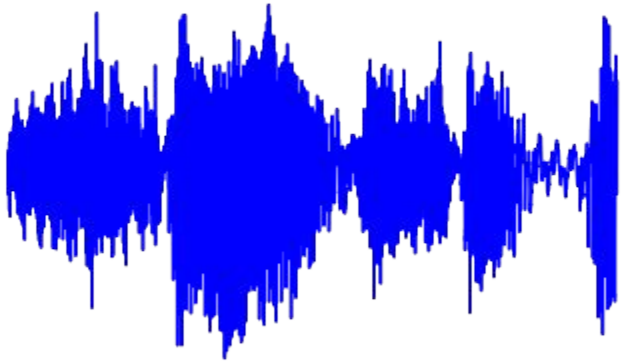


Insight: speech is a hierarchical signal

Phonetic

IH T S AO R S AH W

Acoustic



Insight: speech is a hierarchical signal

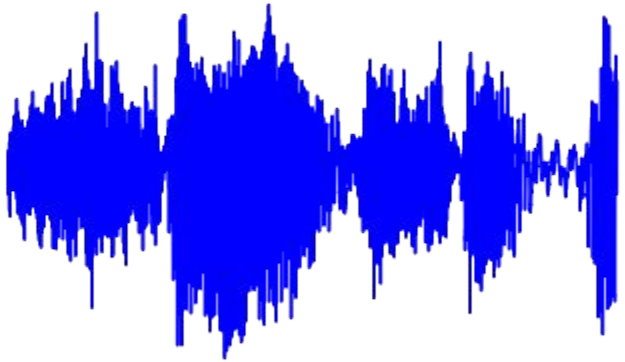
Lexical

It’s hours away

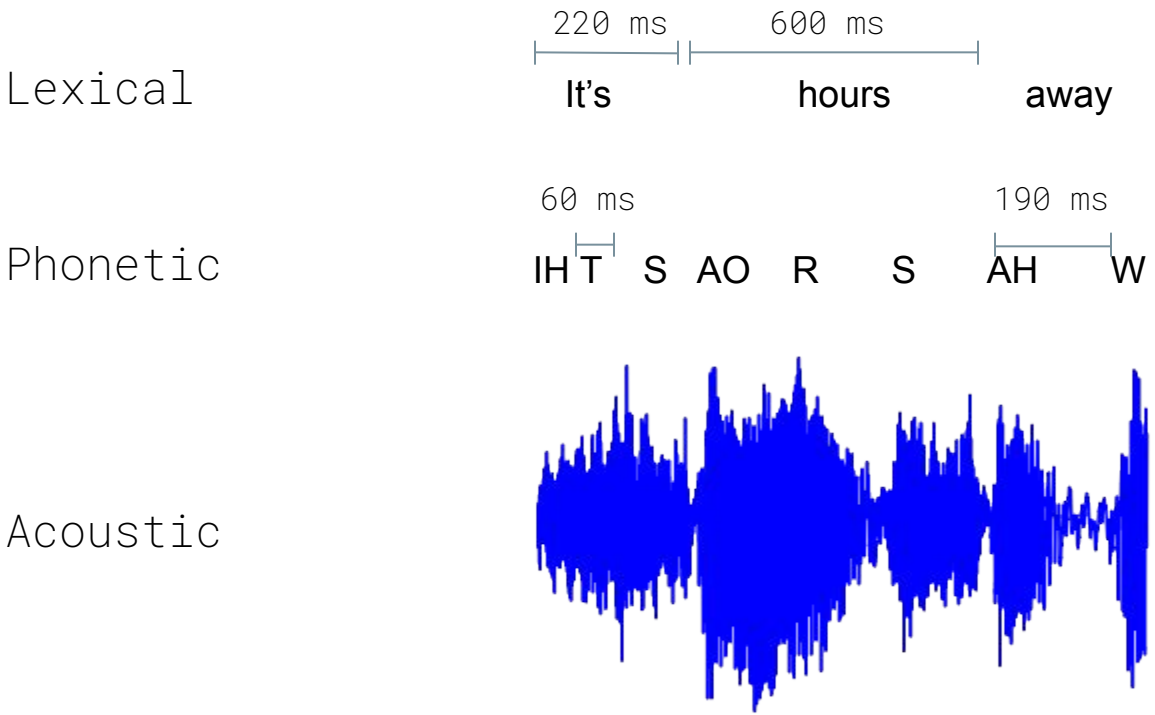
Phonetic

IH T S AO R S AH W

Acoustic



Insight: speech is a hierarchical signal with non-uniform information density



Insight: speech is a hierarchical signal with non-uniform information density

Do our representation learners account for it?

