COGNITIVE CODING OF SPEECH

Reza Lotfidereshgi, Philippe Gournay

Speech and Audio Research Group
Université de Sherbrooke
Sherbrooke (Québec) J1K 2R1 Canada

ABSTRACT

We propose an approach for cognitive coding of speech by unsupervised extraction of contextual representations in two hierarchical levels of abstraction. Speech attributes such as phoneme identity that last one hundred milliseconds or less are captured in the lower level of abstraction, while speech attributes such as speaker identity and emotion that persist up to one second are captured in the higher level of abstraction. This decomposition is achieved by a two-stage neural network, with a lower and an upper stage operating at different time scales. Both stages are trained to predict the content of the signal in their respective latent spaces. A top-down pathway between stages further improves the predictive capability of the network. With an application in speech compression in mind, we investigate the effect of dimensionality reduction and low bitrate quantization on the extracted representations. The performance measured on the LibriSpeech and EmoV-DB datasets reaches, and for some speech attributes even exceeds, that of state-of-the-art approaches.

Index Terms— Unsupervised learning, predictive coding, representation learning, speech compression, neural networks.

1. INTRODUCTION

The human cognitive system is known to have a hierarchical organization, the most cognitively complex operations being performed at the top of the hierarchy. While information mostly flows from the bottom to the top of the hierarchy, this bottom-up flow is often influenced by what is already known at the top of the hierarchy. Furthermore, there is substantial evidence for the predictive nature of this top-down influence [1, 2]. A parallel can be drawn between these defining elements of the cognitive system and the models used in machine learning. One of the first successful applications of deep learning was precisely in the field of automatic learning of hierarchical representations [3, 4]. It was also found that introducing top-down processes in hierarchical models improves the quality of learned representations, thereby increasing the accuracy of recognition systems based on these representations [5, 6]. Predictive coding has also been shown to be a successful strategy in machine learning when processing various data modalities [6, 7].

Unsupervised learning not only reduces the need for labeled datasets, it also makes it possible to build comprehensive hierarchical representations that provide a deep insight into the nature of the input data. This is particularly important in speech compression, where efficiency depends on the completeness and compactness of the representation, which should capture all sorts of speech attributes. Yet despite the great potential of unsupervised learning, domain-specific representation learning, which can only capture a subset of the attributes from labeled data, is still prevalent in the literature. Currently, one of the very few approaches to extract comprehensive speech representations is the Vector Quantized Variational Autoencoder (VQ-VAE) [8]. Its use in recent deep learning-based speech coders and synthesizers [9, 10, 11] substantiates the need for compact and complete speech representations.

In this paper, we propose and evaluate a new approach for unsupervised learning and extraction of speech representations that heavily relies on the principles of cognition. First, 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. Secondly, a top-down pathway between stages is introduced, which has the effect of improving the quality of the representations. Finally, predictive coding is used as the learning strategy. The performance of the proposed approach is measured in terms of classification accuracy for speaker identity, emotions and phonemes. To position the results of the proposed approach with respect to the current state of the art, Contrastive Predictive Coding (CPC) [7] is used as a baseline. We observe that the second stage of the proposed model delivers a compact and remarkably high-quality long-term representation of the speech signal. The quality of the short-term representation extracted by the first stage is improved compared to that of the CPC baseline, especially when the dimension of the representation is reduced. Finally, we demonstrate that the extracted representations are extremely robust to quantization.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

