INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Modeling and Transforming Speech using Variational Autoencoders

Merlijn Blaauw, Jordi Bonada

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
merlijn.blaauw@upf.edu, jordi.bonada@upf.edu

Abstract

Latent generative models can learn higher-level underlying factors from complex data in an unsupervised manner. Such models can be used in a wide range of speech processing applications, including synthesis, transformation and classification. While there have been many advances in this field in recent years, the application of the resulting models to speech processing tasks is generally not explicitly considered. In this paper we apply the variational autoencoder (VAE) to the task of modeling frame-wise spectral envelopes. The VAE model has many attractive properties such as continuous latent variables, prior probability over these latent variables, a tractable lower bound on the marginal log likelihood, both generative and recognition models, and end-to-end training of deep models. We consider different aspects of training such models for speech data and compare them to more conventional models such as the Restricted Boltzmann Machine (RBM). While evaluating generative models is difficult, we try to obtain a balanced picture by considering both performance in terms of reconstruction error and when applying the model to a series of modeling and transformation tasks to get an idea of the quality of the learned features.

Index Terms: generative models, variational autoencoder, acoustic modeling, deep learning

1. Introduction

Generative models are frequently used for machine learning problems in unsupervised or semi-supervised settings. One of their goals is to extract simpler, higher-level features from complex data which can then be used for classification, generating new data or transforming existing data; in this last case the generative model can act as a regularizer which ensures the generated data will behave like the data seen during training.

We can think of these simpler, higher-level features to be (related to) some hidden underlying factors that may be involved in the process that generated the data. The observed data that we focus on in this paper is the frame-wise spectral envelope of speech signals. In this case we know that these envelopes contain phonetic information, speaker identity, voice quality, influences related, etc. While there is some understanding of how these factors influence the spectrum, "handcrafting" features and algorithms to deal with these factors can be difficult [1]. Therefore, we try to approach this problem by modeling and learning from data.

In this paper we will first introduce the variational autoencoder (VAE) and set it in a context of more conventional models, such as the Restricted Boltzmann Machine (RBM) and some other more recent models. While evaluating and comparing generative models is a generally difficult task [2], we evaluate performance considering both reconstruction error and performance in a set of sampling and transformation tasks.

2. The Variational Autoencoder

The variational autoencoder [3, 4] defines a probabilistic generative model,

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}), \tag{1}$$

where our observed data $\mathbf{x}$ is generated by a random process involving some underlying random variables $\mathbf{z}$. In this model the prior $p_\theta(\mathbf{z})$ quantifies what we know about about $\mathbf{z}$ before seeing any data, and the likelihood function $p_\theta(\mathbf{x}|\mathbf{z})$ quantifies how the observed data $\mathbf{x}$ relates to $\mathbf{z}$. Here both $p_\theta(\mathbf{z})$ and $p_\theta(\mathbf{x}|\mathbf{z})$ are parametric families of distributions with parameters θ. The posterior $p_\theta(\mathbf{z}|\mathbf{x})$ can be used to infer the hidden latent variables $\mathbf{z}$ from data, or optimize model parameters θ maximizing marginalized likelihood $p_\theta(\mathbf{x})$.

By defining the prior distribution to be something simple (much simpler than the empirical data distribution) we try to ensure that the latent variables $\mathbf{z}$ are higher-level, simpler features of the data $\mathbf{x}$. Typically a standard Gaussian distribution is used,

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, I). \tag{2}$$

For real-valued data the likelihood, or observation model, is typically modeled using an independent Gaussian distribution,

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mu_\theta(\mathbf{z}), \sigma^2_\theta(\mathbf{z})I), \tag{3}$$

where $\mu_\theta(\mathbf{z})$ and $\sigma^2_\theta(\mathbf{z})$ are non-linear functions of $\mathbf{z}$ modeled using a neural network, e.g. a multi-layer perceptron (MLP).

