# Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data

**Report**

**Stefan Wezel**
stefan.wezel@student.uni-tuebingen.de
4080589
ML4S

## Abstract

Information in sequential data is often distributed over multiple time scales. While if viewed as a single signal, such data might appear noisy. However, patterns can emerge if temporal scales are inspected separately from one another. Hsu et al. [1] leverage this intrinsic structure to learn disentangled representations from sequential data in an unsupervised manner with a proposed factorized hierarchical variational autoencoder (FHVAE). They aim to factorize sequence level and segment level attributes into distinct latent subspaces. Architectural and sequence-dependent priors create an inductive bias to encourage the proposed factorization. Here, we put their work into a formal context, explore the proposed methodology, and reflect critically on their work [1].

## Introduction

Intuitively, disentangled representations are representations that are reflective of the underlying generating factors of observed data in that they are encoded as separate latent subspaces (See Figure 1). This notion is already present in classical factor analysis work, where it is referred to as independent component analysis (ICA) [2].
However, many problems cannot be solved in a linear fashion. The success of deep neural networks (DNN) can be largely attributed to the fact that they are very powerful non-linear function approximators. Thus, making them a promising method to solve the long-standing problem of non-linear ICA. Different methods have been proposed to learn such disentangled representations [3, 4, 5] with varying success. Many of these works focus on image data. However, it has been shown by Locatello et al. [6] that disentangled representations cannot be learned without introducing any kind of supervision or inductive biases. Sequential data, while having been explored less, offers an inherent structure that can be exploited to construct inductive biases as has been proposed by Hsu et al. [1].
Besides technical challenges, this strain of research suffers from the lack of formally defined and agreed upon foundations. The very term of disentangled representations for example is often understood differently in between works. In the following section, we will use the definition proposed by Higgins et al. [7] to put the work by Hsu et al. [1] into formal context.

## Viewing the FHVAE though a Formal Lense

Works on disentanglement often lack a proper formal framework [7]. While claiming to achieve disentanglement, Hsu et al. [1] do not provide any formal foundation for their claim. With no formal

[1] An implementation of Hsu et al. [1]'s method and our experiments can be found at `www.github.com/wastedsummer/SequentialVAE`

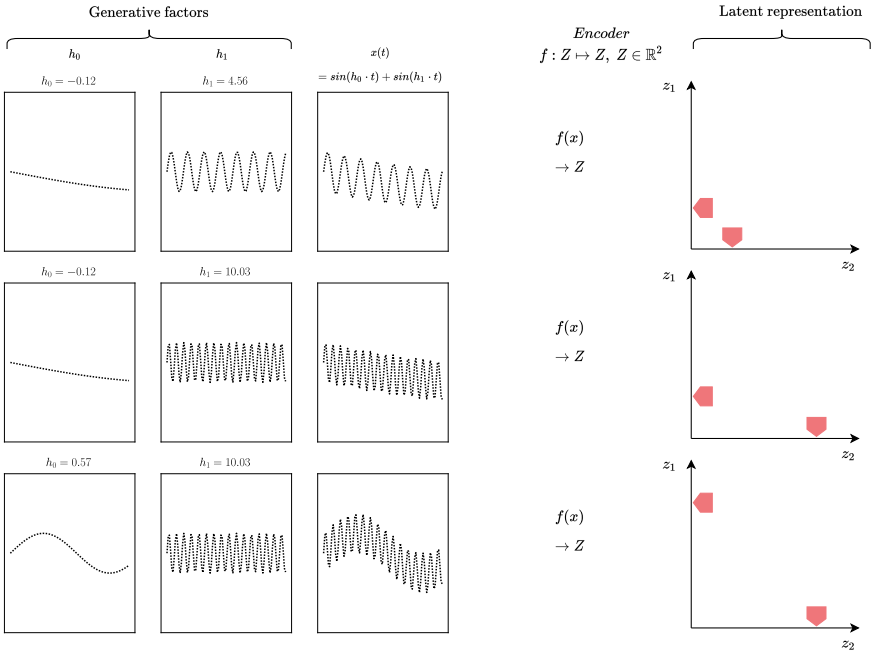


Figure 1: Changing values of generating factors are reflected in the corresponding latent variables.

tools at hand, we cannot formally discuss whether disentanglement was achieved. Thus, we use the following sections to give a theoretical background that defines disentanglement in group-theoretic terms, and describe the problem, the FHVAE was designed to solve, using the established formal framework.

