Interspeech 2018
2-6 September 2018, Hyderabad

Scalable Factorized Hierarchical Variational Autoencoder Training

Wei-Ning Hsu, James Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{wnhsu,glass}@mit.edu

Abstract

Deep generative models have achieved great success in unsupervised learning with the ability to capture complex nonlinear relationships between latent generating factors and observations. Among them, a factorized hierarchical variational autoencoder (FHVAE) is a variational inference-based model that formulates a hierarchical generative process for sequential data. Specifically, an FHVAE model can learn disentangled and interpretable representations, which have been proven useful for numerous speech applications, such as speaker verification, robust speech recognition, and voice conversion. However, as we will elaborate in this paper, the training algorithm proposed in the original paper is not scalable to datasets of thousands of hours, which makes this model less applicable on a larger scale. After identifying limitations in terms of runtime, memory, and hyperparameter optimization, we propose a hierarchical sampling training algorithm to address all three issues. Our proposed method is evaluated comprehensively on a wide variety of datasets, ranging from 3 to 1,000 hours and involving different types of generating factors, such as recording conditions and noise types. In addition, we also present a new visualization method for qualitatively evaluating the performance with respect to the interpretability and disentanglement. Models trained with our proposed algorithm demonstrate the desired characteristics on all the datasets.

Index Terms: unsupervised learning, speech representation learning, factorized hierarchical variational autoencoder

1. Introduction

Unsupervised learning can leverage large amounts of unlabeled data to discover latent generating factors that often lie on a lower dimensional manifold compared to the raw data. A learned latent representation from speech can be useful for many downstream applications, such speaker verification [1], automatic speech recognition [2], and linguistic unit discovery [3, 4]. A factorized hierarchical variational autoencoder (FHVAE) [5] is a variational inference-based deep generative model that learns interpretable and disentangled latent representation from sequential data without supervision by modeling a hierarchical generative process. In particular, it has been demonstrated that an FHVAE trained on speech data learns to encode sequence-level generating factors, such as speaker and channel condition, into one set of latent variables, while encoding segment-level generating factors, such as phonetic content, into another set of latent variables. The ability to disentangle latent factors has been beneficial to a wide range of tasks, including domain adaptation [6], conditional data augmentation [7], and voice conversion [5].

However, the original FHVAE training algorithm proposed in [5] does not scale to datasets of over hundreds of thousands of utterances, making it less applicable to real world settings, where an unlabeled dataset of such size is common. This limitation is mainly due to the following issues: (1) the inference model of the sequence-level latent variable, and (2) the design of the discriminative objective. To be more specific, the original training algorithm reduces the complexity of inferring sequence-level latent variables by maintaining a cache, whose number of entries equals the number of training sequences. In addition, the discriminative objective, which encourages disentanglement, requires computing a partition function that sums over the entries in that cache. The two facts combined lead to significant scalability issues.

In this paper, we propose a hierarchical sampling algorithm to address these issues. In addition, a new method for qualitatively evaluating disentanglement performance based on a t-Distribution Stochastic Neighbor Embedding [8] is also presented. The proposed training algorithm is evaluated on a wide variety of datasets, ranging from 3 to 1,000 hours and involving many different types of generating factors, such as recording conditions and noise types. Experimental results verify that the proposed algorithm is effective on all sizes of datasets and achieves desirable disentanglement performance. The code is available on-line.[1]

2. Limitations of Original FHVAE Training

In this section, we briefly review the factorized hierarchical variational autoencoder (FHVAE) model and discuss the scalability issues of the original training objective.

2.1. Factorized Hierarchical Variational Autoencoders

An FHVAE [5] is a variant of a VAE that models a generative process of sequential data with a hierarchical graphical model. Let $\boldsymbol{X} = \{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ be a sequence of N segments. An FHVAE assumes that the generation of a segment $\boldsymbol{x}$ is conditioned on a pair of latent variables, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, referred to as the *latent segment variable* and the *latent sequence variable* respectively. While $\boldsymbol{z}_1$ is generated from a global prior, similar to those latent variables in a vanilla VAE, $\boldsymbol{z}_2$ is generated from a sequence-dependent prior that is conditioned on a sequence-level latent variable, $\boldsymbol{\mu}_2$, named the *s-vector*. The joint probability of a sequence is formulated as: $p(\boldsymbol{\mu}_2)\prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{z}_1^{(n)}, \boldsymbol{z}_2^{(n)})p(\boldsymbol{z}_1^{(n)})p(\boldsymbol{z}_2^{(n)}|\boldsymbol{\mu}_2)$. With this formulation, an FHVAE learns to encode generating factors that are consistent within segments drawn from the same sequence into $\boldsymbol{z}_2$. In contrast, $\boldsymbol{z}_1$ captures the residual generating factors that changes between segments.