Using the exact posterior given by Bayes' rule $p_\theta(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})/\int p_\theta(\mathbf{x}, \mathbf{z})d\mathbf{z}$ leads to intractable computations. The idea behind variational inference is to approximate the intractable true posterior with some tractable parametric auxiliary distribution $q_\phi(\mathbf{z}|\mathbf{x})$. In the VAE framework this is typically chosen to be an independent Gaussian distribution,

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x})I). \tag{4}$$

Like the generative model, this recognition model is also parameterized by a neural network.

We now have an optimization problem of fitting the approximate posterior to the true posterior by minimizing their Kullback-Leibler (KL) divergence. While this is not a tractable objective itself, when we expand the KL divergence the relationship to the marginalized likelihood becomes clear,

$$\begin{aligned} &D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})] \\ &\quad = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p_\theta(\mathbf{z}|\mathbf{x})] &(5)\\ &\quad = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log q_\phi(\mathbf{z}|\mathbf{x})] - \log p_\theta(\mathbf{x}, \mathbf{z})] + \log p_\theta(\mathbf{x}). &(6)\end{aligned}$$

Given that $D_{KL}(\cdot) \geq 0$, we obtain a lower bound on the marginalized likelihood,

$$\begin{aligned} \log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] &(7)\\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] &(8)\\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})], &(9)\end{aligned}$$

http://dx.doi.org/10.21437/Interspeech.2016-1183

which is the variational objective $\mathcal{L}(\theta, \phi; \mathbf{x})$.

The model can be trained by jointly optimizing θ and ϕ to maximize $\mathcal{L}$ using stochastic gradient ascent, however care must be taken when computing the gradient w.r.t. ϕ. Using the so-called *reparameterization trick* the expectation over $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ is transformed into an expectation over an auxiliary distribution independent of ϕ and a deterministic mapping, e.g. $\mathbf{z} = \mu_\phi(\mathbf{x}) + \sigma_\phi(\mathbf{x}) \odot \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$. Finally, we approximate the expectation term of eq. (9) using Monte Carlo integration. Often one or few samples suffice for typical mini-batch sizes. In many cases the KL divergence term can be solved analytically to further reduce variance of the estimator.

The connection with the classical deterministic autoencoder (AE) becomes clear when viewing the model as a stack of the recognition (encoder) and generation (decoder) MLP networks. Looking at eq. (9), we see that the first term corresponds to a stochastic version of the negative reconstruction error in the classical AE, while the second term is an additional regularizer which encourages the approximate posterior to be close to the prior.

3. Related work

The original VAE publications and many of their derived works only report results on image data. While speech data is not generally considered, there is at least one notable exception [5], where the raw time-domain speech signal is modeled using a time-aware recurrent version of the VAE. Directly modeling the time domain signal is very desirable as it does away with any "handcrafted" vocoder or signal processing. However we feel that at this point working with a more high-level abstraction of speech, such as the spectral envelope, will be beneficial to obtaining higher quality results and require less training data. Another domain in which VAEs have recently been successful is natural language processing, e.g. [6].

Compared to the more conventional generative models used in speech processing, the closest model is the Restricted Boltzmann Machine (RBM), or its multi-layer, stacked counterpart the Deep Belief Network (DBN). These models have been used for tasks such as speech recognition [7], speech synthesis [8], and other tasks like voice conversion [9]. Perhaps the biggest difference between these models and the VAE is that the RBM (in practice) is limited to binary latent variables, while the VAE typically uses continuous latent variables (using alternative gradient estimators to the one described in sec. 2, binary latent variables are possible as well). The Bayesian framework in which the VAE is set is more easily interpretable and allows for a more flexible choice of the distributions involved, compared to the undirected probabilistic RBM model. VAEs offer a tractable lower bound on the log likelihood (and easy approximation of true log likelihood by means of importance sampling [4]), while RBMs use a different approximate objective (contrastive divergence) which is much more difficult to interpret and complicates training. Finally, stacked RBMs (DBNs) have to be trained in a greedy layer-wise fashion, while deep VAEs can be trained end-to-end.