### The Tools of Group Theory

A group is described by a tuple of an operation $\circ$ and a non-empty set $G$. The set has to be closed under the operation, it must contain an identity element, the operation must be associative, and for every element in $G$, there must be an inverse element [8]. Multiple groups $G_i$ can be combined via a direct product $G = G_1 \times ... \times G_n$. The result of this direct product is itself a group.
Groups that are of particular interest for the field of disentangled representation learning are symmetry groups [8]. A symmetry group consists of a set of transformations that leave a given object $X$ invariant. Its operation is the composition of such transformations. A prominent example of a symmetry group is $SE_3$. This symmetry group can be visualized as a set set of vertices $X$ that form an equilateral triangle. Permutations of this set would result in rotations or flips of the triangle. These permutations would be the symmetry transformations of our symmetry group. The operation would be composing multiple permutations.
Another important concept is the group action. It is the result of applying a symmetry transformation to an object. In our triangle example, a group action would be the permuted set of vertices.
If a group action on $X$ is the result of a subset $G_i$ of symmetries $G = G_1 \times ... \times G_n$ and only affects a subset $X_i$ of $X$ but leaves all other $X_{j \neq i}$ unchanged, we say it is a disentangled group action. Such disentangled group actions are of particular interest for disentangled representation learning, and generally are what we want to model. Note, that if we observe such disentangled group actions, we can infer that $G$ can be decomposed into a direct product of symmetry groups $G_i$ without knowing the underlying processes.
To model such a process, we need to find some symmetry preserving mapping $f : X \mapsto Z$. It should not matter, whether $X$ is mapped to $Z$ followed by applying a symmetry $G$ or $G$ is applied to $X$ and

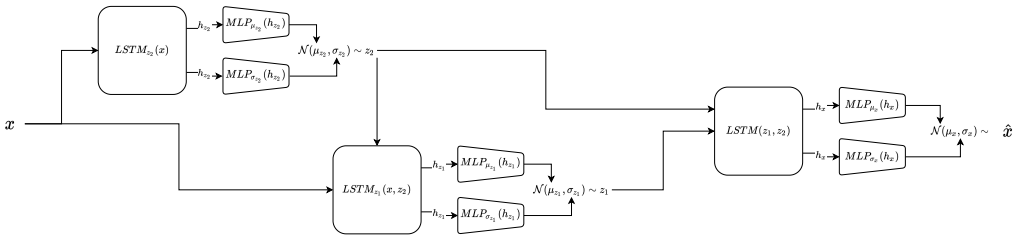


Figure 2: The architecture of the proposed FHVAE. The latent segment variable $z_2$ is conditioned on latent sequence variable $z_1$. Together, they are decoded to parameterize a normal distribution over a reconstructed input $\hat{x}$.

then mapped to $Z$. The resulting space $Z$ should be the same. This is visualized below.

$$\begin{array}{ccc} X & \xrightarrow{G} & X \\ f \downarrow & & f \downarrow \\ Z & \xrightarrow{G} & Z \end{array}$$

Such a mapping is called an equivariant map. Its result is a disentangled representation. The space $Z$ (which we will also refer to as latent space in the following) naturally decomposes into a direct product of independent subspaces $Z_1 \times ... \times Z_n$. Moreover, each subspace $Z_i$ is only affected by symmetry $G_i$ on $X$ (as only $X_i$ is changed) and remains invariant to all other $G_{j \neq i}$.
Note that it is only disentangled with respect to a certain decomposition. This is important, as it means we can only discuss whether disentanglement was achieved if the decomposition is sufficiently clear. Not all decompositions make sense or are possible to model.
As these concepts are rather abstract, we will use the next section to frame the real-world setting of Hsu et al. [1] using group-theoretic terms to build further intuition and a formal foundation to discuss their work.

