INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Predictive Auxiliary Variational Autoencoder for Representation Learning of Global Speech Characteristics

Sebastian Springenberg[1], *Egor Lakomkin*[1], *Cornelius Weber*[1], *Stefan Wermter*[1]

[1] University of Hamburg - Dept. of Informatics, Knowledge Technology

5springe@informatik.uni-hamburg.de, lakomkin@informatik.uni-hamburg.de,
weber@informatik.uni-hamburg.de, wermter@informatik.uni-hamburg.de

Abstract

Unsupervised learning represents an important opportunity for obtaining useful speech representations. Recently, variational autoencoders (VAEs) have been shown to extract useful representations in an unsupervised manner. These models are usually not designed to explicitly disentangle specific sources of information. When processing data of sequential nature which involves multi-timescale information, disentanglement can however be beneficial. In this paper we address this issue by developing a predictive auxiliary variational autoencoder to obtain speech representations at different timescales. We will present an auxiliary lower bound which is used to develop a model that we call the Predictive Aux-VAE. The model is designed to disentangle global from local information into a dedicated auxiliary variable. Learned representations are analysed with respect to their ability to capture global speech characteristics. We observe that representations of individual speakers are separated well in the latent space and can successfully be used in a subsequent speaker identification task where they achieve high classification accuracy, comparable to a fully supervised model. Moreover, manipulating the global variable allows to change global characteristics while retaining the local content during generation which demonstrates the success of our model to disentangle global from local information.

Index Terms: unsupervised learning, speech representation learning, predictive coding, predictive auxiliary VAE, speaker identification

1. Introduction

With the rapidly growing amount of unlabeled data publicly available, unsupervised feature extractors are becoming increasingly appealing as they can help to accelerate various supervised learning scenarios without the need for additional annotated data [1]. A lot of work on variational auto-encoders has lately analysed the processing static data such as images [2, 3]. Much less work has however been done on data of sequential nature such as video, text and audio [4, 5].

Sequences usually contain information at different timescales. When e.g. processing speech audio, overall speech characteristics can be seen as a source of information independent from the content being spoken. Recent work has been concerned with capturing the multi-scaled nature of sequences by introducing additional latent variables [6, 7]. While these models have been shown to be successful in extracting representations at different levels of hierarchy, they involve computationally inefficient caching methods during training [6] or relinquish generative capabilities [7].

This work focuses on applying a predictive auxiliary VAE architecture to speech audio. The model is designed to disentangle global speech characteristics from local information into a dedicated auxiliary variable. It builds upon an extension of the standard evidence lower bound (ELBO) inspired by work on variational inference with auxiliary variables [8, 9]. We will investigate learned representations with respect to their separability in the latent space and their transferability to a subsequent speaker identification task.

2. Related Work

Hsu et al. develop an architecture that aims to disentangle local from global information inherently in the model architecture [6]. They introduce a factorised hierarchical variational autoencoder (FHVAE) that extracts information at different timescales. Two latents are produced by the encoder, one operating on the segment level (local variable) and another operating on the sequence level (global variable). Their model makes use of two RNNs to first encode the global variable z_2 which is then used as a condition for encoding the local variable z_1. The decoder is chosen to be a RNN as well and processes a concatenation of both the local and the global variable to produce a reconstruction. While the local prior distribution $p_\theta(z_1)$ is assumed to be a sequence-independent isotropic Gaussian centered at zero, the global prior distribution $p_\theta(z_2|\mu_2)$ is an isotropic Gaussian centered at μ_2. This results in the global prior being formulated as $p_\theta(z_2|\mu_2) = \mathcal{N}(z_2|\mu_2, \sigma^2_{z_2} I)$, where I denotes the identity matrix. During generation, μ_2, also called the s-vector, first has to be drawn before z_2 and subsequently z_1 can be inferred. Here, a specific s-vector is assigned to each sequence in the training set. The posterior mean $p(\mu_2|X)$ for each training sequence is thus stored in a lookup table and not directly inferred from the input. While this procedure leverages problems associated with extracting a global s-vector from extensively long input sequences, it poses problems related to the cache memory required. In fact, maintaining a cache with entries for every training example renders the model as not scalable to large datasets. Even though the authors address this problem in a follow-up paper [10] by introducing a hierarchical sampling method, keeping a cache and reinitialising it when processing new batches of data appears not to be computationally efficient.