Since computing the true posteriors of $\boldsymbol{Z}_1 = \{\boldsymbol{z}_1^{(n)}\}_{n=1}^{N}$, $\boldsymbol{Z}_2 = \{\boldsymbol{z}_2^{(n)}\}_{n=1}^{N}$, and $\boldsymbol{\mu}_2$ are intractable, an FHVAE introduces an inference model, $q(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{\mu}_2|\boldsymbol{X})$, and factorizes it as: $q(\boldsymbol{\mu}_2|\boldsymbol{X})\prod_{n=1}^{N} q(\boldsymbol{z}_1^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{z}_2^{(n)})q(\boldsymbol{z}_2^{(n)}|\boldsymbol{x}^{(n)})$. We summarize in Table 1 the family of distributions an FHVAE adopts for the generative model and the inference model. All the functions, $f_{\mu_{\boldsymbol{x}}}$, $f_{\sigma^2_{\boldsymbol{x}}}$, $g_{\mu_{\boldsymbol{z}_1}}$, $g_{\sigma^2_{\boldsymbol{z}_1}}$, $g_{\mu_{\boldsymbol{z}_2}}$, and $g_{\sigma^2_{\boldsymbol{z}_2}}$, are neural networks that parameterize mean and variance of Gaussian distributions.

Table 1: *Family of distributions adopted for FHVAE generative and inference models.*

generative model	
$p(\boldsymbol{\mu}_2)$	$\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
$p(\boldsymbol{z}_1)$	$\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
$p(\boldsymbol{z}_2\|\boldsymbol{\mu}_2)$	$\mathcal{N}(\boldsymbol{\mu}_2, \sigma^2_{\boldsymbol{z}_2}\boldsymbol{I})$
$p(\boldsymbol{x}\|\boldsymbol{z}_1, \boldsymbol{z}_2)$	$\mathcal{N}(f_{\mu_{\boldsymbol{x}}}(\boldsymbol{z}_1, \boldsymbol{z}_2), diag(f_{\sigma^2_{\boldsymbol{x}}}(\boldsymbol{z}_1, \boldsymbol{z}_2)))$
inference model	
$q(\boldsymbol{\mu}_2\|\boldsymbol{X})$	$\mathcal{N}(\sum_{n=1}^{N} g_{\mu_{\boldsymbol{z}_2}}(\boldsymbol{x}^{(n)})/(N+\sigma^2_{\boldsymbol{z}_2}), \boldsymbol{I})$
$q(\boldsymbol{z}_1\|\boldsymbol{x}, \boldsymbol{z}_2)$	$\mathcal{N}(g_{\mu_{\boldsymbol{z}_1}}(\boldsymbol{x}, \boldsymbol{z}_2), diag(g_{\sigma^2_{\boldsymbol{z}_1}}(\boldsymbol{x}, \boldsymbol{z}_2)))$
$q(\boldsymbol{z}_2\|\boldsymbol{x})$	$\mathcal{N}(g_{\mu_{\boldsymbol{z}_2}}(\boldsymbol{x}), diag(g_{\sigma^2_{\boldsymbol{z}_2}}(\boldsymbol{x})))$

[1] https://github.com/wnhsu/ScalableFHVAE

10.21437/Interspeech.2018-1034

2.2. Original FHVAE Training

In the variational inference framework, since the marginal likelihood of observed data is intractable, we optimize the variational lower bound, $\mathcal{L}(p, q; \boldsymbol{X})$, instead. We can derive a sequence variational lower bound of $\boldsymbol{X}$ based on Table 1. However, this lower bound can only be optimized at the sequence level, because inferring of $\boldsymbol{\mu}_2$ depends on an entire sequence, and would become infeasible if $\boldsymbol{X}$ is extremely long.