#### Symmetries in Sequential Data

Hsu et al. [1] argue that certain sequential data can be factorized into attributes of different temporal scales. For example, voice recordings can be decomposed into sequence attributes and segment attributes. In this context, sequence attributes are features that remain unchanged over multiple sequences if spoken by the same speaker. Segment attributes, on the other hand, vary in sequences and are independent of the speaker. They are determined by variables such as linguistic content.
In group-theoretic terms, we could state this as the group $G$ acting on audio recordings $X$ is a direct product of $G_{sequence} \times G_{segment}$, because we observe the resulting disentangled group actions of this decomposition. Hsu et al. [1] now want to find a representation that is disentangled with respect to this decomposition. They need to find an equivariant map $f : X \mapsto Z$, where $Z = (z_1, z_2)$ is a latent space, so that i.e. $z_2$ is only affected by $G_{sequence}$ and $z_1$ only is affected by $G_{segment}$. Then, the proposed decomposition of symmetries is reflected in $Z$.
For this specific setting, finding such an equivariant map would allow separating speaker information from content information. This in turn would enable us to reconstruct given content information using another speaker's voice information. Hsu et al. [1] propose this as a qualitative assessment of their disentanglement. Quantitative evaluation of disentanglement is notoriously challenging and various metrics have been proposed by an active field of research [6, 3]. As we will see later, Hsu et al. [1] propose a speaker verification task to quantitatively evaluate the separation of sequence and segment attributes in the latent space.

## FHVAE - Constructing an Equivariant Map

To find an equivariant map for the sequential data setting, Hsu et al. [1] propose the FHVAE (see Figure 2). Its encoder maps to a latent subspace factorized into a latent sequence variable $z_1$ and a latent segment variable $z_2$. To form Gaussian distributions over these latent variable, Hsu et al. [1]

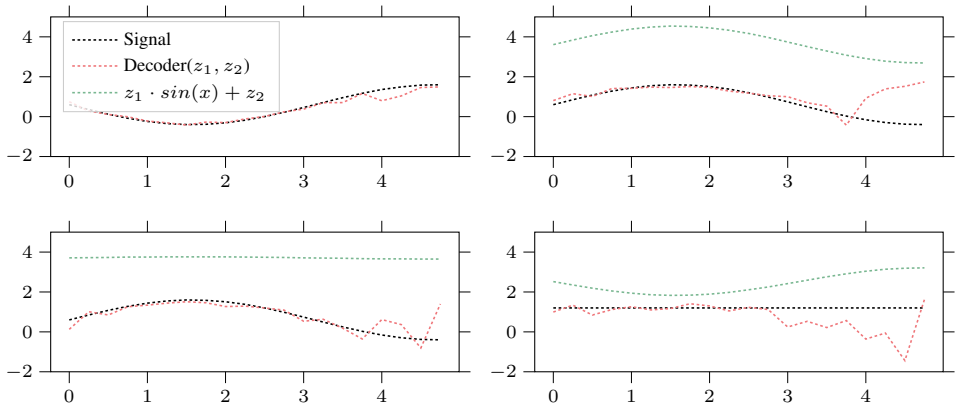


Figure 3: Sample signals reconstructed with the FHVAE. Blue lines are generated using the latent variables $z_1$ and $z_2$ with the known generative process used to create the training data. For these samples, the latent variables do not correspond to the actual generative factors values.

pass an input sequence $x$ to two distinct Long short-term Memory (LSTM) [9] cells $LSTM_{z_1}$ and $LSTM_{z_2}$. The hidden states of those cells are decoded by Multi-layer Perceptrons (MLP) to give the mean and standard deviation for each latent variable. To decode the latent space, a sample is drawn and passed to another LSTM. This LSTMs hidden state is decoded to form a normal distribution over a reconstructed input $\hat{x}$.
To train the FHVAE, Hsu et al. [1] propose a segment variational lower bound, based on Kingma and Welling [10] and add a discriminative objective. This segment variational lower bound can be written as

$$\begin{aligned}\mathcal{L}(\theta, \phi, X) = & \underbrace{log\ p(x|z_1, z_2)}_{\text{reconstruction}} \\ & - \underbrace{D_{KL}(\mathcal{N}(\mu_{z_1}, \sigma_{z_1})||\mathcal{N}(0, 1))}_{\text{regularize } z_1 \text{ with global prior}} \\ & - \underbrace{D_{KL}(\mathcal{N}(\mu_{z_2}, \sigma_{z_2})||\mathcal{N}(\mu_2, 0.5))}_{\text{regularize } z_2 \text{ with seq. dep. prior } \mu_2} \\ & + \underbrace{log\ p(\mu_2) \cdot \frac{1}{seq.\ length}}_{\text{scaled prob. of } \mu_2 \text{ under standard Gaussian prior}} ,\end{aligned}$$