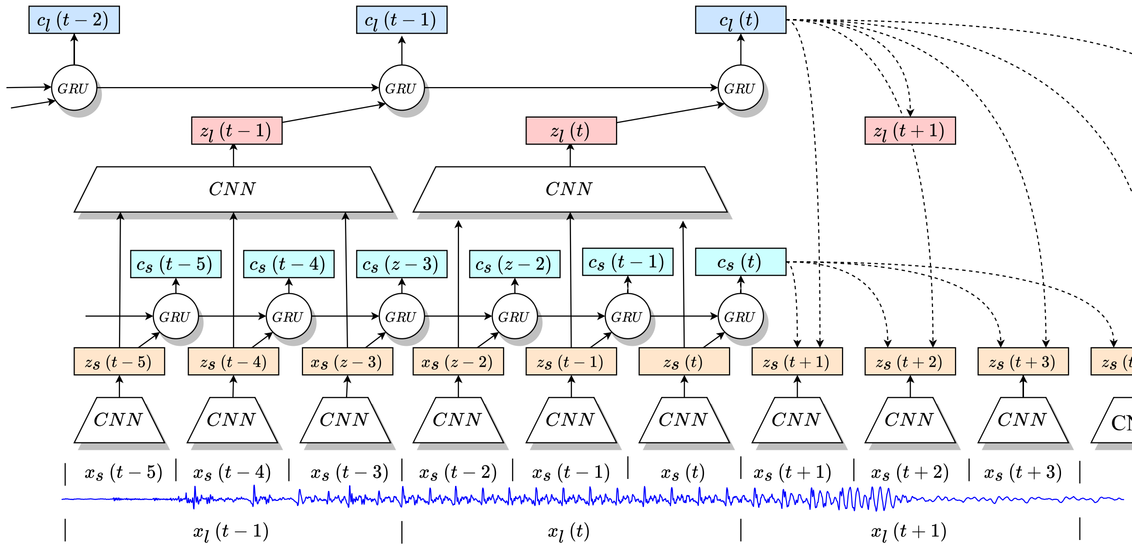


Fig. 1. The architecture and learning algorithm of the cognitive coding model. The ratio between long and short frames in the diagram is chosen to be three for illustration purposes. In this study the actual frame ratio is eight.

2. RELATION TO PRIOR WORK

The proposed Cognitive Coding model utilizes predictive coding in two stages and includes a top-down process between stages. These two stages produce two representations that evolve at a different pace and thus correspond to different levels of abstraction. The representations are extracted by maximizing the mutual information between the latent variables and the speech signal. Finally, the mutual information is maximized by minimizing a contrastive loss.

Mutual information is a fundamental quantity measuring the relationship between random variables. In previous work, it has been used in the formulations of Generative Adversarial Networks (GANs) [12] and Variational Autoencoders (VAEs) to make them learn interpretable representation of data [13, 14, 15]. Noise Contrastive Estimation (NCE) is a method for parameter estimation of probabilistic models by discriminating data from noise [16, 17]. In the model called Contrastive Predictive Coding (CPC) [7], NCE is also formulated as a probabilistic contrastive loss that maximizes the mutual information between the encoded representations and the input data.

In the CPC model, an encoder maps the input data to a sequence of latent variables, and an autoregressive model produces another sequence of latent variables. The InfoNCE loss introduced in [7] optimizes the discrimination of a positive sample from multiple negative samples. In this paper, we optimize a similar objective with consideration of two levels of abstraction and the presence of a top-down process. We also implemented the CPC algorithm as a baseline against which to compare our results.