In [5] the authors proposed replacing the maximum a posterior (MAP) estimation of $\boldsymbol{\mu}_2$'s posterior mean for training sequences with a cache $h_{\mu_{\boldsymbol{\mu}_2}}(i)$, where i indexes training sequences. In other words, the inference model for $\boldsymbol{\mu}_2$ becomes $q(\boldsymbol{\mu}_2|\boldsymbol{X}^{(i)}) = \mathcal{N}(h_{\mu_{\boldsymbol{\mu}_2}}(i), \boldsymbol{I})$. Therefore, the lower bound can be re-written as:

$$\mathcal{L}(p, q; \boldsymbol{X}^{(i)}) = \sum_{n=1}^{N^{(i)}} \mathcal{L}(p, q; \boldsymbol{x}^{(i,n)}|h_{\mu_{\boldsymbol{\mu}_2}}(i)) + \log p(h_{\mu_{\boldsymbol{\mu}_2}}(i)) \tag{1}$$

$$\begin{aligned}\mathcal{L}(p, q; \boldsymbol{x}^{(i,n)}|h_{\mu_{\boldsymbol{\mu}_2}}(i)) &= \mathbb{E}_{q(\boldsymbol{z}_1, \boldsymbol{z}_2|\boldsymbol{x}^{(i,n)})}[\log p(\boldsymbol{x}^{(i,n)}|\boldsymbol{z}_1, \boldsymbol{z}_2)] \\ &- \mathbb{E}_{q(\boldsymbol{z}_2|\boldsymbol{x}^{(i,n)})}[D_{KL}(q(\boldsymbol{z}_1|\boldsymbol{x}^{(i,n)}, \boldsymbol{z}_2)||p(\boldsymbol{z}_1))] \\ &- D_{KL}(q(\boldsymbol{z}_2|\boldsymbol{x}^{(i,n)})||p(\boldsymbol{z}_2|h_{\mu_{\boldsymbol{\mu}_2}}(i))).\end{aligned} \tag{2}$$

We can now sample a batch at the segment level to optimize the following variational lower bound:

$$\mathcal{L}(p, q; \boldsymbol{x}^{(i,n)}) = \mathcal{L}(p, q; \boldsymbol{x}^{(i,n)}|h_{\mu_{\boldsymbol{\mu}_2}}(i)) + \frac{1}{N^{(i)}} \log p(h_{\mu_{\boldsymbol{\mu}_2}}(i)). \tag{3}$$

Furthermore, to obtain meaningful disentanglement between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, it is not desirable to have constant $\boldsymbol{\mu}_2$ for all sequences; otherwise, for each segment, swapping z_1, z_2 would lead to the same objective. To avoid such condition, the following objective is added to encourage $\boldsymbol{\mu}_2$ to be discriminative between sequences:

$$\log p(i|\bar{\boldsymbol{z}}_2^{(i,n)}) := \log \frac{p(\bar{\boldsymbol{z}}_2^{(i,n)}|\bar{\boldsymbol{\mu}}_2^{(i)})}{\sum_{j=1}^{M} p(\bar{\boldsymbol{z}}_2^{(i,n)}|\bar{\boldsymbol{\mu}}_2^{(j)})}, \tag{4}$$

where M is the total number of training sequences, $\bar{\boldsymbol{z}}_2^{(i,n)} = g_{\mu_{\boldsymbol{z}_2}}(\boldsymbol{x}^{(i,n)})$ denotes the posterior mean of $\boldsymbol{z}_2$, and $\bar{\boldsymbol{\mu}}_2^{(i)} = h_{\mu_{\boldsymbol{\mu}_2}}(i)$ denotes the posterior mean of $\boldsymbol{\mu}_2$. This additional discriminative objective encourages $\boldsymbol{z}_2$ from the i-th sequence to be not only close to $\boldsymbol{\mu}_2$ of the i-th sequence, but also far from $\boldsymbol{\mu}_2$ of other sequences. Combining this discriminative objective and the segment variational lower bound with a weighting parameter, α, the objective function that an FHVAE maximizes then becomes:

$$\mathcal{L}^{dis}(p, q; \boldsymbol{x}^{(i,n)}) = \mathcal{L}(p, q; \boldsymbol{x}^{(i,n)}) + \alpha \log p(i|\bar{\boldsymbol{z}}_2^{(i,n)}), \tag{5}$$

referred to as the *discriminative segment variational lower bound*.

2.3. Scalability Issues

The original FHVAE training addressed the scalability issue with respect to sequence length by decomposing a sequence variational lower bound into sum of segmental variational lower bounds over segments. However, here we will show that this training objective is not scalable with respect to the number of training sequences.