Van den Oord et al. develop a contrastive predictive coding (CPC) architecture which relinquishes the decoder completely and performs predictive coding within the latent space [7]. Here, the input sequence x_t is first encoded into a sequence of latent representations z_t. Subsequently, an autoregressive model, e.g. a RNN, summarises all $z_{<t}$ into another latent representation c_t. This vector can be interpreted as a context vector and is used to predict future latents $z_{>t}$. The authors propose to produce multiple predictions into the future at every step, encouraging the latent representation to capture slow features [11] at different levels of abstraction. Predicting further

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2845

into the future than the next step prevents the model to exploit local smoothness of the signal. A common problem when predicting high-dimensional sequential data such as audio is the need for powerful auto-regressive models that usually exploit complex relationships in the data and ignore low-dimensional global context vectors. In order to circumvent this issue, Van den Oord et al. encode the original input and context into compact latent vectors in a way that maximally preserves the mutual information of both. As the model predicts future latents z_{t+k} instead of future observations x_{t+k}, the authors model a density ratio f to ensure the mutual information between x_{t+k} and c_t. The CPC model can be trained to estimate the density ratio by optimising a Noise Contrastive Estimation (NCE) loss. In NCE, density estimation is simplified by applying a binary classifier to distinguish samples of the model distribution from samples generated by a noise distribution. The authors report good results when applying the model to different domains including images, video, reinforcement learning and audio. Relinquishing a decoder however discards generative capabilities.

Hjelm et al. recently proposed a method called Deep InfoMax (DIM) that follows a learning objective similiar to the one presented in this work with the difference of relying on adverserial training to achieve global and local mutual information maximisation as well as prior matching [12].

3. Proposed Model

3.1. Auxiliary ELBO

We extend the standard ELBO commonly used when training VAEs by an auxiliary variable h. The ELBO can be expressed in terms of the likelihood, the prior and the entropy H of the approximate posterior $q(z|x)$:

$$\begin{aligned}&\log p(x)\\&\geq \mathbb{E}_{q(z_t|x_t)}\Big[\log \frac{p(x_t|z_t)p(z_t)}{q(z_t|x_t)}\Big]\\&= \mathbb{E}_{q(z_t|x_t)}\Big[\log p(x_t|z_t) + \log p(z_t)\Big] + \mathrm{H}\Big[q(z_t|x_t)\Big].\end{aligned} \tag{1}$$

It can now be assumed that there is another source of information (e.g. from a global time-scale) provided via a second latent variable $h_t \sim \mathcal{N}(\mu_{h_t}|\sigma^2_{h_t}I)$, where $[\mu_{h_t}, \sigma_{h_t}] = g(x_{t-W:t})$, with W being the window size and g being an arbitrary parameterised function (e.g. a neural network) that sums over all time in the window: $g_\theta(x_{t-W:t}) = \sum_{i=t-W}^{t} \mathrm{f}_\theta(x_t)$, where f is the application of a neural network.

The additional variable is desired to be used in an expressive manner by ensuring that as much information as possible is taken from the global time-scale, while the local time-scale variable facilitates accurate reconstruction. This is achieved by following a derivation similar to existing work on variational inference with auxiliary variables [8, 9].

The auxiliary global variable h_t can be introduced by rewriting the entropy as:

$$\begin{aligned}&\mathrm{H}\Big[q(z_t|x_t)\Big]\\&= -\int q(z_t|x_t)\log q(z_t|x_t)dz_t\\&= \int q(z_t|x_t)\log\frac{1}{q(z_t|x_t)}dz_t\\&= \int q(z_t|x_t)\log\Big[\int_{p(h_t)} p(h_t|x_{W:t})\frac{1}{q(z_t|x_t)}dh_t\Big]dz_t,\end{aligned} \tag{2}$$

By inferring that $\frac{1}{p(x)} = \frac{p(z|x)}{p(x,z)}$ (due to Bayes rule), the following can be deduced:

$$\begin{aligned}&\mathrm{H}\Big[q(z_t|x_t)\Big]\\&= \int q(z_t|x_t)\log\Big[\int p(h_t|x_{W:t})\frac{p(h_t|z_t,x_{w:t})}{q(z_t,h_t|x_t)}dh_t\Big]dz_t.\end{aligned} \tag{3}$$

The posterior entropy can now be lower bounded by applying Jensen's inequality:

$$\begin{aligned}&\mathrm{H}\Big[q(z_t|x_t)\Big]\\&\geq \int q(z_t|x_t)\Big[\int p(h_t|x_{W:t})\log\frac{p(h_t|z_t,x_{w:t})}{q(z_t,h_t|x_{w:t})}dh_t\Big]dz_t\\&= \mathbb{E}_{q(z_t,h_t|x_{W:t})}\Big[\log p(h_t|z_t,x_{w:t}) - \log q(z_t,h_t|x_{w:t})\Big].\end{aligned} \tag{4}$$

