A Convolutional Deep Markov Model for Unsupervised Speech Representation Learning

Sameer Khurana, Antoine Laurent, Wei-Ning Hsu, Jan Chorowski, Adrian Łańcucki, Ricard Marxer, James Glass

A Convolutional Deep Markov Model for Unsupervised Speech Representation Learning

Sameer Khurana[1], *Antoine Laurent*[2], *Wei-Ning Hsu*[1], *Jan Chorowski*[3], *Adrian Lancucki*[4], *Ricard Marxer*[5], *James Glass*[1]

[1]MIT - CSAIL, Cambridge, USA
[2]LIUM - Le Mans University, France
[3]University of Wrocaw, Poland [4]NVIDIA Corporation, Poland [5]LIS - University of Toulon, France
skhurana@mit.edu, antoine.laurent@univ-lemans.fr

Abstract

Probabilistic Latent Variable Models (LVMs) provide an alternative to self-supervised learning approaches for linguistic representation learning from speech. LVMs admit an intuitive probabilistic interpretation where the latent structure shapes the information extracted from the signal. Even though LVMs have recently seen a renewed interest due to the introduction of Variational Autoencoders (VAEs), their use for speech representation learning remains largely unexplored. In this work, we propose Convolutional Deep Markov Model (ConvDMM), a Gaussian state-space model with non-linear emission and transition functions modelled by deep neural networks. This unsupervised model is trained using black box variational inference. A deep convolutional neural network is used as an inference network for structured variational approximation. When trained on a large scale speech dataset (LibriSpeech), ConvDMM produces features that significantly outperform multiple self-supervised feature extracting methods on linear phone classification and recognition on the Wall Street Journal dataset. Furthermore, we found that ConvDMM complements self-supervised methods like Wav2Vec and PASE, improving on the results achieved with any of the methods alone. Lastly, we find that ConvDMM features enable learning better phone recognizers than any other features in an extreme low-resource regime with few labelled training examples.

Index Terms: Neural Variational Latent Variable Model, Structured Variational Inference, Unsupervised Speech Representation Learning

1. Introduction

One of the long-standing goals of speech and cognitive scientists is to develop a computational model of language acquisition [1, 2, 3, 4]. Early on in their lives, human infants learn to recognize phonemic contrasts, frequent words and other linguistic phenomena underlying the language [5]. The computational modeling framework of generative models is well-suited for the problem of spoken language acquisition, as it relates to the classic analysis-by-synthesis theories of speech recognition [6, 7]. Although, generative models are theoretically elegant and informed by theories of cognition, most recent success in speech representation learning has come from self-supervised learning algorithms such as Wav2Vec [8], Problem Agnostic Speech Encoding (PASE) [9], Autoregressive Predictive Coding (APC) [10], MockingJay (MJ) [11] and Deep Audio Visual Embedding Network (DAVENet) [12]. Generative models present many advantages with respect to their discriminative counterparts. They have been used for disentangled representation learning in speech [13, 14, 15]. Due to the probabilistic nature of these models, they can be used for generating new data and hence, used for data augmentation [16, 17] for Automatic Speech Recognition (ASR), and anomaly detection [18].

In this paper, we focus solely on designing a generative model for low-level linguistic representation learning from speech. We propose Convolutional Deep Markov Model (ConvDMM), a Gaussian state-space model with non-linear emission and transition functions parametrized by deep neural networks and a Deep Convolutional inference network. The model is trained using amortized black box variational inference (BBVI) [19]. Our model is directly based on the Deep Markov Model proposed by Krishnan et. al [20], and draws from their general mathematical formulation for BBVI in non-linear Gaussian state-space models. When trained on a large speech dataset, ConvDMM produces features that outperform multiple self-supervised learning algorithms on downstream phone classification and recognition tasks, thus providing a viable latent variable model for extracting linguistic information from speech.

We make the following contributions:

1) Design a generative model capable of learning good quality linguistic representations, which is competitive with recently proposed self-supervised learning algorithms on downstream linear phone classification and recognition tasks.
2) Show that the ConvDMM features can significantly outperform other representations in linear phone recognition, when there is little labelled speech data available.
3) Lastly, demonstrate that by modeling the temporal structure in the latent space, our model learns better representations compared to assuming independence among latent states.