First of all, the original FHVAE training maintains an M-entry cache, $h_{\mu_{\boldsymbol{\mu}_2}}(\cdot)$, that stores the posterior mean of $\boldsymbol{\mu}_2$ for each training sequence. The size of this cache scales linearly in the number of training sequences. Suppose $\boldsymbol{z}_2$ are 32-dimensional 32-bit floating point vectors as in [5]. If the number of training sequences is on the order of 10^8, the cache size would grow to about 10 GB and exhaust the memory of a typical commercial GPU (8 GB). Even worse, when computing the gradient given a batch of training segments, we need to maintain a tensor of size $(bs, |\theta|)$, where bs is the segment batch size, and $|\theta|$ is the total number of trainable parameters involved in the computation of the objective function. Since the computation of the discriminative objective involves the entire cache, $h_{\mu_{\boldsymbol{\mu}_2}}(\cdot)$, the gradient tensor is of size at least bs times larger than the cache. With a batch size of 256, a dataset with 10^5 sequences can exhaust the GPU memory during training.

Second, the denominator of the discriminative objective, $\sum_{i=1}^{M} p(\bar{\boldsymbol{z}}_2^{(i,n)}|\bar{\boldsymbol{\mu}}_2^{(j)})$, marginalizes over a function of posterior mean of $\boldsymbol{\mu}_2$ of all training sequences, which increases the computation time proportionally to the number of sequences for each training step.

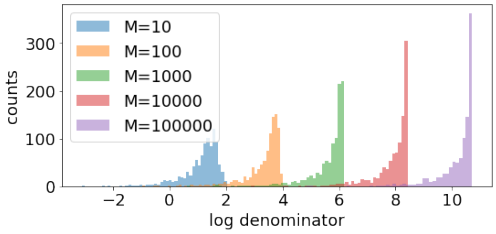

Figure 1: *Histogram of* $\log \sum_{i=1}^{M} p(\bar{\boldsymbol{z}}_2^{(i,n)}|\bar{\boldsymbol{\mu}}_2^{(j)})$ *with respect to different* $M \in \{10^1, 10^2, 10^3, 10^4, 10^5\}$. *Distributions shift by roughly a constant when* M *increases by 10 times, implying the denominator scales proportionally to* M.

Third, from the hyperparameter optimization point of view, the distribution of the denominator, $\sum_{i=1}^{M} p(\bar{\boldsymbol{z}}_2^{(i,n)}|\bar{\boldsymbol{\mu}}_2^{(j)})$, also changes with respect to the number of training sequences. Specifically, the expected value of this terms scales roughly linearly in M, as shown in Figure 1. Such behavior is not desirable, because the α parameter that balances the variational lower bound and the discriminative objective would need to be adjusted according to M.

3. Training with Hierarchical Sampling

In order to utilize the discriminative objective, while eliminating the memory, computation, and optimization issue induced by a large training set, we need to control the size of the cache as well as the denominator summation in the discriminative objective. Both of these can be achieved jointly with a hierarchical sampling algorithm.

Given a dataset of M training sequences, **we maintain a cache of only K entries, where K is a dataset independent hyperparameter, named the *sequence batch size***. We optimize an FHVAE model by repeating the following procedure until the convergence criterion is met: (1) Sample a batch of K sequences from the entire training set. (2) Reset each entry of the cache, $h_{\mu_{\boldsymbol{\mu}_2}}(k)$, with the MAP estimation, $\sum_{n=1}^{\tilde{N}^{(k)}} g_{\mu_{\boldsymbol{z}_2}}(\tilde{\boldsymbol{x}}^{(k,n)})/(\tilde{N}^{(k)} + \sigma_{\boldsymbol{z}_2}^2)$, where $\tilde{N}^{(k)}$ is the number of segments in the k-th sampled sequence, and $\tilde{\boldsymbol{x}}^{(k,n)}$ is the n-th segment of the k-th sampled sequence, for $k = 1, \ldots, K$. (3) Sample B_{seg} batches of segments sequentially from the K sampled sequences. Each batch of segments is used to estimate the discriminative segmental variational lower bound for optimizing the parameters of f_*, g_*, and $h_{\mu_{\boldsymbol{\mu}_2}}$ as before. The only difference is that *the denominator of the discriminative object now sums over the K sampled training sequences*, instead of the entire set of M training sequences. We list the pseudo code in Algorithm 1.

We refer to the proposed algorithm as a hierarchical sampling algorithm, because we first sample at the sequence level, and then at the segment level. The size of the cache is then controlled by the sequence batch size K, instead of the number of training sequences M. The algorithm can also be viewed as iteratively sampling a sub-dataset of K sequences and running the original training algorithm on it. Compared with the proposed algorithm, the original training algorithm can be regarded as a "flat" sampling algorithm, where we sample segments from the entire pool, so it is therefore necessary to maintain a cache of M entries. The proposed algorithm introduces an overhead associated with reseting the cache whenever a new batch of sequences is sampled. However, this cost can be amortized by increasing the number of segment batches B_{seg}, for each batch of sequences.