Apart from VAEs, there have been several other neural network-based generative models proposed recently. One class of models that has received a lot of attention is the Generative Adversarial Net (GAN) [10]. Like VAEs, these models contain a generative network, but instead of a recognition network they are trained using a discriminator which decides whether a randomly generated sample is real or not. This type of model has been able to produce very convincing results in the image domain, often cited to avoid the oversmoothing from which VAEs can suffer. Some of the downsides of GANs are that balancing the relative strength of generative and discriminative networks can complicate training, and that the lack of a recognition network can be a limitation for transformation tasks like the ones discussed in this paper.

Highlighting the flexibility of the model, there have been many recent works that build upon the VAE considered in this paper. These include improved accuracy of the approximate inference using multi-sample objectives [11] or more expressive variational posteriors [12], sequential generative models [13], time recurrent models [5] and combinations with graphical or state-space models [14]. Considering the task of transformation, conditional VAE models have also been proposed for structured output prediction [15].

4. Methodology

4.1. Feature space and pre-processing

We are interested in modeling the timbre of the voice by first extracting its spectral envelope, assuming that (small) variations in pitch do not affect it significantly. The STRAIGHT vocoder [16] allows decomposing a speech signal into F0, and a harmonic and aperiodic spectral envelope. From these features a high quality reconstruction can then be obtained, even after moderate modification. Our initial features are 1025-dimensional (bin-by-bin) harmonic spectral envelopes, sampled every 5 milliseconds, obtained from recordings with a sampling rate of 32 kHz.

We should consider the observation model used when deciding which is an appropriate feature space. Ideally this would be a measure of the perceptual similarity between input and reconstruction. In the image domain, where a pixel-wise loss is generally not the best choice, there has been some work in this area [17, 18]. For spectral envelopes, which are relatively smooth, we expect an independent Gaussian observation model to work quite well. However, using features with a non-linear frequency scale may help allocate more of the model's capacity to the perceptually more relevant bottom range of the spectrum.

We consider two different representations of the spectral envelope, which we call MGC70 and SPEC257. The first representations is a 70-dimensional mel-generalized cepstral (MGC) representation (using parameters $\gamma = 0$ and $\alpha = 0.45$) [19]. By using the mel-scale these features should be able significantly reduce dimensionality while still being able to perform perceptually accurate reconstruction. The latter representation is a 257-dimensional log-spectral envelope obtained by downsampling the original STRAIGHT spectral envelope. The log-spectral distance between input and reconstruction from these reduced-dimensionality features over the dataset used in the experiments (see sec. 5.1) is 2.00 dB and 0.13 dB for MGC70 and SPEC257 respectively.

Additionally, we perform per-frame energy normalization of the initial STRAIGHT spectra. Leading and trailing silences are trimmed from the training utterances. The final features are standardized to have zero mean and unit variance across the training set.

4.2. Model architecture

The baseline model architecture used in this work consists of a standard Gaussian prior, independent Gaussian observation model, and independent Gaussian approximate posterior. The encoder and decoder MLPs both consist of two hidden layers of

512 units each with ELU [20] activations, and μ and σ^2 output layers with linear and softplus activations respectively (to constrain $\sigma^2 > 0$). We use a single stochastic latent layer consisting of 100 units. The expectation in eq. (9) is approximated using a single Monte Carlo sample. Model parameters are optimized using the Adam algorithm [21] with a learning rate of 0.0002, for 1000 epochs, with a batch size of 256. Batch normalization [22] is used for all hidden layers in the encoder and decoder to stabilize training and speed up convergence. We have found these settings to work well for most cases, but they have not been exhaustively optimized.

