JTT Seminar. TALEP

Variable-rate hierarchical representation learning in speech

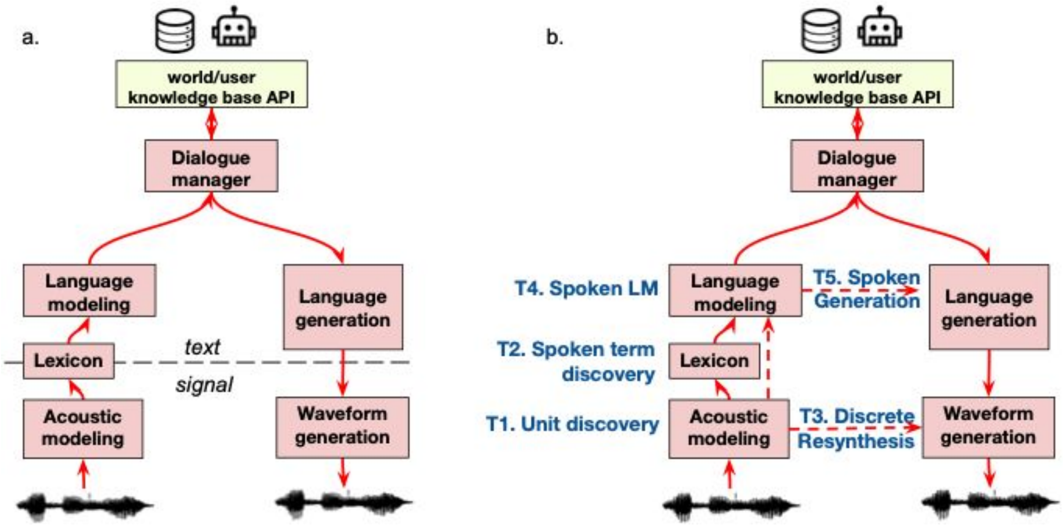
Santiago Cuervo, PhD student

UNIVERSITÉ DE TOULON
Aix*Marseille université
LIS Laboratoire d'Informatique & Systèmes
cnrs

Problem

Background: Zero Resource Speech Challenge

The end-game is to build a spoken dialogue system directly from raw audio recordings. **No text involved.**



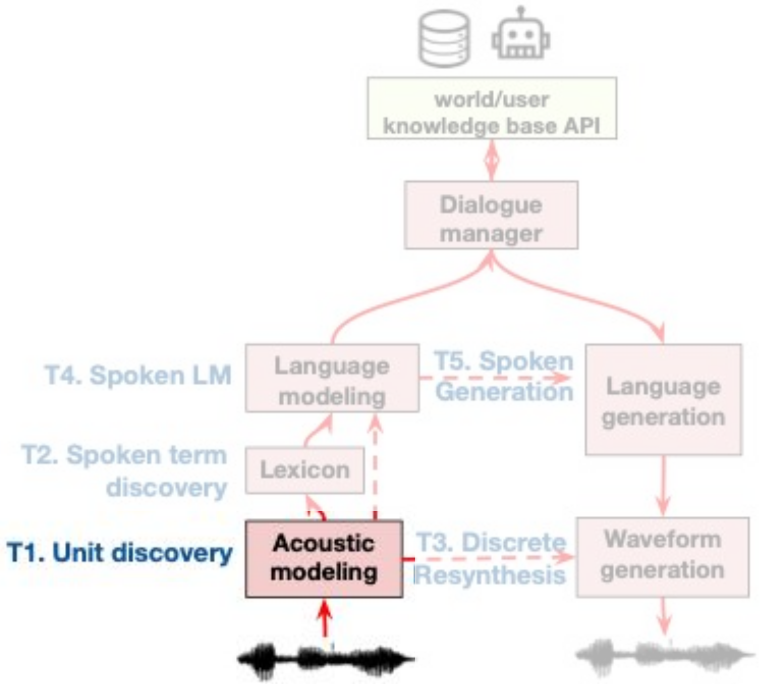
a. Traditional pipeline for a spoken assistant based on textual resources.
b. Pipeline and tasks for the Zero Resource Speech Challenge.
Taken from https://www.zerospeech.com/

Background: Zero Resource Speech Challenge

Motivations:

- Open up applications in low-resource languages, therefore making AI dialogue systems more inclusive.
- Potentially more expressive than text-based systems.
- Could provide hints on human language acquisition.

We focus on Task 1: acoustic unit discovery



We aim to learn representations of speech sounds that retain linguistically relevant information and discard linguistically irrelevant acoustic information, like speaker voice type or recording conditions.

In text based systems, such representations are *phonemes* or *characters*.