3. COGNITIVE CODING OF SPEECH

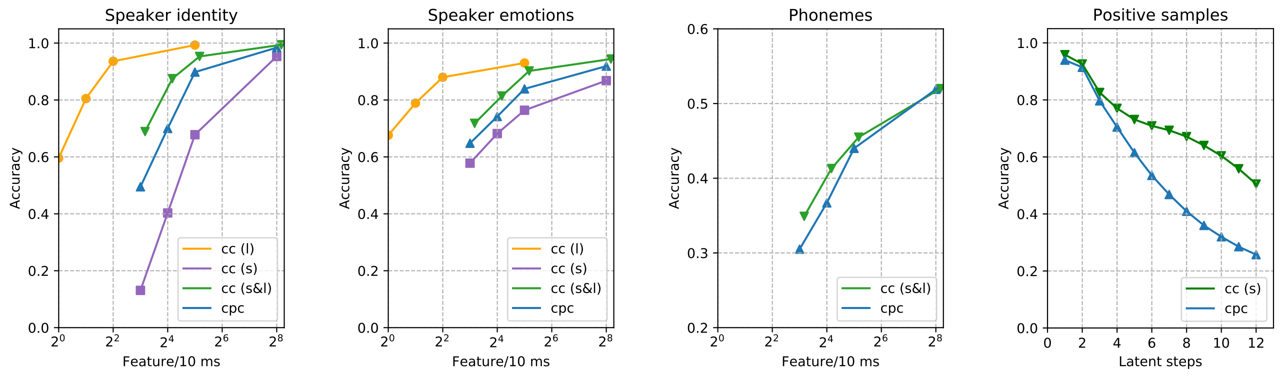
The architecture and learning algorithm of the Cognitive Coding model are illustrated in Fig. 1. The architecture can be described as follows. First, an encoder maps short frames of speech signal $x_s(t)$ to a sequence of latent variables $z_s(t)$ while decreasing the temporal resolution. Then, another 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. In this study, we use layers of Convolutional Neural Networks (CNNs) as encoders. Finally, two autoregressive models map $z_s(t)$ and $z_l(t)$ to two sequences of contextual representations $c_s(t)$ and $c_l(t)$. In this study we use Gated Recurrent Units (GRUs) for the autoregressive models.

We begin by describing the learning algorithm for the lower stage of the model. In this lower stage, the mutual information between both contextual representations and short frames of speech signal can be expressed as:

$$I(x_s; c_s, c_l) = \sum_{x_s, c_s, c_l} p(x_s, c_s, c_l) log \frac{p(x_s|c_s, c_l)}{p(x_s)} \qquad (1)$$

The following unnormalized density ratio captures the mutual information between a future short frame of speech signal at step $t+k$ and both contextual representations:

$$f_k(x_s(t+k), c_s(t), c_l(t)) \propto \frac{p(x_s(t+k)|c_s(t), c_l(t))}{p(x_s(t+k))} \qquad (2)$$



(a) (b) (c) (d)

Fig. 2. Linear classification of attributes and prediction accuracy of positive samples in the loss function. (s: short-term, l: long-term, CC: Cognitive Coding. CPC: Contrastive Predictive Coding.)

As in the CPC model, we do not use a generative model to produce future frames of speech signal. Rather, we use the following quantity to approximate f_k:

$$\exp(z_s^T(t+k)W_s(k)g(c_s(t),c_l(t))) \tag{3}$$

In equation (3), $W_s(k)$ is a linear transformation used for the prediction of $z_s(t+k)$ (k steps in the future) and $g(c_s(t),c_l(t))$ is a function of both contextual representations that constitutes the input of the linear transformation. While a neural network could be used for g to perform a nonlinear transformation, we simply repeat the long-term representation to match the temporal resolution of the short-term representation and concatenate it with the short-term representation to be used as input for the linear prediction of $z_s(t+k)$ by $W_s(k)$. This is perfectly justified because the upper stage of our model produces a long-term representation that is easily interpretable by linear classifiers (see section 4.1).

Finally, the loss function is derived according to noise contrastive estimation which is the categorical cross entropy of classifying one positive sample of short frames of speech signal from $N-1$ negative ones:

$$L_N = \mathop{\mathbb{E}}_{X_s}\left[log\frac{f_k(x_s(t+k),c_s(t),c_l(t))}{\sum_{x_s(j)\in X_s} f_k(x_s(j),c_s(t),c_l(t))}\right] \tag{4}$$

For the upper stage of the model, an equivalent of equations (1-4) can be derived based on long frames of speech signal $x_l(t)$. c_s is omitted from equations (1-2). Furthermore, since there is no top-down pathway in the upper stage, the prediction of $z_l(t+k)$ is based only on the long-term contextual representation $c_l(t)$ and the approximation for the density ratio becomes:

$$\exp(z_l^T(t+k)W_l(k),c_l(t)) \tag{5}$$

The loss function is derived by substituting equation (5) in equation (4), and samples are drawn from long frames of speech signal.