Insight: speech is a hierarchical signal with non-uniform information density

Do our representation learners account for it?

Hierarchy? Only partially. Layers in deep neural nets learn an internal hierarchy of concepts, but the training criteria are only applied at the top level (layer).

Insight: speech is a hierarchical signal with non-uniform information density

Do our representation learners account for it?

Hierarchy? Only partially. Layers in deep neural nets learn an internal hierarchy of concepts, but the training criteria are only applied at the top level (layer).

Information density? Not really. Models usually process data at fixed input-driven rates (eg. pixels in images, frames in speech).

Insight: speech is a hierarchical signal with non-uniform information density

Do our representation learners account for it?

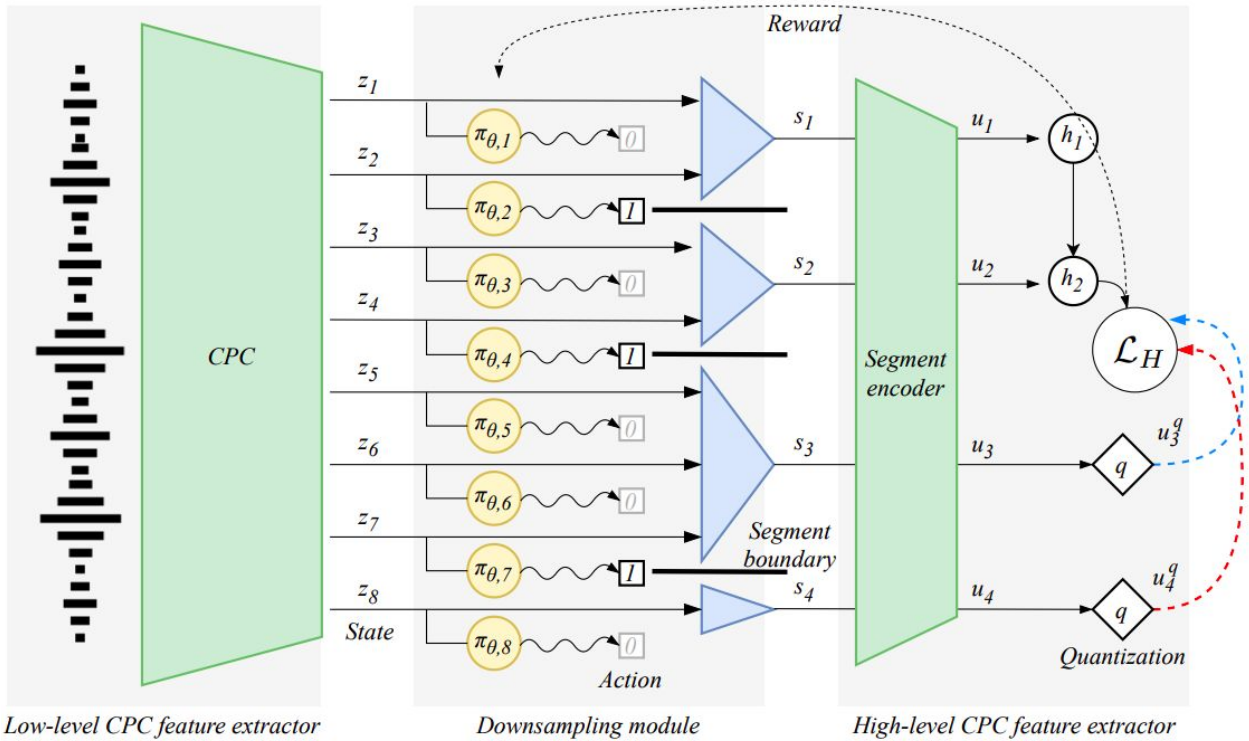
Hierarchy? Only partially. Layers in deep neural nets learn an internal hierarchy of concepts, but the training criteria are only applied at the top level (layer).

Information density? Not really. Models usually process data at fixed input-driven rates (eg. pixels in images, frames in speech).

So we work on representation learners which:

1. At each level extract features at different non-uniform information dependent rates.
2. At each level apply a training criterion.