What defines a unit: Contrastiveness

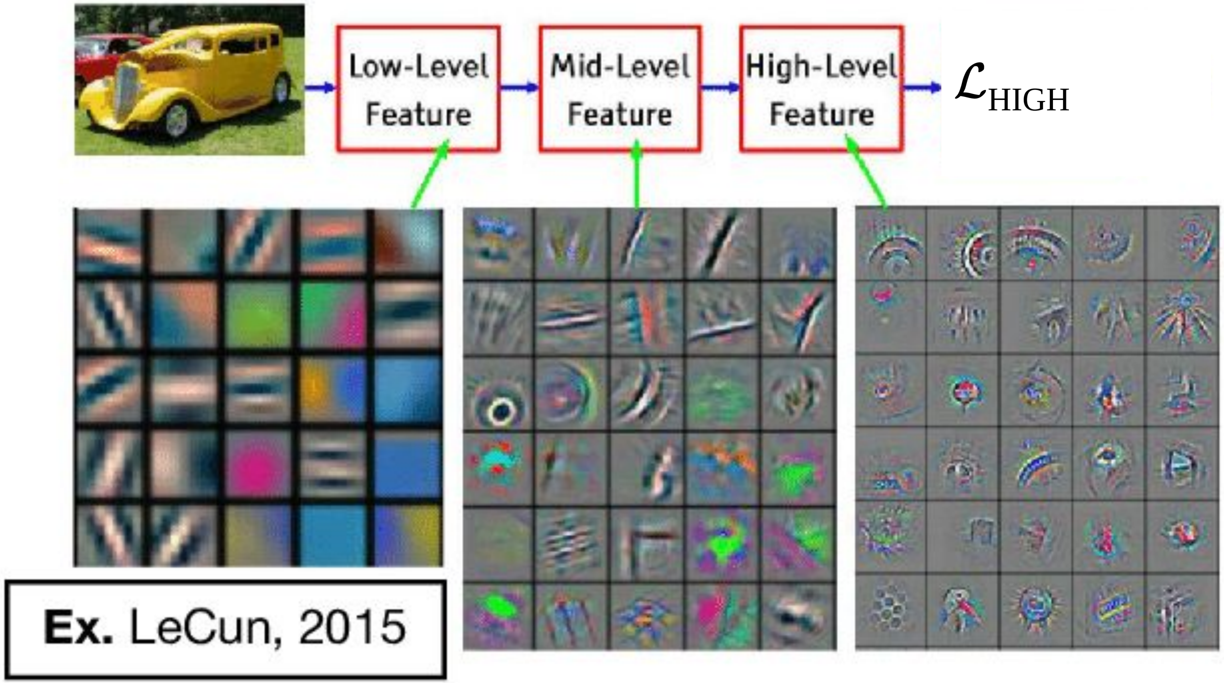
Representations do not have to be phone(me)s, but they should support the same key function: phonemic contrast. *Phonemes* are defined as the smallest element of speech that make a difference in meaning between words (e.g., /bit/ versus /but/).

The Zero Speech benchmark requires representations to distinguish pairs of phonemes, while ignoring non-linguistic variations. Discriminability is computed by running an **ABX discrimination test**:

> *The ABX discriminability of category **A** from category **B** is the probability that **a** and **x** are closer than **b** and **x**, according to some distance metric, when **a** and **x** are from category **A** and **b** is from category **B**.*

Our method

Motivation: emergent hierarchical learning from top-down feedback in DNNs



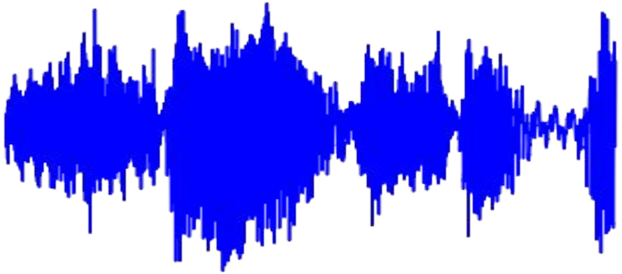
In deep neural networks an internal hierarchy of features (edges, shapes, object parts) emerges from optimizing a training criterion in terms of high-level features (eg. object categories).

Top-down feedback should also promote acoustic unit discovery

Lexical   It's   hours   away → $\mathcal{L}_{\mathrm{HIGH}}(\text{word features})$

Phonetic   IH T S AO R S AH W → $\mathcal{L}_{\mathrm{MID}}(\text{phone features})$

Acoustic



→ $\mathcal{L}_{\mathrm{LOW}}(\text{acoustic features})$

Top-down feedback through backpropagation

Challenge: how to obtain top-down feedback in a zero-resource setup where we only have access to speech recordings?