4. EXPERIMENTS

This section presents experimental results regarding various speech attributes and investigates the effects of dimensionality reduction and quantization on the quality of the representations. Two different datasets were used. First, a 100-hour subset of the LibriSpeech dataset [18] was used to evaluate the performance of the proposed approach on phonemes (a short-term attribute) and on speaker identity (a long-term attribute). We used forced-aligned phoneme labels as well as the test and train split from [7] so that we could obtain comparable results. Secondly, we used the Emov-DB dataset [19] to evaluate the performance of the proposed approach on speaker emotions which is another long-term attribute.

The encoder used in the lower stage consists of five layers of CNNs with filter sizes [10, 8, 4, 4, 4] and with strides [5, 4, 2, 2, 2]. The encoder in the upper stage consists of three layers of CNNs with filter sizes [4, 4, 4] and with strides [2, 2, 2]. Each layer has 512 hidden dimensions with ReLu activations. As a result, the lower and upper encoders downsample their input by a factor of 160 and 8, respectively. We trained on 20480-sample windows of speech signal sampled at 16kHz. As a result, the lower and upper encoders produce z_c and z_l vectors of features once every 10 ms and 80 ms, respectively. We decided that the dimension of the hidden state of GRUs would be either 8, 16, 32 or 256 so that the network can produce representations of various dimensions. Prediction is done twelve steps in the future, which extends the window of prediction up to 120 ms in the future for the lower stage and 960 ms for the upper stage. We trained with a learning rate of 2e-4, using mini batches of 8 samples, and performed approximately 300k updates.

4.1. Linear classification

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. Fig. 2 (a-c) presents the performance of linear classification for speaker identity, emotion and phonemes. Fig. 2 (d) shows the ability of the lower stage of the proposed model to predict positive samples in the loss function up to twelve steps in the future. The results are reported for classifying contextual representations extracted from long frames of signal (l), short frames of signal (s), combined contextual representations (s&l) as well as contextual representations of the CPC model. The following observations can be made based on the results.

Regarding the baseline, the results reported in [7] for the 256-dimension representation which produces 256 features every 10 ms are 97.4% and 64.6% of accuracy for speaker identity classification and phoneme classification, respectively. With our implementation of CPC, we were able to achieve a higher accuracy of 98.4% for speaker identity but a lower accuracy of 51.9% for phonemes.

Since the upper stage of our model produces a set of features for each 80 ms of speech signal, the number of features per 10 ms is 8 times less relative to the lower stage of our model and to the CPC model. For long-term attributes (speaker identity and emotion) the proposed network outperforms CPC in terms of linear classification for combined 256-dimension representations by achieving an accuracy of 99.3% and 94.4% for speaker identity and emotion, respectively. The corresponding accuracy achieved by the CPC model was 98.4% and 91.9%. By reducing the dimensionality of the representations, we observe that a high degree of linear separation between speaker identities and emotions is maintained when considering the features extracted by the upper stage of our model. Features extracted by the lower stage provide lower performance for long-term attributes. Overall this is a desirable effect that we attribute to the top-down pathway which helps decorrelate long-term attributes from short-term representations.

Regarding linear classification of phonemes based on contextual representations, we achieved 52% accuray, a lower performance than the state of the art with forced aligned features provided by [7]. Strangely enough, this is true even with our implementation of CPC baseline model. However, phoneme information is encoded in latent variable z_s which has a smaller receptive field compared to both contextual representations. Besides, not all information is linearly interpretable. In an experiment we used a classifier with one hidden layer on contextual representations and latent variables z_s and z_l and accuracy increased to 64.1%. Features of z_s are also a candidate for dimensionality reduction to encode information in a smaller time scale.