2. The Convolutional Deep Markov Model

2.1. ConvDMM Generative Process

Given the functions; $f_\theta(\cdot), u_\theta(\cdot)$ and $t_\theta(\cdot)$, the ConvDMM generates the sequence of observed random variables, $x_{1:T} = (x_1, \ldots, x_T)$, using the following generative process

$$z_1 \sim \mathcal{N}(0, I) \tag{1}$$
$$z_\tau | z_{\tau-1} \sim \mathcal{N}(t^\mu_\theta(z_{\tau-1}), t^\sigma_\theta(z_{\tau-1})) \quad \tau = 2, \ldots, L \tag{2}$$
$$e_{1:T} = u_\theta(z_{1:L}) \tag{3}$$
$$\mu_{1:T} = f_\theta(e_{1:T}) \tag{4}$$
$$x_t | e_t \sim \mathcal{N}(\mu_t, \gamma) \quad t = 1, \ldots, T \tag{5}$$

where T is a multiple of L, $T = k \cdot L$ and $z_{1:L}$ is the sequence of latent states. We assume that the observed and latent random variables come from a multivariate normal distribution with diagonal covariances. The joint density of latent and observed

variables for a single sequence is

$$p(z_{1:L}, x_{1:T}) = p(x_{1:T}|z_{1:L})p(z_1)\prod_{\tau=2}^{L} p(z_\tau|z_{\tau-1}). \quad (6)$$

For a dataset of i.i.d. speech sequences, the total joint density is simply the product of per sequence joint densities. The scale γ is learned during training.

The transition function $t_\theta : z_{\tau-1} \rightarrow (\mu_\tau, \sigma_\tau)\big|_{\tau=2}^{L}$ estimates the mean and scale of the Gaussian density over the latent states. It is implemented as a Gated Feed-Forward Neural Network [20]. The gated transition function could capture both linear and non-linear transitions.

The embedding function $u_\theta : z_{1:L} \rightarrow e_{1:T}$ transforms and up-samples the latent sequence to the same length as the observed sequence. It is parametrized by a four layer CNN with kernels of size 3, 1024 channels and residual connections. We use the activations of the last layer of the embedding CNN as the features for the downstream task. This is reminiscent of kernel methods [21] where the raw input data are mapped to a high dimensional feature space using a user specified feature map. In our case, the CNN plays a similar role, mapping the low-dimensional latent vector sequence, $z_{1:L} \in \mathbb{R}^{L\times 16}$, to a high dimensional vector sequence, $e_{1:T} \in \mathbb{R}^{T\times 1024}$, by repeating the output activations of the CNN k times, where $k = T/L$. In our case, k is 4 which is also the downsampling factor of the encoder function (§ 2.2). A similar module was used in Chorowski et. al [22], where they used a single CNN layer after the latent sequence.

The emission function $f_\theta : e_t \rightarrow (\mu_t)\big|_{t=1}^{T}$ (a decoder) estimates the mean of the likelihood function. It is a two-layered MLP with 256 hidden units and residual connections. We employ a low capacity decoder to avoid the problem of posterior collapse [23], a common problem with high capacity decoders.

2.2. ConvDMM Inference

The goal of inference is to estimate the posterior density of latent random variables given the observations $p(z|x)$. Exact posterior inference in non-conjugate models like ConvDMM is intractable, hence we turn to Variational Inference (VI) for approximate inference. We use VI and BBVI interchangeably throughout the rest of the paper. In VI, we approximate the intractable posterior $p(z|x)$ with a tractable family of distributions, known as the variational family $q_\phi(z|x)$, indexed by ϕ. In our case, the variational family takes the form of a Gaussian with diagonal covariance. Next, we briefly explain the Variational Inference process for ConvDMM.

Given a realization of the observed random variable sequence $x_{1:T} = x_1, \ldots, x_T$, the initial state parameter vector $\hat{z}_1$, and the functions $g_\phi(\cdot)$ and $c_\phi(\cdot)$, the process of estimating the latent states can be summarized as:

$$h_{1:L} = g_\phi(x_{1:T}) \quad (7)$$

$$\hat{z}_\tau|\hat{z}_{\tau-1}, x \sim \mathcal{N}(c_\phi^\mu(h_\tau, \hat{z}_{\tau-1}), c_\phi^\sigma(h_\tau, \hat{z}_{\tau-1})) \quad \tau = 2 \text{ to } L. \quad (8)$$

Let $T = k * L$, where k is the down-sampling factor of the encoder, $g_\phi : x_{1:T} \rightarrow h_{1:L}$ is the encoder function, c_ϕ is the combiner function that provides posterior estimates for the latent random variables.