4. Experimental Setup

We evaluate our training algorithm on a wide variety of datasets, ranging from 3 to 1,000 hours, including both clean and noisy, close-talking and distant speech. In this section, we describe the datasets, and introduce FHAVE models and their training configurations.

Algorithm 1 Training with Hierarchical Sampling

Input: $\{\boldsymbol{X}^{(i)}\}_{i=1}^{M}$: training set; K: sequence batch size; bs: segment batch size; B_{seg}: number of segment batches; f_*/g_*: decoders/encoders; $h_{\mu_{\boldsymbol{\mu}_2}}$: cache of K entries; Optim: gradient descent-based optimizer

1: **while** not converged **do**
2: sample a batch of K training sequences, $\{\tilde{\boldsymbol{X}}^{(k)}\}_{k=1}^{K}$
3: **for** $k = 1 \dots K$ **do**
4: $h_{\mu_{\boldsymbol{\mu}_2}}(k) \leftarrow \sum_{n=1}^{N^{(k)}} g_{\mu_{\boldsymbol{z}_2}}(\tilde{\boldsymbol{x}}^{(k,n)})/(N + \sigma^2_{\boldsymbol{z}_2})$
5: **end for**
6: **for** $1 \dots B_{seg}$ **do**
7: sample segments $\{\tilde{\boldsymbol{x}}^{(k_b,n_b)}\}_{b=1}^{bs}$ from $\{\tilde{\boldsymbol{X}}^{(k)}\}_{k=1}^{K}$
8: $l_{dis}(b) \leftarrow -\log \frac{p(g_{\mu_{\boldsymbol{z}_2}}(\tilde{\boldsymbol{x}}^{(k_b,n_b)})|h_{\mu_{\boldsymbol{\mu}_2}}(k_b))}{\sum_{k=1}^{K} p(g_{\mu_{\boldsymbol{z}_2}}(\tilde{\boldsymbol{x}}^{(k_b,n_b)})|h_{\mu_{\boldsymbol{\mu}_2}}(k))}$
9: $l_{gen}(b) \leftarrow -\mathcal{L}(p, q; \tilde{\boldsymbol{x}}^{(k_b,n_b)})$
10: $loss \leftarrow \sum_{b=1}^{bs}(l_{gen}(b) + \alpha \cdot l_{dis}(b))/bs$
11: $f_*, g_*, h_{\mu_{\boldsymbol{\mu}_2}} \leftarrow \text{Optim}(loss, \{f_*, g_*, h_{\mu_{\boldsymbol{\mu}_2}}\})$
12: **end for**
13: **end while**
14: **return** f_*, g_*

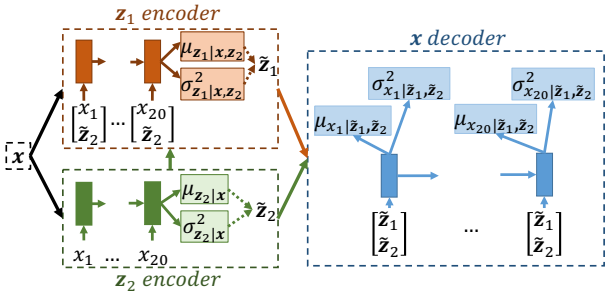

Figure 2: *The proposed FHVAE architecture consists of two encoders (orange and green) and one decoder (blue).* $\boldsymbol{x} = [x_1, \cdots, x_{20}]$ *is a segment of 20 frames. Dotted lines in the encoders denote sampling from parametric distributions.*

4.1. Datasets

Four different corpora are used for our experiments, which are TIMIT [9], Aurora-4 [10], AMI [11], and LibriSpeech [12]. TIMIT contains broadband 16kHz recordings of phonetically-balanced read speech. A total of 3,696 utterances (3 hours) are presented in the training partition based on the Kaldi [13] recipe, where `sa` utterances are excluded. In addition, manually labeled time-aligned phonetic transcripts are available, which we use to study the disentanglement performance between phonetic and speaker information achieved by FHVAE models.

Aurora-4 is another broadband corpus designed for noisy speech recognition tasks, based on the Wall Street Journal corpus (WSJ0) [14]. Six different types of noise are artificially mixed with clean read speech of two different microphone types, amounting to a total of 4,620 utterances (10 hours) for the development set. Following the training setup in [5], we train our FHVAE models on the development set, because the noise and channel information on the training set is not available.