We also investigated the effect of the top-down pathway on the prediction of positive samples in the lower stage and compared the performance of our model with that of the CPC baseline in the same setup. Fig. 2 (d) shows that the proposed approach is able to predict positive samples of short frames more efficiently beyond 3 latent steps.

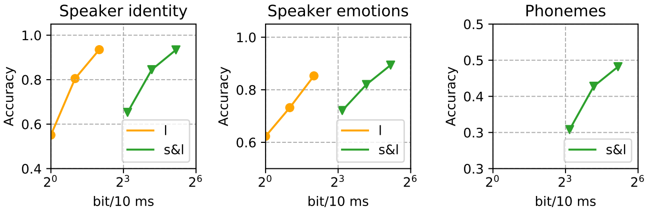


Fig. 3. Linear classification of quantized features.

4.2. Quantization

In this study, we also investigated the compressibility of the features. Since each stage predicts twelve time steps in the future, the contextual representations have a slow-evolving nature and we observe that the features exhibit a high degree of temporal dependency. For this reason, we decided we would quantize the features using 1-bit Δ-modulation. The initial values of the features are encoded on 5 bits. Fig. 3 shows the results obtained when the features are quantized for the most interesting configurations from Fig. 2. We only consider representations with 32 dimensions and less because they are the most likely to be used in speech compression applications. For the majority of the cases, the performance of the linear classification is within 5% of the corresponding performance from Fig. 2. Most notably, we observe that our model can encode long-term speech attributes such as speaker identity and emotion with more that 50% accuracy at bitrates as low as 100 bit/s.

5. CONCLUSION

In this paper, we presented a new model for cognitive coding of speech that combines several principles of cognition. Specifically: (1) it produces a hierarchy of representations that correspond to different levels of abstraction; (2) it uses the predictive coding principle; and (3) it includes a top-down pathway between levels of abstractions. The hierarchy of representations captures a wide variety of speech attributes over a broad range of time scales. Experiments show that this hierarchy is also easily interpretable, well suited for compression, and remarkably robust to quantization. This cognitive coding model should therefore find applications in high-quality speech synthesis, voice transformation and speech compression.

6. REFERENCES

[1] Jerome S Bruner, "On perceptual readiness.," *Psychological review*, vol. 64, no. 2, pp. 123, 1957.

[2] Micha Heilbron and Maria Chait, "Great expectations: is there evidence for predictive coding in auditory cortex?," *Neuroscience*, vol. 389, pp. 54–73, 2018.

[3] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[4] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[5] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5968–5977.

[6] Haiguang Wen, Kuan Han, Junxing Shi, Yizhen Zhang, Eugenio Culurciello, and Zhongming Liu, "Deep predictive coding network for object recognition," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5266–5275.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[8] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," *arXiv preprint arXiv:1711.00937*, 2017.

[9] Cristina Gârbacea, Aäron van den Oord, Yazhe Li, Felicia SC Lim, Alejandro Luebs, Oriol Vinyals, and Thomas C Walters, "Low bit-rate speech coding with vq-vae and a wavenet decoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 735–739.

[10] Jonah Casebeer, Vinjai Vale, Umut Isik, Jean-Marc Valin, Ritwik Giri, and Arvindh Krishnaswamy, "Enhancing into the codec: Noise robust speech coding with vector-quantized autoencoders," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 711–715.

[11] Xin Wang, Shinji Takaki, Junichi Yamagishi, Simon King, and Keiichi Tokuda, "A vector quantized variational autoencoder (vq-vae) autoregressive neural f_0 model for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 157–170, 2019.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[13] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 2180–2188.

[14] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[15] Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[16] Michael Gutmann and Aapo Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[17] Aapo Hyvarinen and Hiroshi Morioka, "Unsupervised feature extraction by time-contrastive learning and nonlinear ica," *Advances in Neural Information Processing Systems*, vol. 29, pp. 3765–3773, 2016.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[19] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.