We parameterize the encoder g_ϕ using a 13-layer CNN with kernel sizes (3, 3, 3, 3, 3, 4, 4, 3, 3, 3, 3, 3, 3), strides (1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1) and 1024 hidden channels. The encoder down-samples the input sequence by a factor of four. The last layer of the encoder with $h_{1:L}$ as its hidden activations has a receptive field of approximately 50. This convolutional architecture is inspired by [22], but other acoustic models such as Time-Depth Separable Convolutions [24], VGG transformer [25], or ResDAVENet [12] could be used here. We leave this investigation for future work.

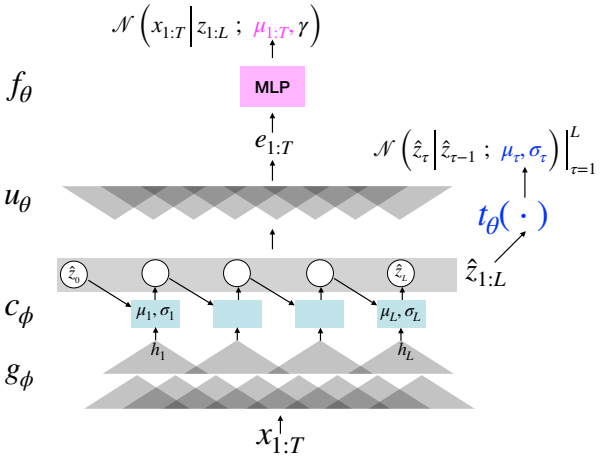

Figure 1: *An illustration of the ConvDMM.*

The combiner function c_ϕ provides structured approximations of the variational posterior over the latent variables by taking into account the prior Markov latent structure. The combiner function follows [20]:

$$h_{\text{combined}} = \frac{1}{2}(\tanh(W\hat{z}_{t-1} + b) + h_t)$$
$$\mu_t = W_\mu h_{\text{combined}} + b_\mu$$
$$\sigma_t = \text{softplus}(W_\sigma h_{\text{combined}} + b_\sigma).$$

It uses tanh non-linearity on z_{t-1} to approximate the transition function. Future work could investigate sharing parameters with the generative model as in Maaløe et. al's Bidirection inference VAE (BIVA) [26]. We note that structured variational inference in neural variational models is an important area of research in machine learning, with significant recent developemnts [27, 28]. Structured VAE has also been used for acoustic unit discovery [29], which is not the focus of this work.

2.3. ConvDMM Training

ConvDMM like other VAEs is trained to maximize the bound on model likelihood, known as the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x)||p_\theta(z))$$

where $p_\theta(x|z) = \prod_{t=1}^{T} p(x_t|e_t)$ is the Gaussian likelihood function and $p(x_t|e_t)$ is given by the ConvDMM generative process in Section 2.1. The Gaussian assumption lets us use the *reparametrization trick* [30] to obtain low-variance unbiased Monte-Carlo estimates of the expected log-likelihood, the first term in the R.H.S of the ELBO. The KL term, which is also an expectation can be computed similarly and its gradients can be obtained analytically. In our case, we use the formulation of Equation 12, Appendix A., in Krishnan et. al [20], to compute the KL term analytically.

Table 1: *Phone Classification (FER) and Recognition (PER) when trained on a subset of Wall Street Journal. Suffixes* -50, -360, -960 *denote the amount of hours of LibriSpeech training data used for training the unsupervised model.*

% of Labeled Data	1%		2%		5%		10%		50%		Low Shot (0.1%)
	FER	PER	FER	PER	FER	PER	FER	PER	FER	PER	PER
GaussVAE-960	-	55.8	-	50.1	-	48.2	-	45.9	-	42.5	-
Supervised Transfer-960	-	17.9	-	16.4	-	14.4	-	12.8	-	10.8	25.8 ($\pm$ 0.96)
***Self Supervised Learning*:**											
MockingJay-960 [11]	40.0	53.2	38.5	48.8	37.5	45.5	37.0	44.2	36.7	43.5	-
PASE-50 [9]	34.7	61.2	33.5	50.0	33.2	49.0	32.8	49.0	32.7	48.2	80.7 ($\pm$ 2.65)
Wav2Vec-960 [8]	19.8	37.6	19.1	27.7	18.8	24.5	18.6	23.9	18.5	22.9	78.0 ($\pm$ 10.4)
***Audio-Visual Self Supervised Learning*:**											
RDVQ (Conv2) [12]	31.6	44.1	30.8	42.4	30.5	41.1	30.1	41.3	30.2	40.6	52.6 ($\pm$ 0.95)
***Proposed Latent Variable Model*:**											
ConvDMM-50	29.6	37.8	28.6	35.4	27.9	31.3	27.9	30.3	27.0	29.1	-
ConvDMM-360	28.2	34.8	27.0	30.8	26.4	28.2	25.9	27.7	25.7	26.7	-
ConvDMM-960	27.7	32.5	26.6	30.0	26.0	28.1	26.0	27.1	25.6	26.0	50.7 ($\pm$ 0.57)
***Modeling PASE and Wav2Vec features using the proposed generative model*:**											
ConvDMM-960-PASE-50	-	35.4	-	32.6	-	30.6	-	29.3	-	28.4	55.3 ($\pm$ 3.21)
ConvDMM-Wav2Vec-960	-	28.6	-	25.7	-	22.3	-	21.2	-	20.4	40.7 ($\pm$ 0.42)