The AMI corpus consists of 100 hours of meeting recordings, recorded in three different meeting rooms with different acoustic properties, and with multiple attendants. Multiple microphones are used for each session, including individual headset microphones (IHM), and far-field microphone arrays. IHM and single distant microphone (SDM) recordings from the training set are mixed to form a training set for the FHVAE models, including over 200,000 utterances according to the segmentation provided in the corpus.

The largest corpus we evaluate on is the LibriSpeech corpus, which contains 1,000 hours of read speech sampled at 16kHz. This corpus is based on the LibriVox's project, where world-wide volunteers record public domain texts to create free public domain audio books.

4.2. Training and Model Configurations

Speech segments of 20 frames, represented with 80-dimensional log Mel-scale filter bank coefficients (FBank), are used as inputs to FHVAE models. We denote each segment with $\boldsymbol{x} = [x_1, \cdots, x_{20}]$. The variance of $\boldsymbol{z}_2$'s prior is set to $\sigma^2_{\boldsymbol{z}_2} = 0.25$, and the dimension of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are both 32. Figure 2 illustrates the detailed encoder/decoder architectures of the proposed FHVAE model. The conditional mean and variance predictor for each variable (i.e., $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\boldsymbol{x}$) shares a common stacked LSTM pre-network, followed by two different single-layer affine transform networks, μ_* and σ^2_*, predicting the conditional mean and variance respectively. Specifically, a stacked LSTM with 2 layers and 256 memory cells are used for all three pre-networks, illustrated in Figure 2 with blocks filled with dark colors. Affine transform networks of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ encoders take as input the output from the last time step of both layers, which sums to 512 dimension. As for the $\boldsymbol{x}$ decoder, the affine transform network takes as input the LSTM output of the last layer from each time step t, and predicts the probability distribution of the corresponding frame $p(x_t|\boldsymbol{z}_1, \boldsymbol{z}_2)$. The same sampled $\tilde{\boldsymbol{z}}_1$ and $\tilde{\boldsymbol{z}}_2$ from the posterior distributions are concatenated and used as input for the LSTM decoder at each step. Sampling is done by introducing auxiliary input variables for the reparameterization trick [15], in order to keep the entire network differentiable with respect to the objective.

FHVAE models are trained to optimize the discriminative segment variational lower bound with $\alpha = 10$. We set sequence batch size $K = 2000$ for TIMIT and Aurora-4, and $K = 5000$ for the others. Adam [16] with $\beta_1 = 0.95$ and $\beta_2 = 0.999$ is used to optimize all models. Tensorflow [17] is used for implementation. Training is done for 500,000 steps, terminating early if the segmental variational lower bound on a held-out validation set is not improved for 50,000 steps.

5. Results and Discussions

5.1. Time and Memory Complexity

One feature of our proposed training algorithm is to control memory complexity. We found that a training set with over 100,000 sequences would exhaust a single 8GB GPU memory when using the original training algorithm. Hence, it was not feasible for the AMI and the LibriSpeech corpus, while the proposed algorithm does not suffer from the same problem. Another feature of hierarchical sampling is to control the time complexity of computing the discriminative loss. To study how sequence batch size affects the optimization step (line 8 in Algorithm 1), we evaluate the processing time of that step by varying K from 20 to 20,000 and show the results in Table 2.

We can observe that when $K \leq 2000$, the time complexity of computing the discriminative loss is fractional compared to computing the variational lower bound. However, when $K > 2000$, the increased computation time grows proportional to the sequence batch size, so that computation of the discriminative loss starts to dominate the time complexity. In practice, given a new encoder/decoder architecture, we can investigate the computation overhead resulting from the discriminative loss using such a method, and it is possible to determine some K that introduces negligible overhead for optimization.

Table 2: *Processing time of the optimization step with different sequence batch size* K.

K	10	100	1000	2000	5000	10000	20000
Time (ms)	84	84	86	87	103	147	230

5.2. Evaluating Disentanglement Performance

To examine whether an FHVAE is successfully trained, we need to inspect its performance at disentangling sequence-level generating factors (e.g. speaker identity, noise condition, and channel condition) from segment-level generating factors (e.g. phonetic content) in the latent space. For quantitative evaluation, we reproduce the speaker verification experiments in [5]. The FHVAE model trained with hierarchical sampling achieves 1.64% equal error rate on TIMIT, matching the performance of the original training algorithm (1.34%). In the following sections, we proceed with two qualitative evaluation methods.