Assuming further that $q(z_t, h_t|x_{w:t})$ factorizes as $q(z_t, h_t|x_{w:t}) = q(z_t|x_t, h_t)p(h_t|x_{w:t})$ leads to the following expression:

$$\begin{aligned}&\mathrm{H}\Big[q(z_t|x_t,h_t)\Big]\\&\geq \mathbb{E}_{q(z_t,h_t|x_{W:t})}\Big[\log p(h_t|x_{w:t}) - \log q(z_t,h_t|x_{w:t})\Big]\\&= \mathbb{E}_{q(z_t,h_t|x_{W:t})}\Big[\log p(h_t|z_t,x_{w:t})\Big]\\&\quad + \mathbb{E}_{p(h_t|x_{W:t})}\Big[H\Big[q(z_t|x_t,h_t)\Big]\Big] + H\Big[p(h_t|x_{w:t})\Big],\end{aligned} \tag{5}$$

where $\mathbb{E}_{q(z_t,h_t|x_{W:t})}\Big[\log p(h_t|z_t,x_{w:t})\Big] = \mathbb{E}_{q(z_t|x_t)}\Big[-\mathrm{CE}\Big[p(h_t|x_{W:t})||p(h_t|z_t,x_{w:t})\Big]\Big]$ corresponds to a cross-entropy term, encouraging h_t to be predictable from z_t.

By combining the lower bound in equation 5 with the one in equation 1 and introducing hyperparameters α and β, the following learning objective can be obtained:

$$\begin{aligned}&L(x_{w:t})\\&= \mathbb{E}_{q(z_t|x_{W:t})}\Big[\log p(x_t|z_t)\Big]\\&\quad + \beta\Big(\mathbb{E}_{q(h_t|x_{W:t})}\Big[\log p(z_t) + H\Big[q(z_t|x_t,h_t)\Big]\Big]\Big)\\&\quad + \alpha\Big(\mathbb{E}_{q(z_t|x_{W:t})}\Big[\log p(h_t|z_t,x_{w:t})\Big] + H\Big[p(h_t|x_{w:t})\Big]\Big),\end{aligned} \tag{6}$$

where, like in the $\beta-VAE$ objective [3], β acts as as a regularisation term on the KL divergence between the local-timescale variable and the prior and α is a multiplier for the mutual information between the local-timescale and the global-timescale variable.

3.2. Predictive Aux-VAE

The multi-timescale Aux-VAE shown in Figure 1 builds upon the variational lower bound derived in the previous section. By introducing an auxiliary variable h, the architecture does now consist of four probability distributions that can all be parameterised by neural networks:

1. $p(h|x)$, the global-timescale network
2. $p(z_t|x_t, h)$, the local-timescale network that receives the global-timescale variable as additional input
3. $p(h|z_t, x_{w:t})$, the predictor network predicting the global-timescale variable from the local-timescale variable

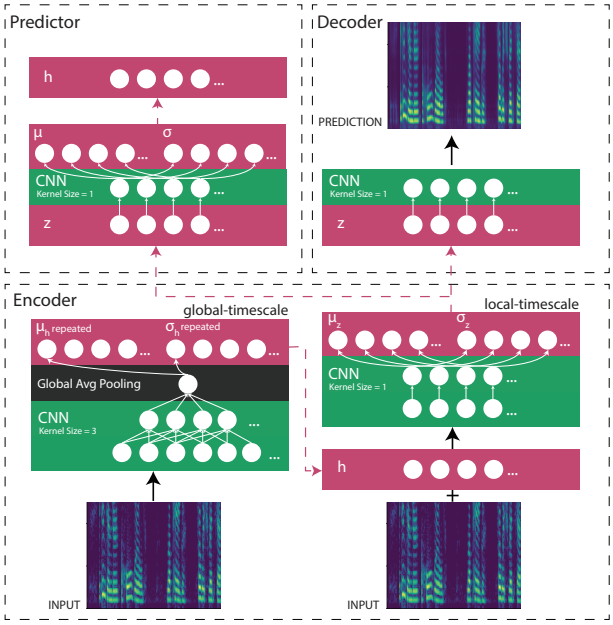

Figure 1: *Aux-VAE architecture; A global timescale auxiliary variable h is first obtained by the encoder and then used together with the input to obtain a local timescale variable z. z is used as input to the decoder for speech frame predictions as well as to the predictor which contributes to the mutual information of z and h*

4. $p(x_{t+1}|z_t)$, the decoder network

We define the global-timescale network, representing $p(h|x)$, to be a CNN with 3 consecutive layers of convolution with a kernel size of $k = 3$, followed by a convolution with a kernel size of $k = 1$. This allows us to extract information from multiple timesteps $t_{t-w:t}$ with a window size of w. Subsequently, global average pooling is performed in order to obtain a global latent variable h that encodes information from the whole speech sequence.