[1] $\pm$ refers to the standard deviation in the results

The model is trained using the Adam optimizer with a learning rate of 0.001 for 100 epochs. We reduce the learning rate to half of its value if the loss on the development set plateaus for three consecutive epochs. L2 regularization on model parameters with weight 5e-7 is used during training. To avoid latent variable collapse we use KL annealing [23] with a linear schedule, starting from an initial value of 0.5, for the first 20 epochs of training. We use a mini-batch size of 64 and train the model on a single NVIDIA Titan X Pascal GPU.

3. Experiments

3.1. Evaluation Protocol and Dataset

We evaluate the learned representations on two tasks; phone classification and recognition. For phone classification, we use the ConvDMM features, the hidden activations from the last layer of the embedding function, as input to a softmax classifier, a linear projection followed by a softmax activation. The classifier is trained using Categorical Cross Entropy to predict framewise phone labels. For phone recognition the ConvDMM features are used as input to a softmax layer which is trained using Connectionist Temporal Classification (CTC) [31] to predict the output phone sequence. We do not fine-tune the ConvDMM feature extractor on the downstream tasks. The performance on the downstream tasks is driven solely by the learned representations as there is just a softmax classifier between the representations and the labels. The evaluation protocol is inspired by the unsupervised learning works in the Computer Vision community [32, 33], where features extracted from representation learning systems trained on ImageNet are used as input to a softmax classifier for object recognition. Neural Networks for supervised learning have always been seen as feature extractors that project raw data into a linearly separable feature space making it easy to find decision boundaries using a linear classifier. We believe that it is reasonable to expect the same from unsupervised representation learning methods and hence, we compare all the representation learning methods using the aforementioned evaluation protocol.

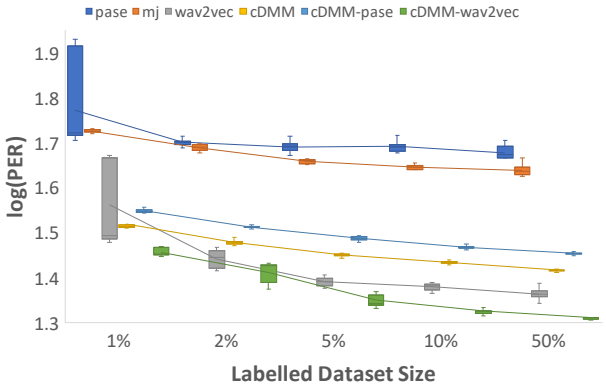

Figure 2: *PER on WSJ eval92 dataset using features extracted from different models.*

We train ConvDMM on the publicly available LibriSpeech dataset [34]. To be comparable with other representation learning methods with respect to the amount of training dataset used, we train another model on a small 50 hours subset of LibriSpeech. For evaluation, we use the Wall Street Journal (WSJ) dataset [35].

3.2. Results & Discussion

Table 1 presents framewise linear phone classification (FER) and recognition (PER) error rates on the WSJ eval92 dataset for different representation learning techniques. ConvDMM is trained on Mel-Frequency Cepstral Coefficients (MFCCs) with concatenated delta and delta-delta features. ConvDMM-50 and PASE-50 are both trained on 50 hours of LibriSpeech, ConvDMM-960, Wav2Vec-960 and MockingJay-960

are trained on 960 hours. ConvDMM-360 is trained on the 360 hours of the clean LibriSpeech dataset. RDVQ is trained on the Places400k spoken caption dataset [4]. We do not train any of the representation learning systems that are compared against ConvDMM on our own. We use the publicly available pretrained checkpoints to extract features. The linear classifiers used to evaluate the features extracted from unsupervised learning systems are trained on different subsets of WSJ train dataset, ranging from 4 mins (0.1%) to 40 hours (50%). To study the effect of modeling temporal structure in the latent space as in ConvDMM, we train a Gauss VAE which is similar to the ConvDMM except that it does not contain the transition model and hence, is a traditional VAE with isotropic Gaussian priors over the latent states [30].