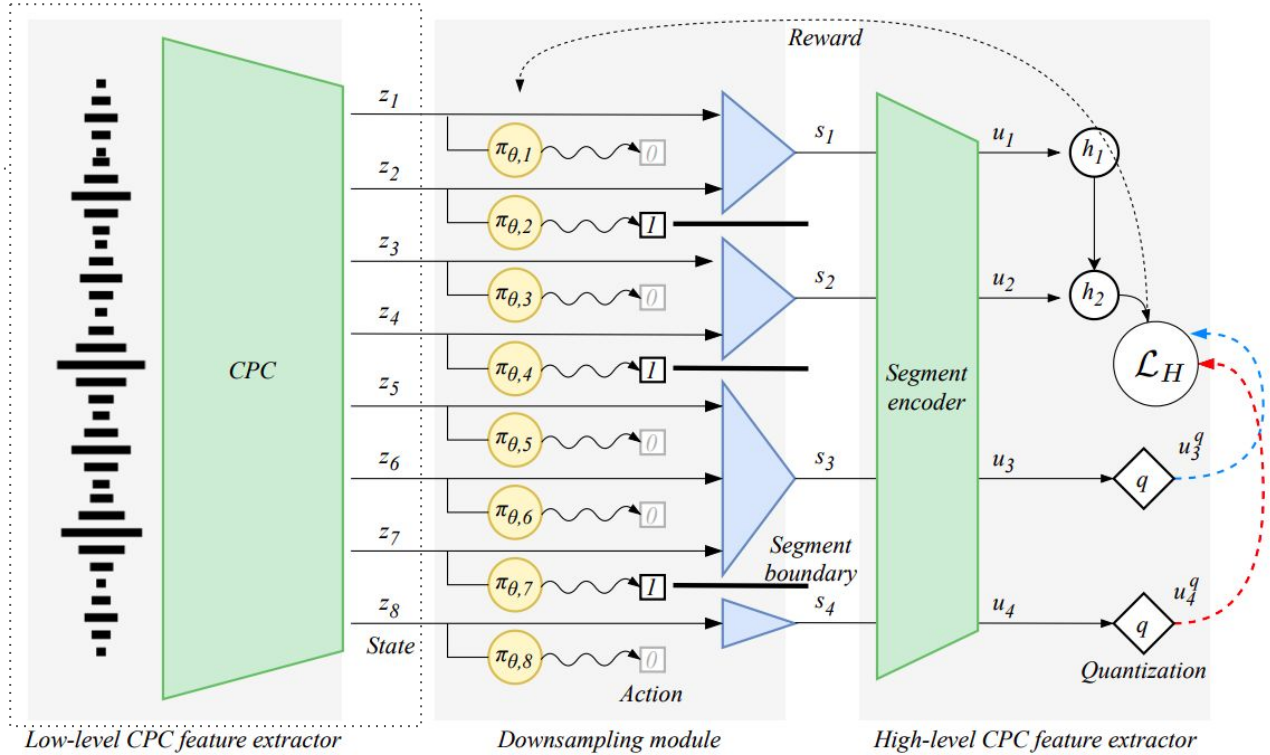
High-level description of our current approach



S. Cuervo et al. "Variable-rate hierarchical CPC leads to acoustic unit discovery in speech"", NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHIsE

High-level description of our current approach

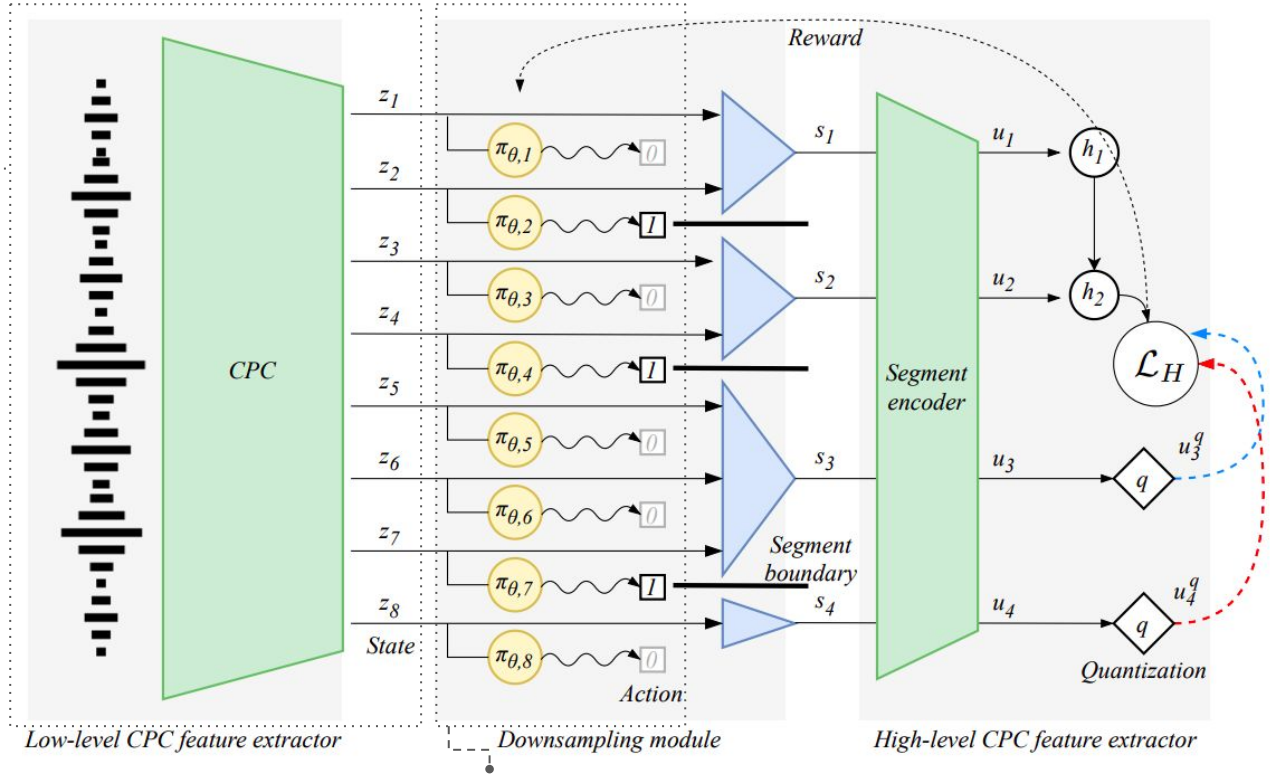
A low-level representation learner extracts features at uniform sampling rates.



S. Cuervo et al. "Variable-rate hierarchical CPC leads to acoustic unit discovery in speech"", NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHIsE

High-level description of our current approach

A low-level representation learner extracts features at uniform sampling rates.