The reference RBM model (following [9]) is a Gaussian-Bernoulli RBM (GB-RBM) with 1539 continuous observed variables (513 bin spectrum, augmented with delta and delta-delta dynamic features) and 2048 binary latent variables. The observed Gaussian units use fixed unit variances and to stabilize training samples are fixed to their mean values. Training is performed according to approximate maximum likelihood using the contrastive divergence algorithm with a single step of Gibbs sampling. Parameters are optimized using stochastic gradient descent with a learning rate of 0.0001, 0.5 momentum for the first 5 epochs and 0.9 afterwards, a batch size of 10, for a total of 60 epochs. Reconstruction is done according to maximum output probability considering dynamic features [23]. This more elaborate parameter generation scheme helps smooth out excessive frame-to-frame variance from which the RBM can suffer. For the VAE model we found this not necessary.

4.3. Training methodology

When training the VAE model, one issue can be the relative dominance of the reconstruction and KL divergence terms in eq. (9) over the course of training. In the case that the KL divergence term is overly dominant, a great number of latent units become inactive in the reconstruction. When this happens $q_\phi(z_u|\mathbf{x})$ for a given latent unit z_u becomes very close to its prior $p(z_u)$ and thus does not encode any information about $\mathbf{x}$. Likewise, if the reconstruction term is overly dominant, most or all latent units will be active in the reconstruction, but their distribution will be very far from the prior distribution. The former case will often not be able to produce accurate reconstructions, while the latter case will. However in the latter case the model will behave much like a deterministic AE and the learned features will generally not be of interest.

The first problem is mentioned in some other works on image and natural language processing [24, 6]. The KLD term is too strong early on in training, causing the model to get stuck in a state where most latent units are inactive and the reconstruction does not improve. However, for spectral envelope data we encountered the second problem. The model very quickly learns how to accurately reconstruct its input, at which point the variances of the independent Gaussian observation model become smaller and smaller. The model thus achieves a higher approximate log likelihood by increasing the reconstruction term at the cost of also increasing the KLD penalty, but at a lower rate. One explanation for this might be that this type of data is inherently simpler to reconstruct. Spectral envelopes are relatively smooth 1d curves, while images are highly structured 2d data. Different samples in a speech dataset may have greater coherence compared to image datasets, i.e. it is likely that many frames have rather similar spectral envelopes, especially the lower frequencies which tend to have relatively low noise.

We propose to reduce the effects of this problem by limiting likelihood of the Gaussian observation model $p_\theta(\mathbf{x}|\mathbf{z})$ by constraining its variances to be greater than a minimum *variance floor*. We achieve this by using an offset softplus activation on the generation network's variance output,

$$\sigma^2(y) = \zeta + \log(1 + e^y), \tag{10}$$

where y is the linear output activation and $\zeta \geq 0$ is an additional hyper-parameter to be tuned which controls the variance floor (labeled "vfloor" in the experiments).

Additionally, we can make reconstruction more difficult for the network by using the denoising VAE objective [25],

$$\mathcal{L}_{dvae} = \mathbb{E}_{p(\tilde{\mathbf{x}}|\mathbf{x})}\, \mathbb{E}_{q_\phi(\mathbf{z}|\tilde{\mathbf{x}})}[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\tilde{\mathbf{x}})], \tag{11}$$

where $p(\tilde{\mathbf{x}}|\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \lambda I)$ is the Gaussian corruption distribution, with noise level $\lambda \geq 0$. That is, noise is injected at the input of the encoder, while the decoder still tries the reconstruct the uncorrupted input. This may have the additional benefit of improving generalization performance of the model. The additional expectation in eq. (11) is is approximated using a single Monte Carlo sample.

5. Experiments

5.1. Dataset

The dataset used in these experiments consisted of a set of 123 Spanish utterances from a professional female speaker. The utterances were optimized to cover most frequently occurring diphones and were spoken at an approximately constant pitch and cadence. All recordings were studio quality and resampled to 32 kHz for the experiments. Train set and validation set were split 80/20 (90.3k/23.9k samples, 451.8/119.5 seconds).

