COGNITIVE CODING OF SPEECH

Reza Lotfidereshgi, Philippe Gournay

Speech and Audio Research Group, Université de Sherbrooke, Quebec, Canada

1. Objectives

- Unsupervised extraction of contextual representations in hierarchical levels of abstraction.
- Capture all sorts of attributes, including:
  - Attributes that persist less than one hundred milliseconds, such a phoneme identity.
  - Attributes that persist up to one second, such as speaker identity and emotion.
- Employ several known principles of cognition.

2. Prior work

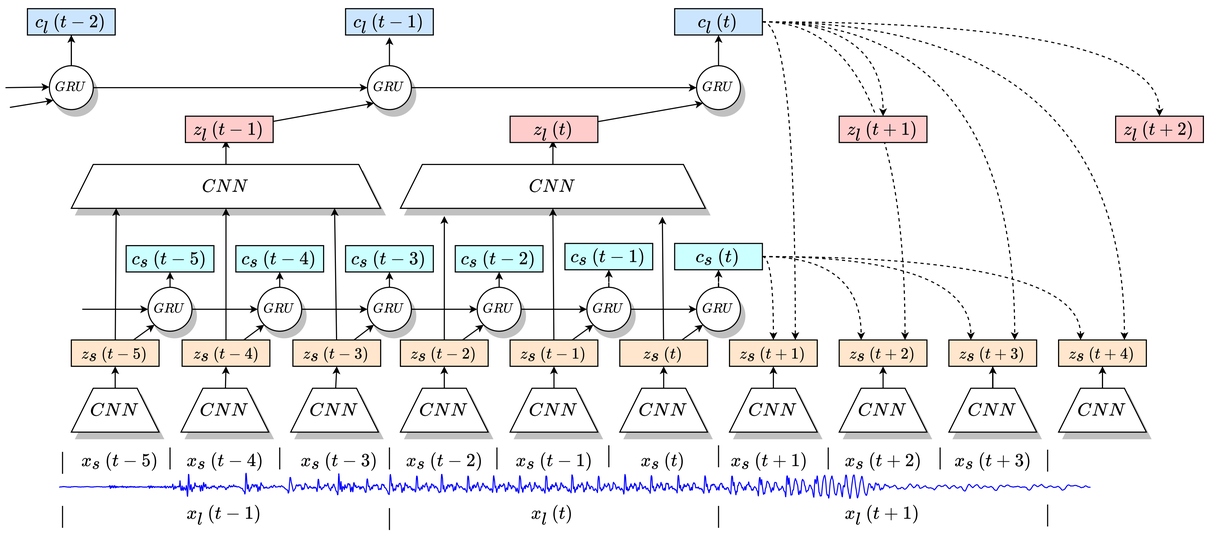
- Mutual information has been used in the formulations of many models for representation learning.
- Noise Contrastive Estimation (NCE) is a method for parameter estimation of probabilistic models by discriminating data from noise.
- In the Contrastive Predictive Coding (CPC) model, NCE is formulated as a probabilistic contrastive loss that maximizes the mutual information between the encoded representations and the input data.
  - https://arxiv.org/pdf/1807.03748.pdf
- Results in this paper are compared with the CPC model.

3. Theories of cognition

This study relies on several principles of cognition:

- A two-stage neural network model is used to extract representations in two levels of abstraction, with a lower stage and an upper stage processing information from short and long frames of data, respectively.
- A top-down pathway between stages is introduced, which has the effect of improving the quality of the representations.
- Predictive coding is used as the learning strategy.

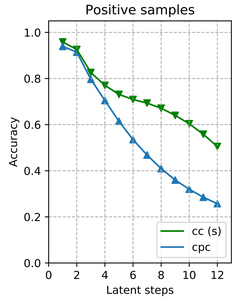
4. The architecture and learning algorithm



1. An encoder maps short frames of speech signal $x_s(t)$ to a sequence of latent variables $z_s(t)$ while decreasing the temporal resolution.
2. Another encoder maps the first sequence of latent variables $z_s(t)$ to another set of latent variables $z_l(t)$ while further decreasing the temporal resolution and increasing the receptive field to match long frames of speech signal.
3. Two autoregressive models map $z_s(t)$ and $z_l(t)$ to two sequences of contextual representations $c_s(t)$ and $c_l(t)$.
4. Approximations of density ratio captures the mutual information between a future frames of speech signal at step $t+k$ and contextual representations (prediction is done up to twelve steps in the future). The model is trained to optimize a loss based on NCE.

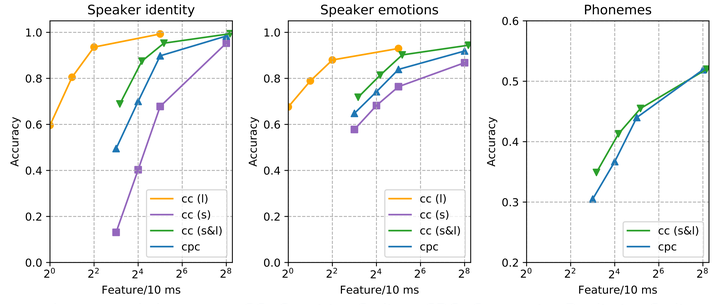
5. Positive samples

The proposed approach is able to predict positive samples of short frames more efficiently beyond 3 latent steps.



6. Linear classification of attributes

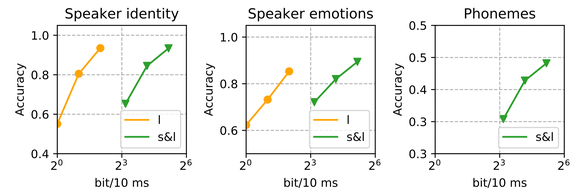
The performance of our model is measured by training linear classifiers for various speech attributes to show to what extent the extracted features are linearly interpretable.



s: short-term, l: long-term, CC: Cognitive Coding. CPC: Contrastive Predictive Coding.

7. Quantization

The performance of the proposed model is also measured by training linear classifier on quantized features. Features are quantized using 1-bit Δ-modulation.



8. Important Results

- The hierarchy of representations captures a wide variety of speech attributes over a broad range of time scales.
- Extracted hierarchical representations have multiple desirable properties including:
  - Easily interpretable.
  - Resilient to dimensionality reduction and well suited for compression.
  - Remarkably robust to quantization.

9. Outlooks

- Based on the results, this cognitive coding model could find applications in:
  - High-quality speech synthesis
  - Voice transformation
  - Speech compression
- In a follow-up study, we investigated the application of this model for speech compression in a paper with the title *Practical Cognitive Speech Compression*.
  - https://arxiv.org/pdf/2203.04415.pdf

10.Contact Information

- Web: http://www.gel.usherbrooke.ca/audio/
- Email: Reza.Lotfi.Dereshgi@USherbrooke.ca
  Philippe.Gournay@USherbrooke.ca
- Phone: +1 819 821-8000

UNIVERSITÉ DE SHERBROOKE