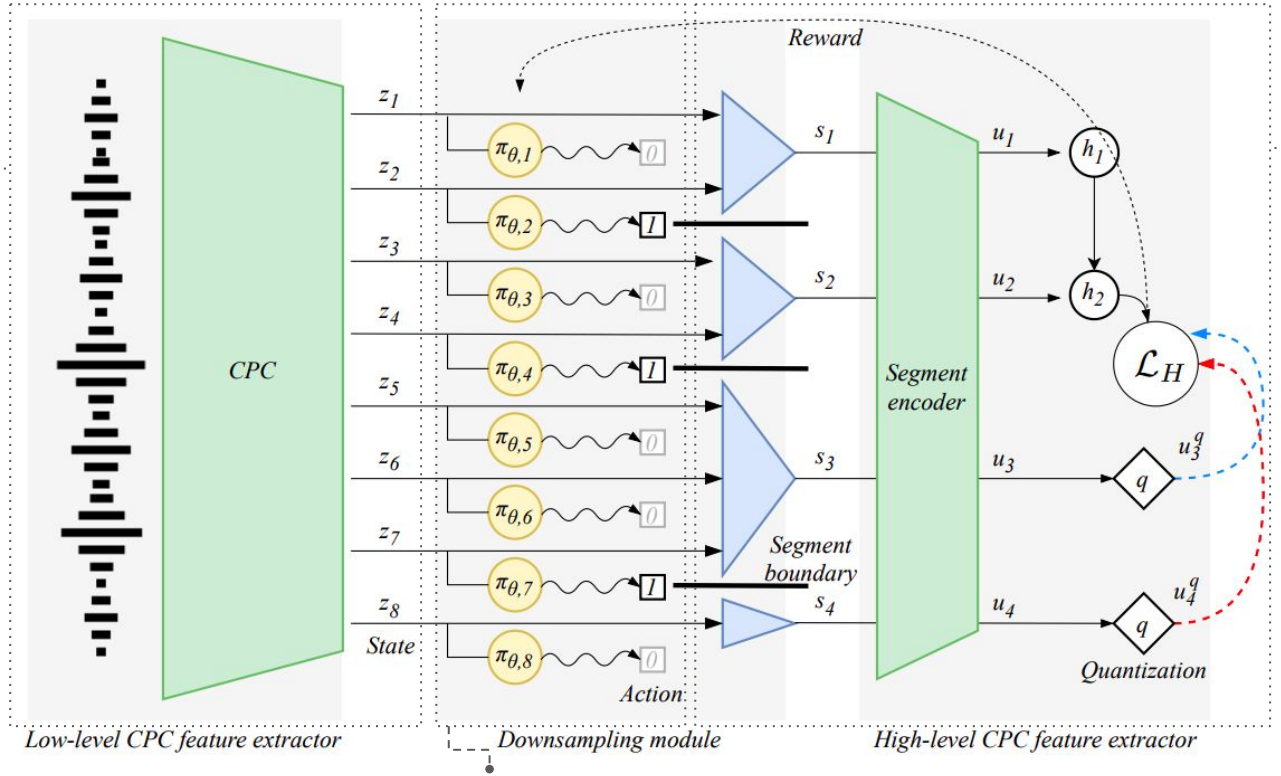


A reinforcement learning agent segments the signal and compresses the low-level feature segments into pseudo-unit representations.

S. Cuervo et al. "Variable-rate hierarchical CPC leads to acoustic unit discovery in speech"", NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHIsE

High-level description of our current approach

A low-level representation learner extracts features at uniform sampling rates.

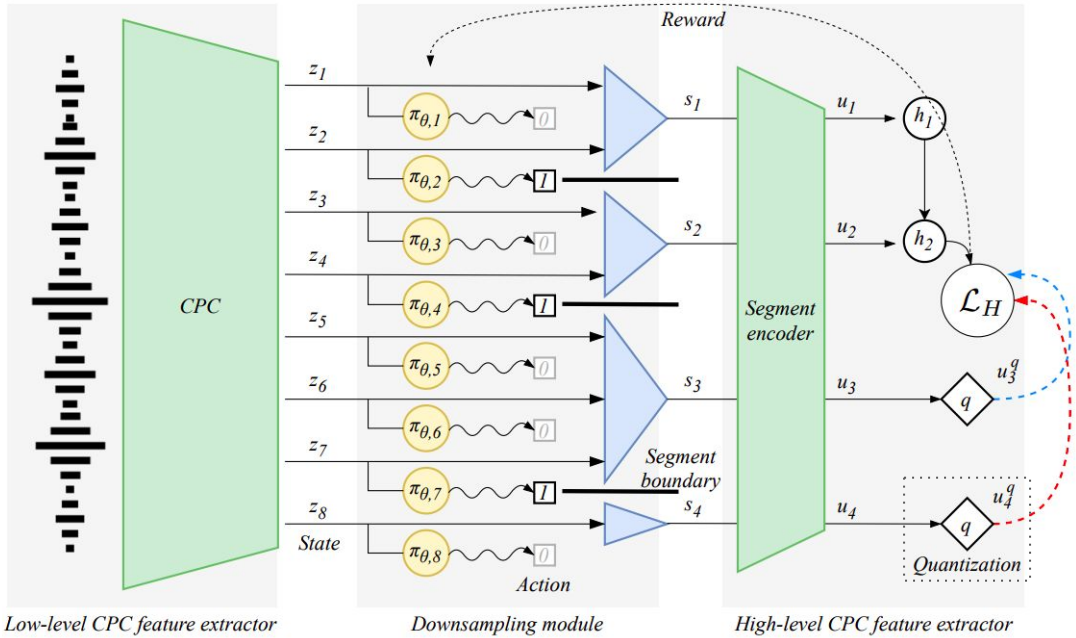


A high-level representation learner models the signal at the level of pseudo-units in a LM fashion. Its predictive power is used as reward for the RL agent.

A reinforcement learning agent segments the signal and compresses the low-level feature segments into pseudo-unit representations.