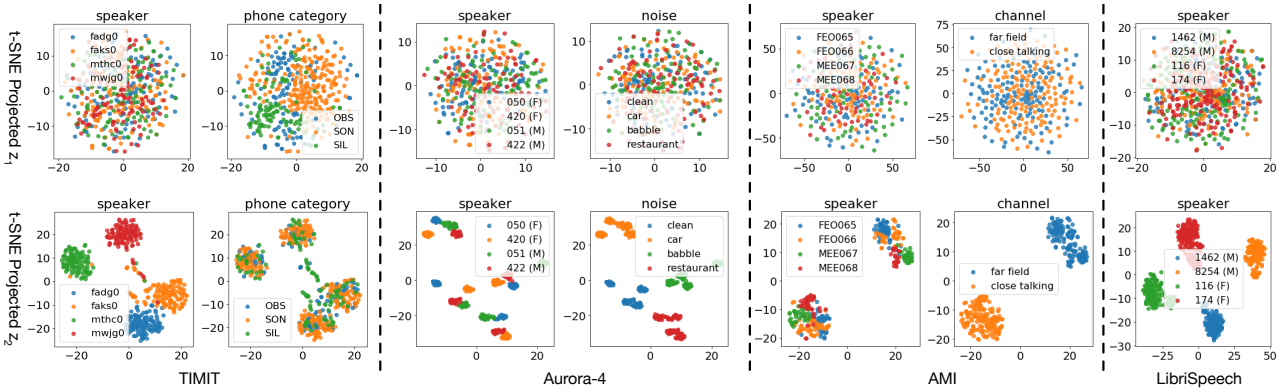

Figure 3: *Scatter plots of t-SNE projected $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ with models trained on TIMIT/Aurora-4/AMI/LibriSpeech. Each point represents one segment. Different colors are used to code segments of different labels with respect to the generating factor shown at the title of each plot.*

5.2.1. *t-SNE Visualization of Latent Variables*

We start with selecting a batch of labeled segments $(\boldsymbol{x}, y)$, where y denotes the values of the associated generating factors, for example y = (phone-id, speaker-id). We then infer $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ of these segments, and project them separately to a two-dimensional space using t-Distributed Stochastic Neighbor Embedding (t-SNE) [8]. Each generating factor is used to color-code both projected $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. Successful disentanglement would result in segments of the same sequence-level generating factors forming clusters in the projected $\boldsymbol{z}_2$ space but not in the projected $\boldsymbol{z}_1$ space, and vice versa.

For all four datasets, speaker label, a sequence-level generating factor, is available for each segment. Since time-aligned phonetic transcripts are available for TIMIT, it is also possible to derive phone labels, which is a segment-level generating factor. Following [18], we further reduce the 61 phonemes to three phonetic subsets: sonorant (SON), obstruent (OBS), and silence (SIL) for better color-coding. In addition, noise types can be obtained for Aurora-4, and microphone types can be obtained for AMI, which are both sequence-level generating factors.

Results of t-SNE projections for models trained on each dataset are shown in Figure 3, where each point represents one segment. It can be observed that in each of the projected $\boldsymbol{z}_2$ spaces, segments of the same sequence-level generating factors (speaker/noise/channel) always form clusters. When segments are generated conditioned on multiple sequence-level generating factors, as in Aurora-4 and AMI, the segments actually cluster hierarchically. In contrast, the distribution of projected $\boldsymbol{z}_1$'s does not vary between different values of these generating factors, which implies that $\boldsymbol{z}_1$ does not contain much information about them. Opposite phenomenon can be observed from the phone category-coded plots, where segments belonging to the same phonetic subset cluster in the projected $\boldsymbol{z}_1$ space, but not in the projected $\boldsymbol{z}_2$ space. These results suggest that FHVAEs trained with hierarchical sampling can achieve desirable disentanglement for these conditions.