The core idea of our method: we learn a proxy to high-level features

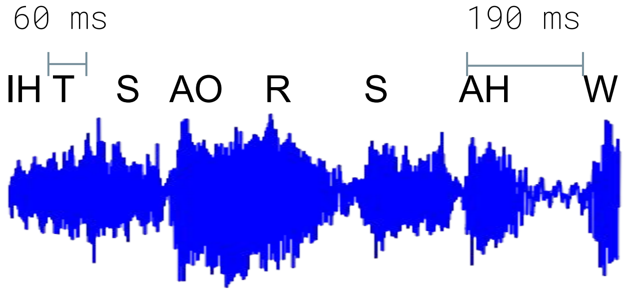
We learn to extract representations that satisfy three properties of linguistic high-level features (phonemes, syllables, words, etc.):

1. They are discrete (ish).

2. They allow for language modeling. ie. the past is predictive of the future.

May the Force be with ________

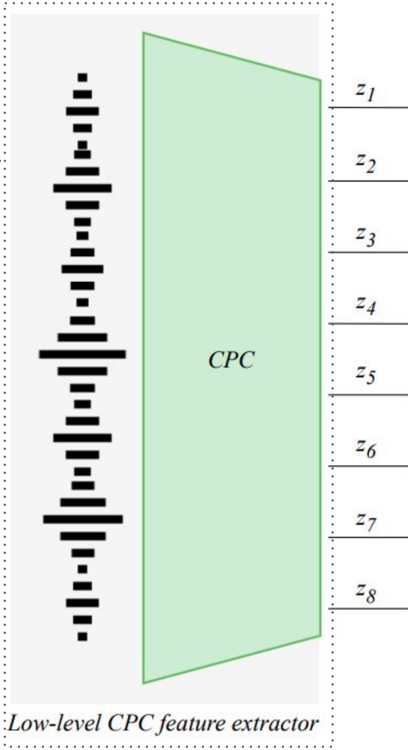
3. They have a non-uniform information density in time:



Our method: Variable-rate hierarchical CPC

S. Cuervo et al. “Variable-rate hierarchical CPC leads to acoustic unit discovery in speech”, NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHlsE

Our method: Variable-rate hierarchical CPC

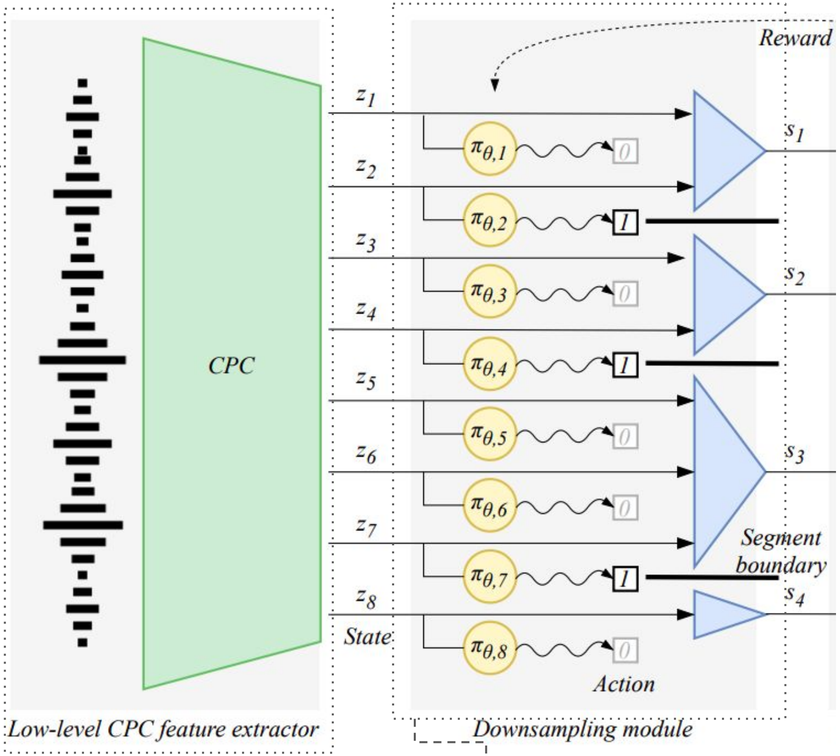


1. A typical representation learner extracts low-level acoustic features.

S. Cuervo et al. "Variable-rate hierarchical CPC leads to acoustic unit discovery in speech", NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHlsE

Our method: Variable-rate hierarchical CPC

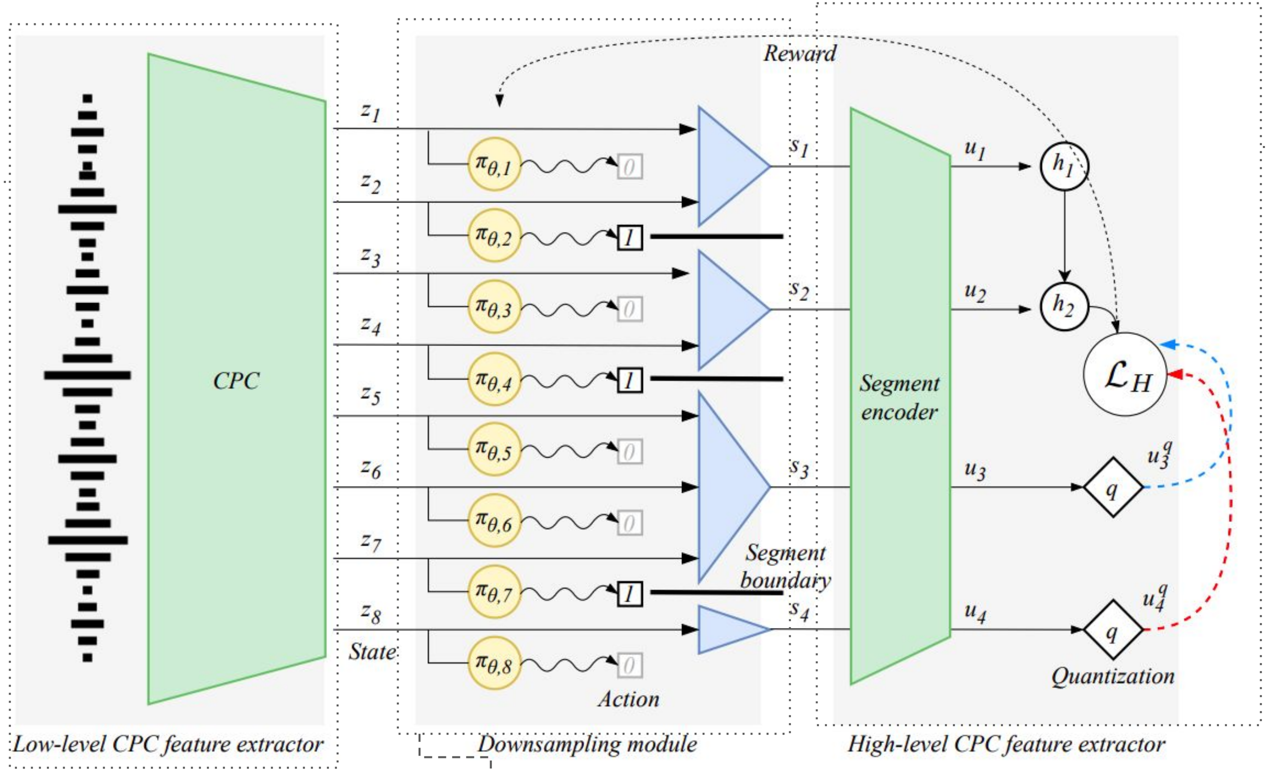
1. A typical representation learner extracts low-level acoustic features.



2. A learnable module downsamples the sequence of acoustic features to produce proxy high-level features with a non-uniform sampling rate.

S. Cuervo et al. "Variable-rate hierarchical CPC leads to acoustic unit discovery in speech", NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHlsE

Our method: Variable-rate hierarchical CPC



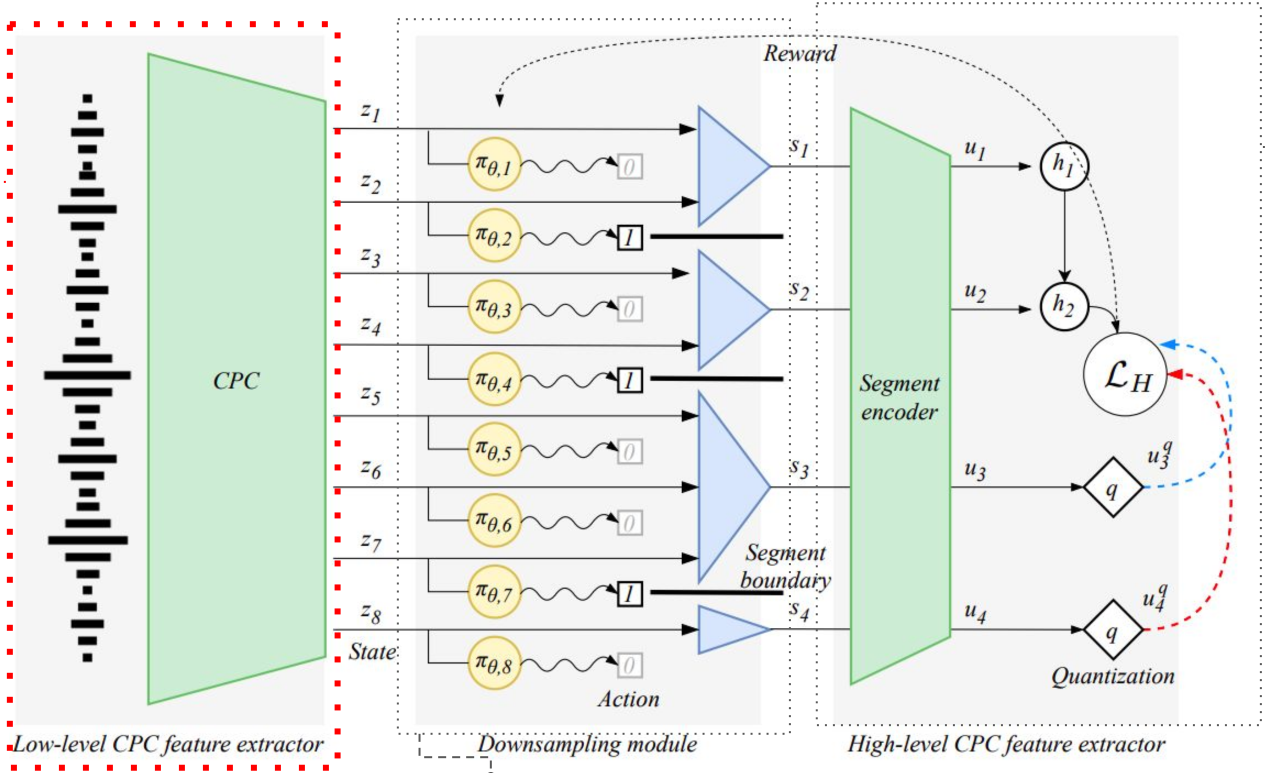
1. A typical representation learner extracts low-level acoustic features.