To generate the numbers in the table we perform the following steps. Consider, for example, the column labelled 1% as we describe how the numbers are generated for different models (rows). We randomly pick 1% of the speech utterances in the WSJ train dataset. This is performed three times with different random seeds, yielding three different 1% data splits of labelled utterances from the WSJ train dataset. We then train linear classifiers on the features extracted using different representation learning systems, on each of the three splits five times with different random seeds. This gives us a total of 15 classification and recognition error rates. The final number is the mean of these numbers after removing the outliers. Any number greater than $q_3 + 1.5 * iqr$ or less than $q_1 - 1.5 * iqr$, where q_1 is the first Quartile, q_3 is the third Quartile and iqr is the inter-quartile range, is considered an outlier. We follow the same procedure to create different training splits, 2%, 5%, 10%, 50%, from the WSJ train dataset and present classification error rates in the table for all splits. Figure 2 shows the box plot for the PER on WSJ eval92 dataset using features extracted from different models.

In terms of PER, ConvDMM-50 outperforms PASE by 23.4 percentage points (pp), MockingJay by 15.4pp and RDVQ by 6.3pp under the scenario when 1% of labeled training data is available to the linear phone recognizer, which corresponds to approximately 300 spoken utterances ($\approx$ 40 mins). Compared to Wav2Vec, ConvDMM lags by 0.2pp, but the variance in Wav2Vec results is very high as can be seen in Figure 2. Under the 50% labeled data scenario, ConvDMM-50 outperforms MockingJay by 14.4pp, PASE by 19.1pp, RDVQ by 11.5pp and lags Wav2Vec by 6.2pp. The gap between ConvDMM-50 and RDVQ widens in the 50% labeled data case. ConvDMM-960 similarly outperforms all the methods under the 1% labeled data scenario, outperforming Wav2Vec, the second best method, by 5.1pp. Also the variance in the ConvDMM-960 results is much lower than Wav2Vec (See Figure 2). ConvDMM systematically outperforms the Gauss VAE which does not model the latent state transitions, showing the value of prior structure.

ConvDMM-PASE which is the ConvDMM model built on top of PASE features instead of the MFCC features, outperforms PASE features by 25.8pp under the 1% labeled data scenario. A significant gap exists under all data scenarios. Similar results can be observed with ConvDMM-Wav2Vec model, but the improvements over Wav2Vec features is not as drastic, probably due to the fact that Wav2Vec already produces very good features. For low shot phone recognition with 0.1% labeled ($\approx$ 4 mins), ConvDMM-960 significantly outperforms all other methods. Surprisingly, RDVQ shows excellent performance under this scenario. ConvDMM-Wav2Vec-960 performs 10pp better than ConvDMM-960 trained on MFCC features and 38pp better than Wav2Vec features alone. We could not get below 90% PER with MockingJay and hence, skip reporting the results.

Lastly, we compare the performance of features extracted using unsupervised learning systems trained on LibriSpeech vs features extracted using the fully supervised system neural network acoustic model trained on the task of phone recognition on 960 hours of labeled data (See the row labeled Supervised Transfer-960). The supervised system has the same CNN encoder as the ConvDMM. There is a glaring gap between the supervised system and all other representation learning techniques, even in the very few data regime (0.1%). This shows there is still much work to be done in order to reduce this gap.

4. Related Work

Another class of generative models that have been used to model speech but not explored in this work are the autoregressive models. Autoregressive models, a class of explicit density generative models, have been used to construct speech density estimators. Neural Autoregressive Density Estimator (NADE) [36] is a prominent earlier work followed by more recent Wavenet [37], SampleRNN [38] and MelNet [39]. An interesting avenue of future research is to probe the internal representations of these models for linguistic information. We note that, Waveglow, a flow based generative model is recently proposed as an alternative to autoregressive models for speech [40].