After having acquired a global variable h, a local variable z_t is computed by the local-timescale CNN, representing $p(z_t|x_t, h_t)$, which involves 3 layers of convolution with kernel size $k = 1$. Input is provided by a concatenation of the speech input and the global variable. Here, the global variable is appended to every speech frame. In order to make information from the global variable persist in the local variable, a third, predictor network, representing $p(h_t|z_t, x_{w:t})$, is used. We apply information hiding and predict the global variable from the local variable only to obtain a more expressive global encoding. The predictor network is a 2-layer CNN that performs convolution with a kernel size of $k = 1$.

The decoder network processes the local variable z_t to predict the next speech frame at timestep $t + 1$. It consists of four convolution layers with a kernel size of $k = 1$.[1]

3.3. Training

The model is trained using the Adam optimiser [13] with a learning rate of $l = 0.001$. Training is stopped after a total amount of 60 epochs as it is found to converge after around 50 epochs. We use the same LibriSpeech [14] test and train set as Van den Oord et al. [7]. Preprocessing involves downsampling the speech audio to a sample rate of 11khz and then performing a short-time Fourier transform. Resulting amplitude spectrograms are transformed to the decibel range and then normalised to the range $[0, 1]$.

[1] An implementation of the model can be found at: https://github.com/sspringenberg/Speech-Aux-VAE.

4. Experiments and Results

4.1. Latent Space Visualisation

We use t-SNE [15] to visualise the latent representations and assess how well individual speakers cluster are separated from each other in the latent space. Here, representations of all examples in the test set that contain speech from 10 selected speakers are obtained with the Aux-VAE architectures and then projected onto a two-dimensional space using t-SNE. True speaker ID labels are indicated by different colours. The global timescale representation h is expected to capture information about the whole speech sequence. t-SNE visualisations reveal that h indeed appears to capture global speech characteristics as speaker clusters are formed (see figure 2). We further average the local timescale representation z over time and create t-SNE plots of the results. As expected, speakers also cluster in z because the neural architecture and objective function are designed in a way that information from h is also present in z. Information about global structures is thus expected to propagate from the global to the local variable, which we will also observe when modifying h during generation. Resulting t-SNE plots of the global variable can be found in figure 2.

4.2. Global Latent Manipulation for Speaker Transformation

In order to evaluate whether information from the global timescale variable h is propagated to the local timescale variable z and then picked up by the decoder, we manipulate the global code and observe changes in speech frame predictions by examining the resulting spectrogram. A straightforward method to validate that information from h is used and not completely ignored in the generative process is adding Gaussian noise drawn from a standard normal distribution. Figure 3 shows that when noise is added, we can indeed observe that predictions are affected. While local structures are retained in the spectrogram, global structures are altered.

Speaker transformation is performed by interchanging the global representations h obtained for a male and a female speaker. Figure 4 shows the resulting transformation. We can observe that the fundamental frequency is raised and harmonic content is added without changing the content being spoken.

4.3. Speaker Identification

Applying representations learned in an unsupervised manner to a downstream task is a common technique for analysing their expressiveness. As in this work we are generally interested in extracting global speech characteristics, speaker identification represents a suitable application to investigate whether representations capture relevant features and are transferable to a classification task. Latent representations are first obtained from the encoders of the VAE architectures and then serve as input to a linear classifier. Using a simple linear classifier explicitly evaluates linear separability. We perform speaker identification on the same LibriSpeech test and train set as Van den Oord et al. [7] which allows us to compare results. Resulting speaker

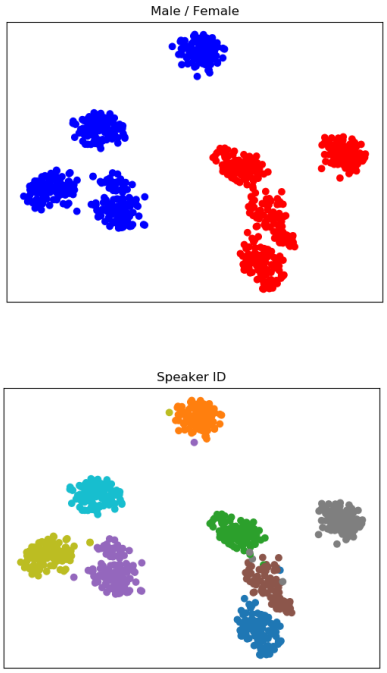

Figure 2: *t-SNE visualisations of global latent representations h obtained for 10 speakers by the multi-timescale Aux-VAE with respect to speaker ID and gender*