2. A learnable module downsamples the sequence of acoustic features to produce proxy high-level features with a non-uniform sampling rate

3. A high-level criterion at the level of proxy high-level features imposes the properties of discreteness and language modeling.

S. Cuervo et al. "Variable-rate hierarchical CPC leads to acoustic unit discovery in speech", NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHlsE

Our method: Variable-rate hierarchical CPC



1. A typical representation learner extracts low-level acoustic features.

2. A learnable module downsamples the sequence of acoustic features to produce proxy high-level features with a non-uniform sampling rate

3. A high-level criterion at the level of proxy high-level features imposes the properties of discreteness and language modeling.

S. Cuervo et al. “Variable-rate hierarchical CPC leads to acoustic unit discovery in speech”, NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHlsE

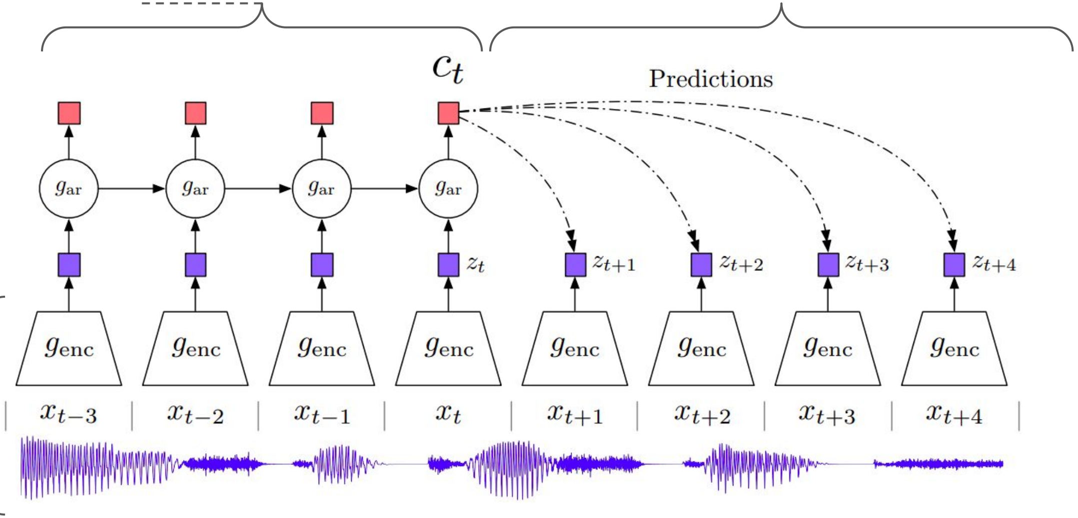
Contrastive Predictive Coding (CPC)

An autoregressive model (eg. an RNN, transformer) summarizes the history of latents into a context representation

We predict the future based on the context. The model is trained through Noise Contrastive Estimation (NCE): the prediction should be closer to the target than to some randomly sampled distractors

$$-\sum_{n=1}^{N} \frac{\exp(p_n^T z_{t+n})}{\sum_{z_i \in \mathcal{N}} \exp(p_n^T z_i)}$$