5.2. Quantitative experiments

The results of a series of reconstruction experiments with different model configurations is summarized in table 1. Looking at the baseline models without the use of any variance flooring or input noise, we see the KLD is overly high (resulting in poor features). The model seems to be able to handle high dimensional data well, with SPEC257 slightly outperforming MGC70 features. The variance floor and noise level have a similar effect on the reconstruction/KLD trade-off, without either being significantly better. The qualitative experiments show results for the SPEC257 model with $\zeta = 0.4$ and $\lambda = 1.0$. The RBM performs a little worse than the VAE in terms of reconstruction error and has higher frame-to-frame variance. The qualitative experiments could not be compared due to the RBM's latent space being binary and not allowing smooth interpolation.

5.3. Latent space sampling

One experiment by which generative models are often evaluated in the image domain is to generate random samples and visually inspecting if the resulting image looks realistic (and sufficiently different from the nearest neighbor in the training set). In our case a sample would correspond to a spectral envelope for a single frame which is difficult to evaluate in isolation. Instead we generate random latent variables that smoothly vary in time using a Gaussian diffusion process. The resulting time sequence of spectra can then be synthesized and listened to. While not evaluated formally, we consider that the resulting sounds could conceivably be produced by the source speaker, albeit lacking the temporal structure and pitch inflections of natural speech. Some sound examples can be found at: http://www.dtic.upf.edu/~mblaauw/IS2016_VAE

Table 1: Quantitative results of resynthesis experiments on held-out data. The log likelihood, $\log p_\theta(\mathbf{x})$, reconstruction term, $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$, and KL divergence term, $D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]$, were estimated using importance sampling with 500 samples [4, appendix E]. The number of active latent units is estimated using the metric, $A_u = \mathrm{Cov}_{\mathbf{x}}(\mathbb{E}_{q_\phi(z_u|\mathbf{x})}[z_u])$ and $A_u > 10^{-2}$, where u is the latent unit of interest [11]. The LSD and MCD35 columns are reconstruction log-spectral distance and 35-dimensional mel-cepstral distance respectively, both w.r.t. the input prior to dimensionality reduction. The global variance (GV) column is per-utterance variance over time of reconstructed features (averaged over all features, all utterances, and as a percentage of GV of input).

	log lik.	**Rec. term**	**KLD term**	**Num. active**	**LSD (dB)**	**MCD35 (dB)**	**GV (%)**
Baseline A (MGC70)	-43.38	11.10	62.21	40	3.48	2.43	87.13
+ vfloor $\zeta = 0.6$, noise $\lambda = 0.0$	-79.06	-67.23	15.52	27	5.03	3.69	77.89
+ vfloor $\zeta = 0.4$, noise $\lambda = 1.0$	-78.49	-73.57	10.25	18	6.25	4.46	75.66
+ vfloor $\zeta = 0.2$, noise $\lambda = 1.0$	-81.52	-67.91	15.02	25	5.03	3.74	76.07
+ vfloor $\zeta = 0.15$, noise $\lambda = 1.5$	-77.34	-72.19	10.68	17	6.24	4.44	74.96
Baseline B (SPEC257)	206.63	364.03	178.00	77	0.67	0.51	100.15
+ vfloor $\zeta = 0.6$, noise $\lambda = 0.0$	-208.53	-195.82	16.34	26	3.59	2.26	87.42
+ vfloor $\zeta = 0.4$, noise $\lambda = 1.0$	-170.76	-162.06	14.58	20	4.22	2.62	82.20
+ vfloor $\zeta = 0.2$, noise $\lambda = 1.0$	-109.91	-99.49	22.33	29	3.81	2.38	84.14
+ vfloor $\zeta = 0.15$, noise $\lambda = 1.5$	-112.81	-109.80	18.13	22	4.39	2.77	80.67
RBM	-	-	-	-	5.71	4.11	110.38