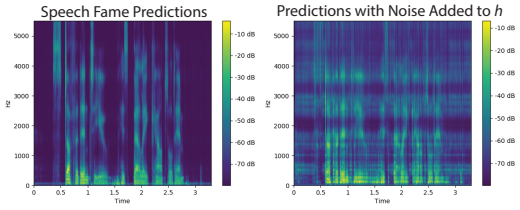

Figure 3: *Effect on the generative process in Aux-VAE when adding Gaussian noise to the global variable h*

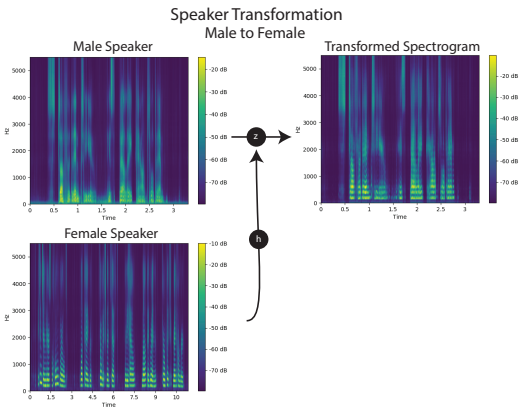

Figure 4: *Multi-timescale VAE transformation from male to female speaker by interchanging the global representation h.*

identification accuracy is given in table 1.

Classification accuracy on the basis of representations obtained by the Predictive Aux-VAE is very good and comparable to the results obtained with the CPC model by Van den Oord et al. [7]. Similar to the CPC model, accuracy is only marginally worse than with a model that is trained in a fully supervised manner, which underlines linear separability and expressiveness of the learned representations. Even though classification based on the global variable h obtained by the Aux-VAE is good, it seems that additional information present in z can further help to increase accuracy. Using the time-averaged local representation z results in a slightly better accuracy than reported for the CPC model.

Table 1: *LibriSpeech speaker identification accuracy when using learned representations; h denotes the use of the global and z denotes the use of the local variable, d denotes feature depth*

Model	Accuracy
Random initialization [7]	1.87
MFCC features [7]	17.4
Predictive Aux-VAE; h, $d = 128$	92.1
Predictive Aux-VAE; h, $d = 256$	95.3
Predictive Aux-VAE; z, $d = 256$	98.1
CPC [7]	97.4
Supervised [7]	98.5

5. Conclusion

In this work we have developed a Predictive Aux-VAE model that introduces an auxiliary variable to encode global structures into a dedicated representation. In order to develop this architecture we have presented an auxiliary lower bound inspired by already existing work on auxiliary variational inference [8, 9]. This lower bound can be of interest for different tasks and purposes. When applied to speech, the Predictive Aux-VAE is able to obtain expressive representations that can be used in a speaker identification scenario where they achieve classification accuracy comparable to the recently proposed CPC model by Van den Oord et al. [7] and perform only marginally worse than a fully supervised model. Different to the CPC model, the Predictive-Aux-VAE involves a decoder and retains generative capabilities. While this is might not be necessary for learning representations, it allows us to investigate the global variable's influence on the local variable and the generative process. Global latent manipulation during generation changes global structures but retains local information about the content being spoken. This supports our finding that global information is successfully disentangled from local information into the auxiliary variable.

6. Acknowledgements

The authors gratefully acknowledge support from the German Research Foundation (DFG) under project CML (TRR 169) and the EU under project SECURE (No 642667).

7. References

[1] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–36.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[3] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[4] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *arXiv preprint arXiv:1704.04222*, 2017.

[5] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016.

[6] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in Neural Information Processing Systems*, 2017, pp. 1878–1889.

[7] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[8] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Auxiliary deep generative models," *arXiv preprint arXiv:1602.05473*, 2016.

[9] K. Hausman, J. T. Springenberg, Z. Wang, N. Heess, and M. Riedmiller, "Learning an embedding space for transferable robot skills," *International Conference on Learning Representations*, 2018.

[10] W.-N. Hsu and J. Glass, "Scalable factorized hierarchical variational autoencoder training," *arXiv preprint arXiv:1804.03201*, 2018.

[11] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Computation*, vol. 14, no. 4, pp. 715–770, 2002.

[12] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations (ICLR)*, 2019.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[15] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.