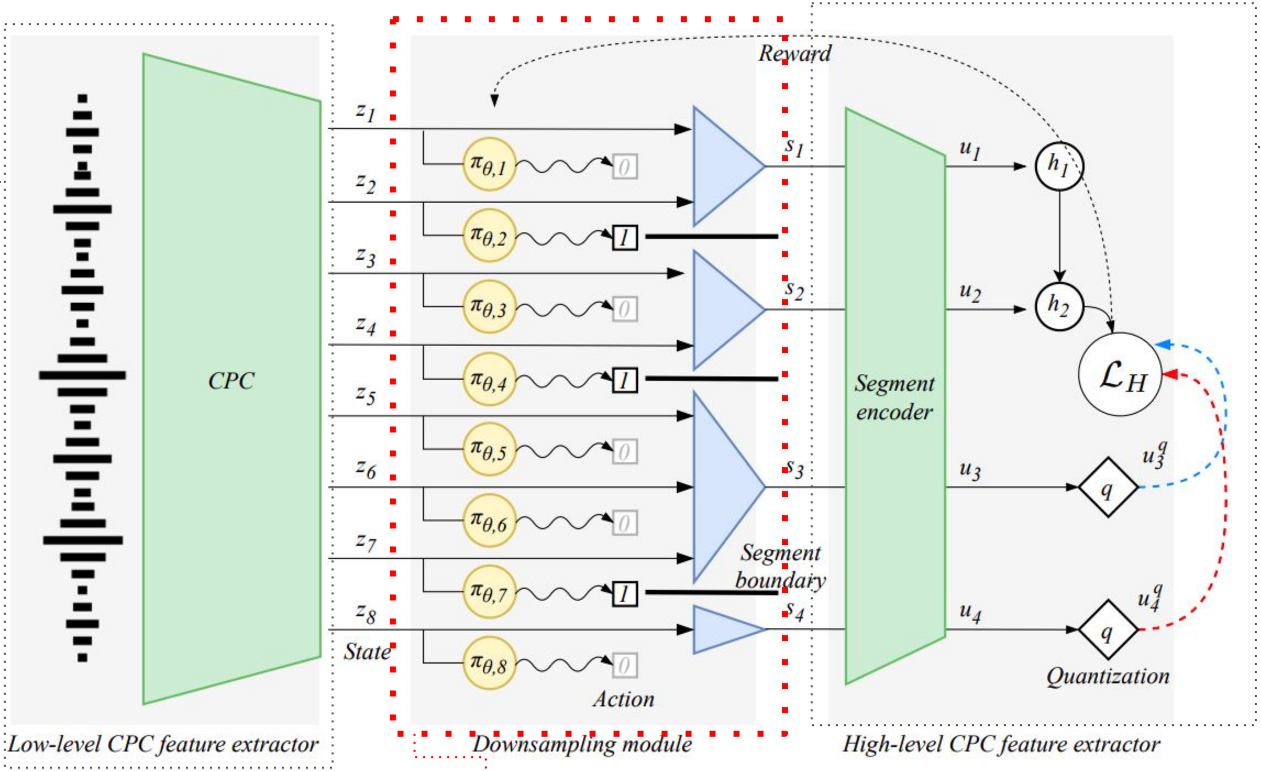
A strided convolutional encoder transforms the input sequence into a sequence of latent representations



van den Oord et al (2018). “Representation learning with contrastive predictive coding”. https://arxiv.org/abs/1807.03748

Our method: Variable-rate hierarchical CPC

1. A typical representation learner extracts low-level acoustic features.



2. A learnable module downsamples the sequence of acoustic features to produce proxy high-level features with a non-uniform sampling rate

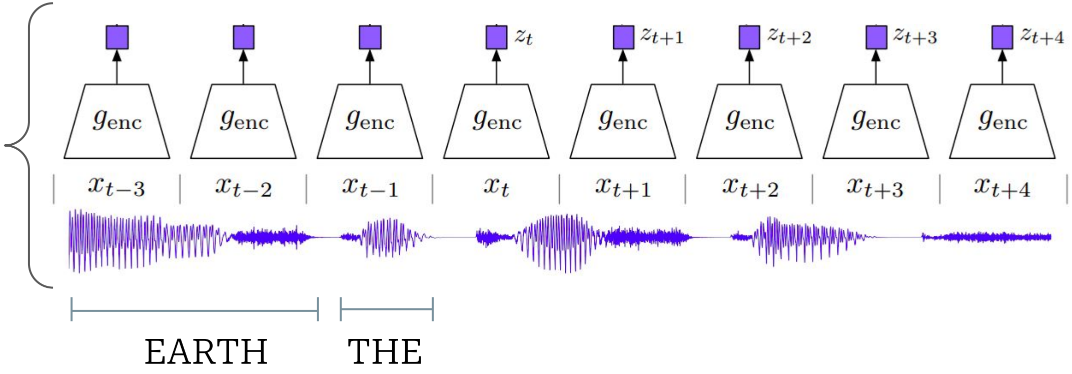
3. A high-level criterion at the level of proxy high-level features imposes the properties of discreteness and language modeling.

S. Cuervo et al. "Variable-rate hierarchical CPC leads to acoustic unit discovery in speech", NeurIPS 2022. https://openreview.net/pdf?id=Jk8RVjnHlsE

Extracting proxy features at a non-uniform sampling rate

We need to match the **variable sampling rate** of the features at the level we aim to model.

In CPC the convolutional encoder produces uniformly sampled representations, which is fine for the continuous acoustic features.