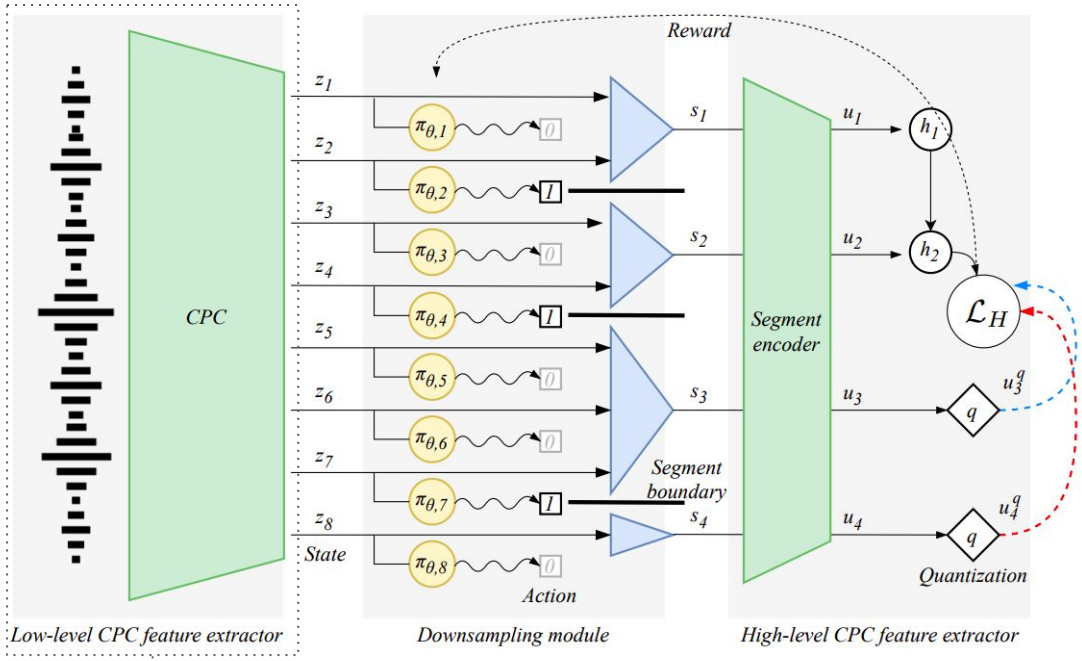
S. Cuervo et al. "Variable-rate hierarchical CPC leads to acoustic unit discovery in speech"", NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHIsE

Variable-rate hierarchical Contrastive Predictive Coding **training**



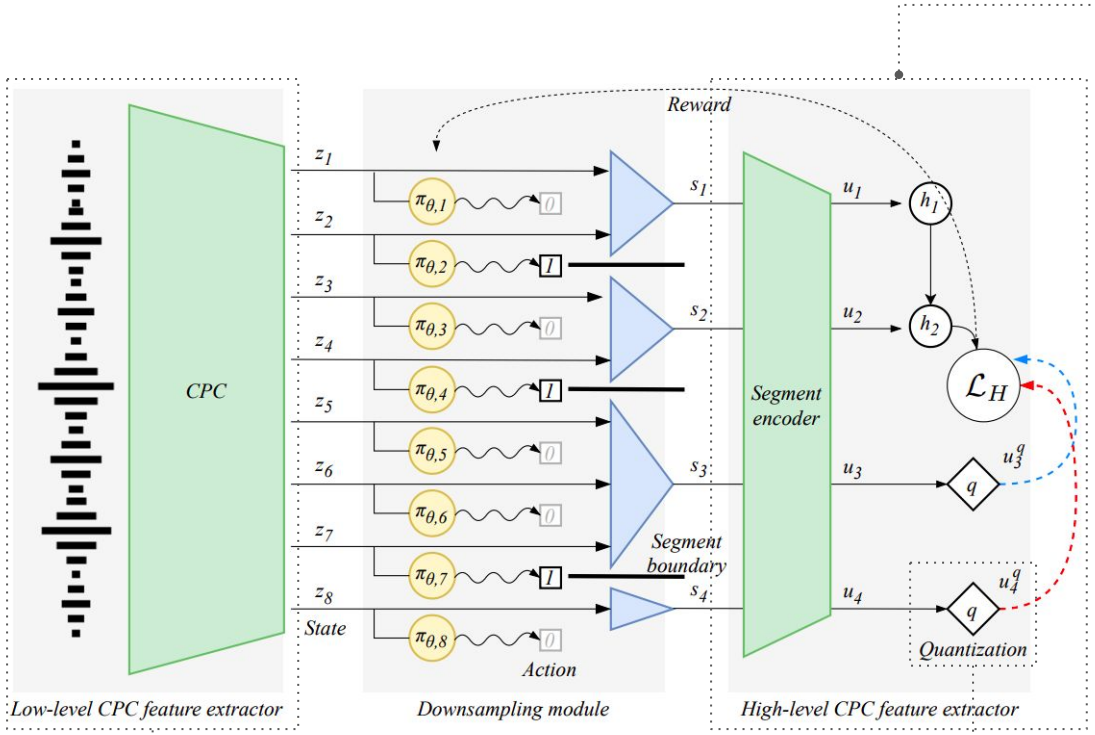
$$\mathcal{L} = \mathcal{L}_L + \mathcal{L}_H + \mathcal{L}_Q + \mathcal{L}_\pi + \mathcal{L}_{\bar{l}}$$