5. Conclusions

In this work, we design the Convolutional Deep Markov Model (ConvDMM), a Gaussian state-space model with non-linear emission and transition functions parametrized by deep neural networks. The main objective of this work is to demonstrate that generative models can reach the same, or even better, performance than self supervised models. In order to do so, we compared the ability of our model to learn linearly separable representations, by evaluating each model in terms of PER and FER using a simple linear classifier. Results show that our generative model produces features that outperform multiple self-supervised learning methods on phone classification and recognition task on Wall Street Journal. We also find out that these features can achieve better performances than all other evaluated features when learning the phone recogniser with very few labelled training examples. Another interesting outcome of this work is that by using self-supervised extracted features as input of our generative model, we produce features that outperforms every other one in the phone recogniser task. Probably due to enforcing temporal structure in the latent space. Lastly, we argue that features learned using unsupervised methods are significantly worse than features learned by a fully supervised deep neural network acoustic model, setting the stage for future work.

6. References

[1] S. Goldwater, T. L. Griffiths, and M. Johnson, "A bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[2] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[3] L. Ondel, L. Burget, and J. Černockỳ, "Variational inference for

acoustic unit discovery," *Procedia Computer Science*, vol. 81, pp. 80–86, 2016.

[4] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems*, 2016, pp. 1858–1866.

[5] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *Cognition*, vol. 173, pp. 43–59, 2018.

[6] M. Halle and K. Stevens, "Speech recognition: A model and a program for research," *IRE transactions on information theory*, vol. 8, no. 2, pp. 155–159, 1962.

[7] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code." *Psychological review*, vol. 74, no. 6, p. 431, 1967.

[8] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[9] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *arXiv preprint arXiv:1904.03416*, 2019.

[10] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

[11] A. T. Liu, S. wen Yang, P.-H. Chi, P. chun Hsu, and H. yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," 2019.

[12] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1elCp4KwH

[13] 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.

[14] S. Khurana, S. R. Joty, A. Ali, and J. Glass, "A factorial deep markov model for unsupervised disentangled representation learning from speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6540–6544.

[15] Y. Li and S. Mandt, "Disentangled sequential autoencoder," *arXiv preprint arXiv:1803.02991*, 2018.

[16] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 16–23.

[17] W.-N. Hsu, H. Tang, and J. Glass, "Unsupervised adaptation with interpretable disentangled representations for distant conversational speech recognition," *arXiv preprint arXiv:1806.04872*, 2018.

[18] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=Hkxzx0NtDB

[19] R. Ranganath, S. Gerrish, and D. M. Blei, "Black box variational inference," *arXiv preprint arXiv:1401.0118*, 2013.

[20] R. G. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models," in *Thirty-first aaai conference on artificial intelligence*, 2017.

[21] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, pp. 1171–1220, 2008.

[22] J. Chorowski, R. J. Weiss, S. Bengio, and A. 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.

[23] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[24] A. Hannun, A. Lee, Q. Xu, and R. Collobert, "Sequence-to-sequence speech recognition with time-depth separable convolutions," *arXiv preprint arXiv:1904.02619*, 2019.

[25] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for asr," *arXiv preprint arXiv:1904.11660*, 2019.

[26] L. Maaløe, M. Fraccaro, V. Liévin, and O. Winther, "Biva: A very deep hierarchy of latent variables for generative modeling," in *Advances in neural information processing systems*, 2019, pp. 6548–6558.

[27] M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta, "Composing graphical models with neural networks for structured representations and fast inference," in *Advances in neural information processing systems*, 2016, pp. 2946–2954.

[28] W. Lin, N. Hubacher, and M. E. Khan, "Variational message passing with structured inference networks," *arXiv preprint arXiv:1803.05589*, 2018.

[29] J. Ebbers, J. Heymann, L. Drude, T. Glarner, R. Haeb-Umbach, and B. Raj, "Hidden markov model variational autoencoder for acoustic unit discovery." in *INTERSPEECH*, 2017, pp. 488–492.

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

[31] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[32] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019.

[33] O. J. Hénaff, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord, "Data-efficient image recognition with contrastive predictive coding," *arXiv preprint arXiv:1905.09272*, 2019.

[34] V. Panayotov, G. Chen, D. Povey, and S. 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.

[35] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[36] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, "Neural autoregressive distribution estimation," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 7184–7220, 2016.

[37] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[38] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[39] S. Vasquez and M. Lewis, "Melnet: A generative model for audio in the frequency domain," *arXiv preprint arXiv:1906.01083*, 2019.

[40] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.