where $\theta$ is the set of decoder parameters, $\phi$ are encoder parameters, and $X$ are training samples. Note that while $z_1$ gets regularized through a standard Gaussian prior, $z_2$ is discouraged to deviate from a Gaussian distribution centered at $\mu_2$. Hsu et al. [1] introduce $\mu_2^{(i)}$ as sequence $(i)$-dependent prior to encourage factorization. This sequence-dependent prior can be retrieved through a differentiable lookup table. The discriminative objective

$$log\ p(sequence\ id^{(i)}|z_2^{(i,n)}) = log\ \frac{p(\mu_{z_2}^{(i,n)}|\mu_2^{(i)})}{\sum_{j=1}^{n} p(\mu_{z_2}^{(i,n)}|\mu_2^{(j)})}$$

is used to build an expressive lookup table, where similarity of sequences is reflected by $\mu_2^{(i)}$ close in Euclidean space.
Hsu et al. [1] refer to the resulting objective

$$\mathcal{L}^{dis}(\theta, \phi; x) = \mathcal{L}(\theta, \phi, X) + \alpha \cdot log\ p(sequence\ id|z_2)$$

as discriminative segment variational lower bound, where hyperparameter $\alpha$ determines the weight of the discriminative objective.

Table 1: Comparison of speaker verification equal error rate (EER) on the TIMIT test set.

| Method | Dimension | $\alpha$ | EER |
|---|---|---|---|
| i-vector | 100 | - | 9.52% |
| $\boldsymbol{\mu}_2$ | 32 | $10^1$ | **2.38%** |
| $\boldsymbol{\mu}_1$ | 32 | $10^1$ | 22.47% |

## Discussion and Future Work

Hsu et al. [1] evaluate the FHVAE on different tasks. They propose an unsupervised speaker verification task to measure performance, and to some extent, quantitatively assess the grade of disentanglement. They outperform an i-vector baseline [2]. To identify speakers, they use $\mu_2$ which ideally should only encapsulate speaker-dependent attributes. As a sanity check, they further introduce a segment vector $\mu_1$. This variable is based on $z_1$ and thus, should ideally only contain information independent of the speaker. The results are displayed in Table 1. While using $\mu_2$ results in the lowest EER, they still achieve below random-baseline EER using $\mu_1$, indicating leakage of sequence-level information into $z_1$. This observation reveals a larger problem. While Hsu et al. [1] provide visual and audio qualitative examples to demonstrate the grade of disentanglement, they fail in providing sufficient quantitative evidence. Moreover, their work lacks a formal description and justification of the proposed factorization.
To further investigate the level of disentanglement, we evaluate the FHVAE on a toy dataset with a known generative process. While achieving good reconstruction of signals, the latent variables are not reflective of the generative factors. Sample results are shown in Figure 3.
To improve disentanglement, we propose to further exploit the used data by introducing a cross-reconstruction loss [11]. Additionally, we hypothesize that using a hierarchical activation function, as proposed by Shen et al. [12] could further encourage a clean factorization.

## Conclusion

Hsu et al. [1] propose an architecture suited to approximate an equivariant map with respect to a sequence-segment decomposition. They exploit the hierarchical nature of certain sequential data to form an inductive bias towards factorization of different temporal scales. We hypothesize that the notion of a sequence-dependent prior could be transferred to other settings. While achieving good performance on considered benchmark tasks, their work lacks quantitative evidence of achieved disentanglement. We advocate for the need for more foundational, formal work on disentanglement. We hypothesize that this would allow to fairly discuss, evaluate and compare methods. Otherwise, this emerging field might get ahead of itself, proposing elaborate architectures without sufficient formal frameworks, metrics, and baseline methods.

[2] According to Hsu et al. [1], i-vector is used in state-of-the-art speaker verification approaches. It is a low dimensional subspace of a Gaussian mixture universal background model which ideally only contains speaker information.

## References


[1] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. *arXiv preprint arXiv:1709.07902*, 2017.

[2] Pierre Comon. Independent component analysis, 1992.

[3] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

[5] Tejas D Kulkarni, Will Whitney, Pushmeet Kohli, and Joshua B Tenenbaum. Deep convolutional inverse graphics network. *arXiv preprint arXiv:1503.03167*, 2015.

[6] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[7] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[8] Georgia Benkart. Abstract algebra, by in herstein. *The American Mathematical Monthly*, 94(8): 804–806, 1987.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[11] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019.

[12] Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. Ordered neurons: Integrating tree structures into recurrent neural networks. *arXiv preprint arXiv:1810.09536*, 2018.