Variable-rate hierarchical Contrastive Predictive Coding **training**



$$\mathcal{L} = \mathcal{L}_L + \mathcal{L}_H + \mathcal{L}_Q + \mathcal{L}_\pi + \mathcal{L}_{\bar{l}}$$

Variable-rate hierarchical Contrastive Predictive Coding **training**



$$\mathcal{L}_H = -\sum_k \sum_{m=1}^{M} \frac{\exp\left(p_m^T u_{k+m}^q\right)}{\sum_{i \in \{k+m-1, k+m+1\}} \exp\left(p_m^T u_i^q\right)}$$

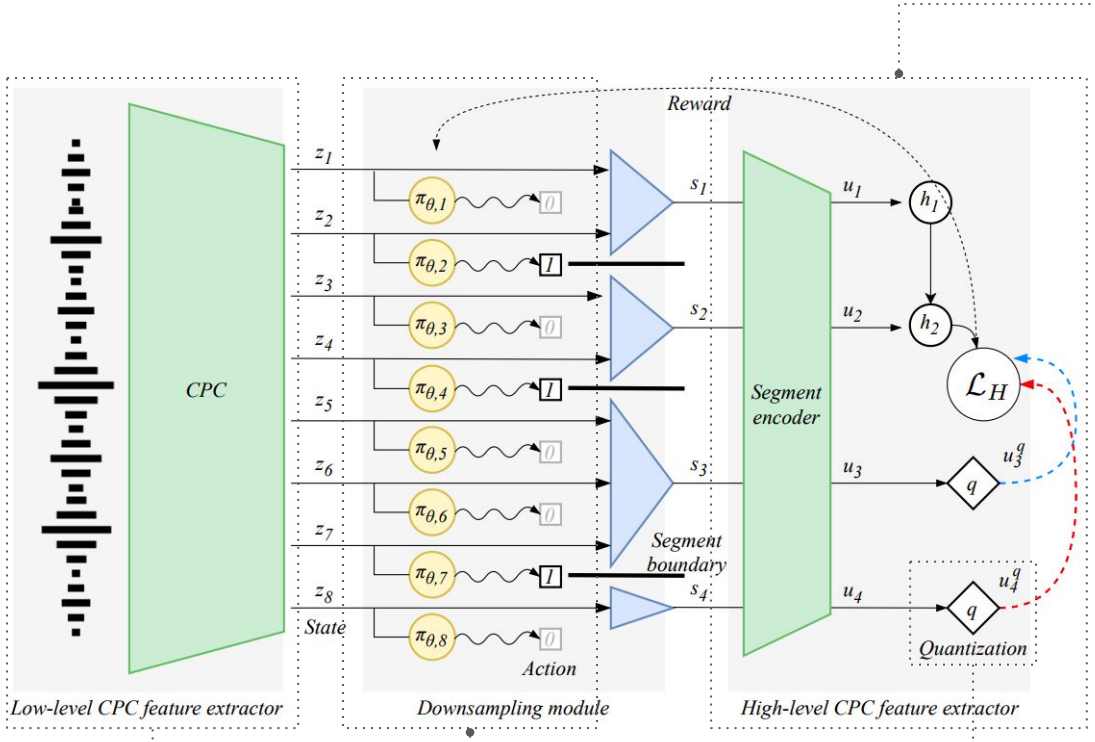
It enforces a **prior of discreteness** into the high-level representations through quantized prediction targets:

$$u_k^q = e_i : \min_i ||u_k - e_i||$$

, and adjacent negative sampling (contrastiveness).

$$\mathcal{L} = \mathcal{L}_L + \mathcal{L}_H + \mathcal{L}_Q + \mathcal{L}_\pi + \mathcal{L}_{\bar{l}}$$

Variable-rate hierarchical Contrastive Predictive Coding **training**



$$\mathcal{L} = \mathcal{L}_L + \mathcal{L}_H + \mathcal{L}_Q + \mathcal{L}_\pi + \mathcal{L}_{\bar{l}}$$

$$\mathcal{L}_H = -\sum_k \sum_{m=1}^{M} \frac{\exp\left(p_m^T u_{k+m}^q\right)}{\sum_{i \in \{k+m-1, k+m+1\}} \exp\left(p_m^T u_i^q\right)}$$

It enforces a **prior of discreteness** into the high-level representations through quantized prediction targets:

$$u_k^q = e_i : \min_i ||u_k - e_i||$$

, and adjacent negative sampling (contrastiveness).