5.2.2. *Reconstructing Re-combined Latent Variables*

Given two segments, $\boldsymbol{x}^A$ and $\boldsymbol{x}^B$, we sample a segment $\boldsymbol{x}^C \sim p(\boldsymbol{x}|\tilde{\boldsymbol{z}}_1^A, \tilde{\boldsymbol{z}}_2^B)$, where $\tilde{\boldsymbol{z}}_1^A$ is a sampled latent segment variable conditioned on $\boldsymbol{x}^A$, and $\tilde{\boldsymbol{z}}_2^B$ is a sampled latent sequence variable conditioned on $\boldsymbol{x}^B$. With a successfully trained FHVAE, $\boldsymbol{x}^C$ should exhibit the segment-level attributes of $\boldsymbol{x}^A$, and the sequence-level attributes of $\boldsymbol{x}^B$. Due to space limitations, we only show results of the model trained on the AMI corpus in Figure 4. Eight segments are sampled for $\boldsymbol{x}^A$, as shown in the upper right corner of the figure. Among these segments, the four leftmost ones are close-talking while the rest are far-field, and the first, second, fifth, and sixth from left are female speakers while the rest are males. For $\boldsymbol{x}^B$, three segments of different sequence-level generating factors are sampled, as shown on the left half of the figure. The segments used to infer $\tilde{\boldsymbol{z}}_2^B$ are highlighted in red boxes; we show the surrounding frames of those segments to better illustrate how sequence-level generating factors affect realization of observations.

Samples of $\boldsymbol{x}^C$ generated by re-combining latent variables are shown in the lower right corner of the figure. It can be clearly observed that $\boldsymbol{x}^C$ presents the same sequence-level generating factors as $\boldsymbol{x}^B$,[2] whose latent sequence variable $\boldsymbol{x}^C$ conditions on. Meanwhile, the phonetic content of $\boldsymbol{x}^C$ stays consistent with $\boldsymbol{x}^A$,[3] whose latent segment variable $\boldsymbol{x}^C$ conditions on. The clear differentiation of generating factors encoded in each sets of latent variables again corroborates the success of our proposed algorithm in training FHVAE models.

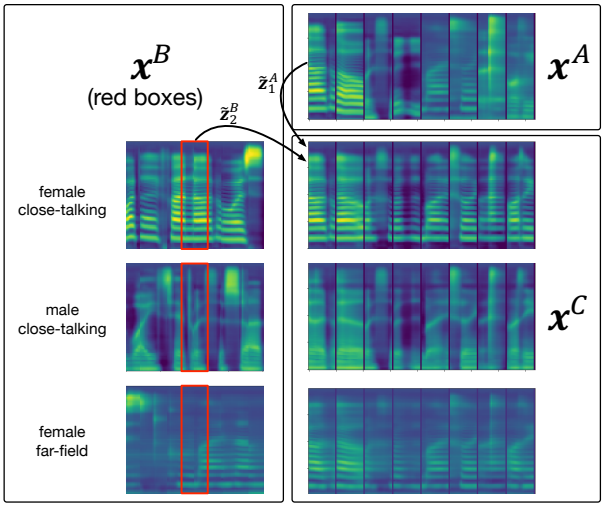

Figure 4: *Results of decoding re-combined latent variables. A segment in the $\boldsymbol{x}^C$ block is generated conditioned on the latent segment variable of a segment in the block $\boldsymbol{x}^A$ of the same column, and conditioned on the latent sequence variable of a red-box highlighted segment in the block $\boldsymbol{x}^B$ of the same row.*

6. Conclusions and Future Work

In this paper, we discuss the scalability limitations of the original FHVAE training algorithm in terms of runtime, memory, and hyperparameter optimization, and propose a hierarchical sampling algorithm to address this problem. Comprehensive study on the memory and time complexity, as well as disentanglement performance verify the effectiveness of the proposed algorithm on all scales of datasets, ranging from 3 to 1,000 hours. In the future, we plan to extend FHVAE applications, such as ASR domain adaptation [7] and audio conversion [5] to larger scales.

[2] In these images, harmonic spacing is the clearest cue for fundamental frequency differences. Far-field recordings tend to have lower signal-to-noise ratios, which results in blurrier images.

[3] Phonetic content can usually be determined by the spectral envelope, and relative position of formants.

7. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] S. Tan and K. C. Sim, "Learning utterance-level normalisation using variational autoencoders for robust automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 43–49.

[3] J. F. Drexler, "Deep unsupervised learning from speech," 2016.

[4] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Interspeech*, 2017, pp. 1273–1277.

[5] ——, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in Neural Information Processing Systems*, 2017.

[6] W.-N. Hsu and J. Glass, "Extracting domain invariant features by unsupervised learning for robust automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[7] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *Automatic Speech Recognition and Understanding (ASRU), 2017 IEEE Workshop on*. IEEE, 2017.

[8] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms." *Journal of machine learning research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[10] D. Pearce, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," Ph.D. dissertation, Mississippi State University, 2002.

[11] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

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

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[14] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

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

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

[17] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[18] A. K. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, 1999.