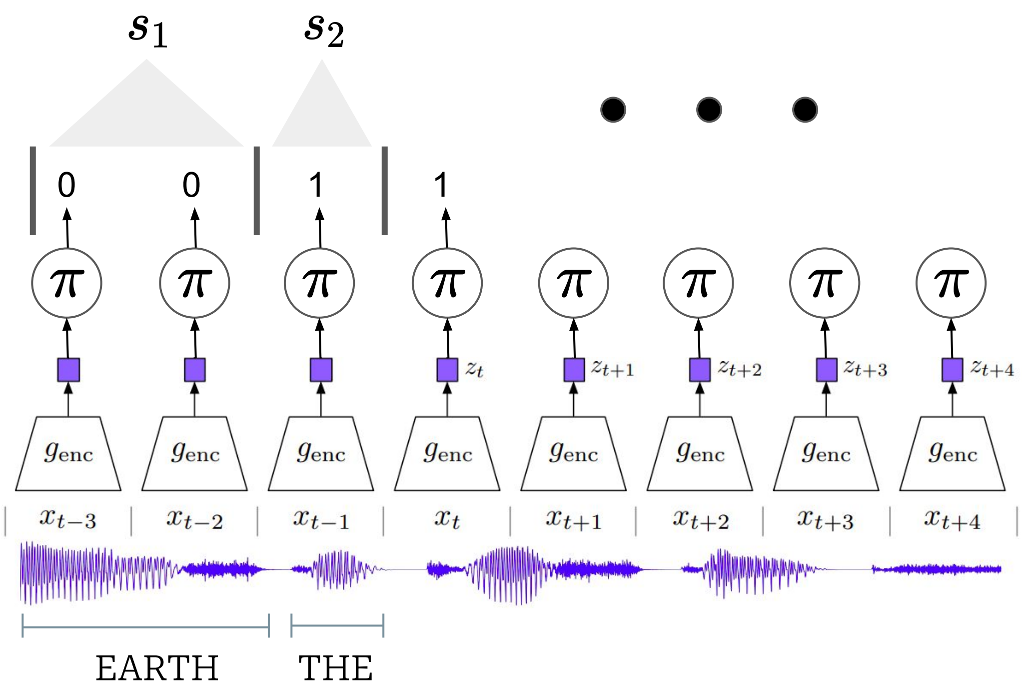
But high-level features have variable lengths. Eg. articles are shorter than most words and vowels last longer than consonants.

We learn to downsample

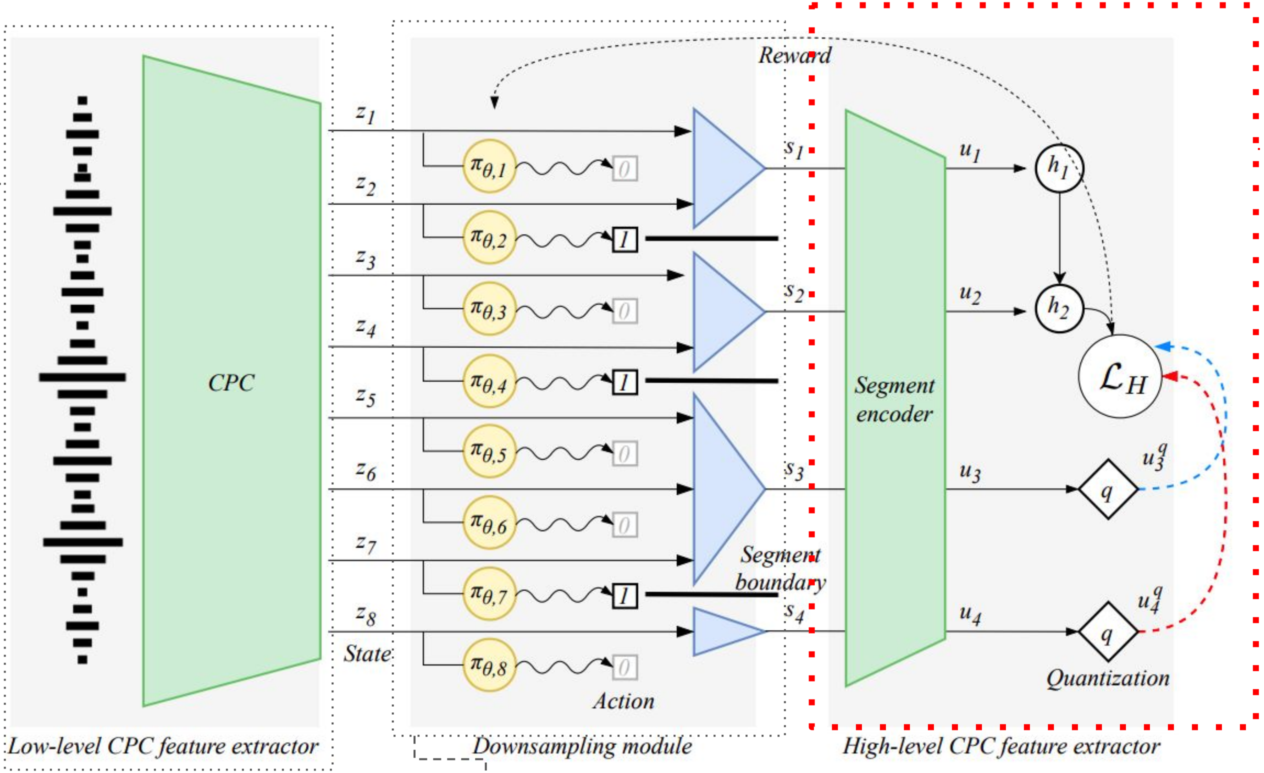
We insert a boundary predictor π that outputs a binary indicator for each element of the sequence of acoustic features. The indicator represents the presence of a high-level feature boundary.