It is a reinforcement learning policy that minimizes the high-level loss:

$$\mathcal{L}_\pi = \mathbb{E}_b[\mathcal{L}_H(b)|z, \theta] = \sum_b \pi_\theta(b|z)\mathcal{L}_H(b)$$

therefore has to predict segments which satisfy the high-level prior of discreteness. It also enforces a **prior of average unit length:**

$$\mathcal{L}_{\bar{l}} = \left\| \mathbb{E}_{b_t \sim \pi_\theta} \left[\sum_{t=1}^{\bar{l}} b_t \right] - 1 \right\| = \left\| \left(\sum_{t=1}^{\bar{l}} \pi_\theta(b_t) \right) - 1 \right\|$$

Low-level representations evaluation

- We evaluate the downstream performance of low-level representations in the tasks of **frame-wise linear phone classification** and **CTC phone transcription** in the test split of LibriSpeech train-clean-100, and the **ABX task** in the ZeroSpeech 2021 dev-clean set.
- Overall our method improves phone discriminability when compared against multiple CPC-based hierarchical and non-hierarchical baselines, including a hierarchical model that uses supervised phone boundaries for downsampling

Architecture	Model	Frame accuracy ↑	Phone accuracy ↑	ABX within ↓	ABX across ↓
Single level	CPC [Rivière et al., 2020]	67.50	83.20	6.68	8.39
	ACPC [Chorowski et al., 2021]	68.60	83.33	5.37	7.09
	Two-level CPC no downsampling	67.49	83.38	6.66	8.34
Multi-level	SCPC [Bhati et al., 2021]	43.79	68.38	20.18	16.26
	Two-level CPC w. downsampling	67.92	83.39	6.66	8.32
	mACPC [Cuervo et al., 2022]	70.25	83.35	5.13	6.84
	Ours	**72.57**	**83.95**	**5.08**	**6.72**
	Downsampling (supervised)	71.01	84.70	5.07	6.68

High-level representations evaluation

- We evaluate the downstream performance of high-level representations in the tasks of **phone transcription** in the test split of LibriSpeech train-clean-100. We additionally report the average sampling rate of the representations to evaluate compression.
- Our model gives the best results in phone accuracy and has the lowest average sampling rate among unsupervised methods with variable downsampling.

Downsampling	Model	Avg. sampling rate (Hz) ↓	Phone accuracy ↑
None	Two-level CPC no downsampling	100	83.41
Constant	Two-level CPC with downsampling	10.94	67.75
Variable	SCPC [Bhati et al., 2021]	15.91	55.49
	mACPC [Cuervo et al., 2022]	14.47	69.66
	Ours	**12.32**	**78.93**
	Downsampling (supervised)	10.87	85.74

Phone segmentation evaluation

Results on the test split of LibriSpeech train clean 100 and TIMIT test split. Our model produces segmentations competitive with the state-of-the-art, while being robust to non-speech events.

Dataset	Architecture	Model	Precision	Recall	F1	R-val
LibriSpeech clean 100	Single level	[Kreuk et al., 2020]	61.12	82.53	70.23	61.87
	Multi-level	mACPC [Cuervo et al., 2022]	59.15	**83.17**	69.13	57.71
		SCPC [Bhati et al., 2021]	64.05	83.11	72.35	66.40
		Ours	**79.94**	77.92	**78.91**	**81.98**
TIMIT (non-speech removed)	Single level	[Kreuk et al., 2020]	84.80	**85.77**	85.27	87.35
	Multi-level	mACPC [Cuervo et al., 2022]	84.63	84.79	84.70	86.86
		SCPC [Bhati et al., 2021]	**85.31**	85.36	**85.31**	**87.38**
		Ours	80.08	81.40	80.73	83.50

Conclusions & Where do we go from here?

Important takeaways:

- We have shown that accounting for the structure of the signal (hierarchy and spatial distribution of information) improves disentanglement of frame-level representations.
- Our objective function incorporating soft constraints of discreteness and average unit-duration leads to the unsupervised discovery of unit-boundaries that coincide with a human-made phonetic segmentation.

Interesting future research directions:

- Further analyze the effect of top-down feedback on the representations.
- Explore other high-level tasks to improve the quality of high-level representations.
- Going beyond phonetic: discovering higher-level units.

Thank you

Paper: https://openreview.net/pdf?id=Jk8RVjnHlsE
Code: https://github.com/chorowski-lab/hCPC