Representations between boundaries are pooled to produce a compressed representation s which is the one we'll use for the high-level criterion.

Now we just need a suitable loss to train π so that it learns to detect high-level feature boundaries.



Our method: Variable-rate hierarchical CPC



1. A typical representation learner extracts low-level acoustic features.

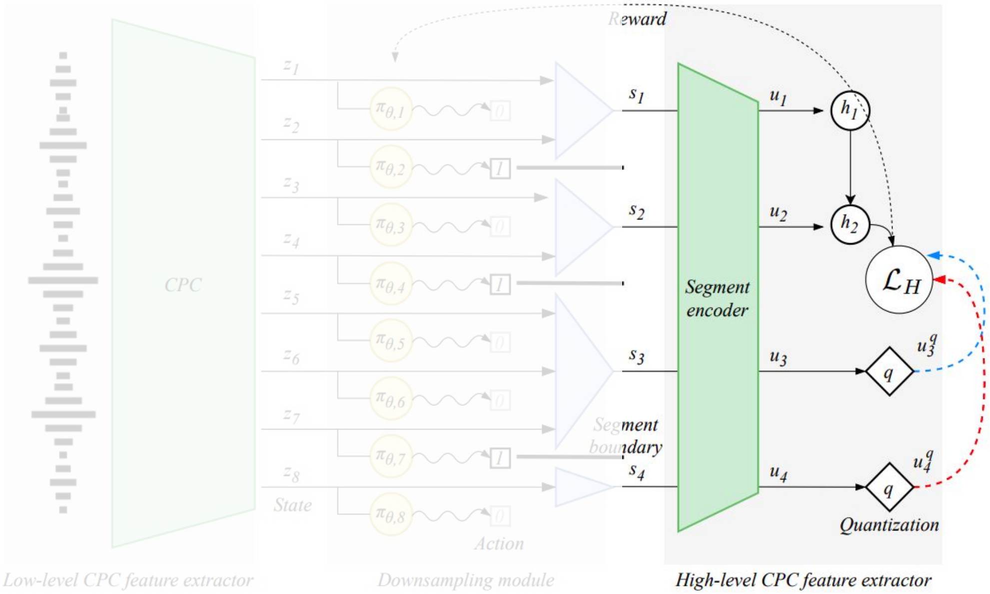
2. A learnable module downsamples the sequence of acoustic features to produce proxy high-level features with a non-uniform sampling rate

3. A high-level criterion at the level of proxy high-level features imposes the properties of discreteness and language modeling.

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

A CPC loss with a prior of discreteness

We design a CPC loss for high level modeling:



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

It encodes a prior of discreteness on the representations by enforcing contiguous elements of the sequence to be clearly distinguishable from each other.

Discreteness is also promoted by quantization of the prediction targets:

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

This loss performs representation learning while forcing the downsampling module to compress the signal so that it results in discrete-like sequences.

There is a catch: we can't get loss gradients to the boundary predictor

Our boundary predictor outputs a discrete (binary) value b, so our loss is not differentiable with respect to it.

Solution: we consider our boundary predictor π as a stochastic reinforcement learning policy and train it using policy gradient algorithms to minimize the expected value of $\mathcal{L}_H$.

We use REINFORCE to estimate the gradient as:

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

$$\nabla_\theta \mathcal{L}_\pi = \mathbb{E}_b\left[\mathcal{L}_H(b)\nabla_\theta \log(\pi_\theta(b))\right]$$

Williams (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". https://link.springer.com/article/10.1007/BF00992696

An extra perk of the stochastic policy formulation

The model of the boundary detector as a stochastic policy allows us to promote an average (proxy) high-level feature length (or sampling rate) in a differentiable way. We define an extra regularization term as:

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

Where $\bar{l}$ is the expected average proxy feature length.

Training the model

With each component being differentiable, we can train the model end-to-end. We used as loss function simply the sum of all the loss and regularization terms from each element:

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

CPC loss from acoustic feature extractor

CPC loss from proxy high-level criterion

K-means loss from the quantizer in the high-level criterion

Policy gradient loss from the downsampler

Average sampling-rate regularization term

Experimental results

Low-level representations evaluation

- We evaluate the downstream performance of acoustic 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
Multi-level	Two-level CPC no downsampling	67.49	83.38	6.66	8.34
	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 contributions:

- We have showed that top-down feedback from the right level of abstraction improves low-level representations’ disentanglement.
- We have proposed a reward function for self-supervised acoustic unit discovery.

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.
- Expanding to other modalities.

Thank you

NeurIPS 2022 paper: https://openreview.net/pdf?id=Jk8RVjnHIsE
ICASSP 2022 paper: https://ieeexplore.ieee.org/document/9746102
Code: https://github.com/chorowski-lab/hCPC