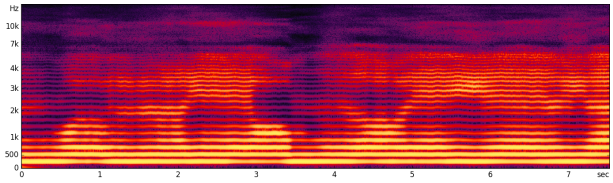

Figure 1: Reconstructed spectrogram of a continuous stream of speech-like sounds generated by exploring the trained model's latent space using a Gaussian diffusion process.

5.4. Latent space interpolation

One desirable property of a generative model is that an interpolation in the latent space leads to a reasonable interpolation in the observed space. For speech one would expect formant frequencies to transition smoothly when interpolating between latent representations of different vowels for instance. This is in fact one of the goals of certain "handcrafted" spectral representations such as Line Spectral Pairs (LSP) [26]. In fig. 2 we compare the results of linearly interpolating single frames from the vowel sequence [a], [e], [i], [o], [u] in SPEC257 observed space, in VAE latent space, and a reference recording. The STRAIGHT reconstructions use vowel timing, F0 and aperiodic components of the reference recording. The VAE result shows a clearer transitioning of formant frequencies compared to the observed space interpolation, even though the concept of a formant is never explicitly introduced in the system.

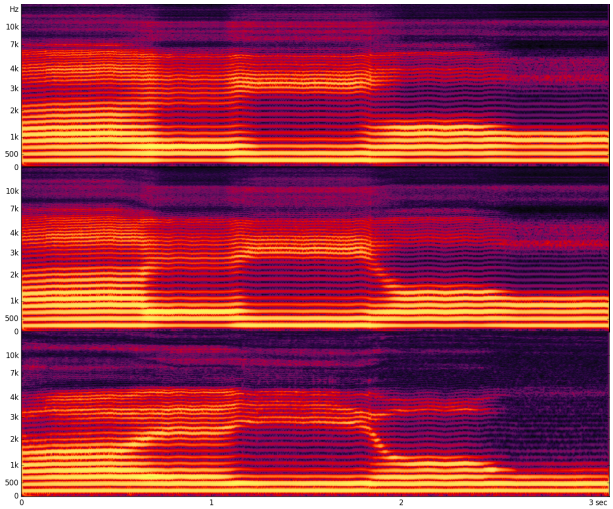

Figure 2: Top: Reconstructed spectrogram from interpolation in observed SPEC257 space. Middle: Reconstruction from interpolation in VAE latent space. Bottom: STRAIGHT-SPEC257 vocoder resynthesis of reference recording (different speaker).

6. Conclusions

In this work we have considered the variational autoencoder for modeling frame-wise spectral envelopes of speech signals. While our findings are still preliminary, in our experiments the VAE model could achieve similar or slightly better reconstruction errors compared to competitive models such as the RBM. While evaluating the usefulness of the learned latent representation is not a straight-forward task, we propose a set of simple problems considering the downstream task of transforming speech. The first is observing formant movements when interpolating between latent representations of different vowels, and the second is generating a continuous stream of samples from the model using a Gaussian diffusion process. We considered some of the issues that arose during training due to the different nature of the speech data compared to the usual image data. In order to better train the model for this type of data we propose using variance flooring in the observation model and using a denoising objective. We feel that the VAE shows promise and its flexibility should allow for many future improvements and applications.

7. Acknowledgements

We would like to thank the developers of the libraries used in this work: Theano, Lasagne and Parmesan. All experiments were carried out on a TITAN X GPU kindly donated by the NVIDIA Corporation.

8. References

[1] Y. Stylianou, "Voice transformation: A survey," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-09)*, April 2009, pp. 3585–3588.

[2] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *Proceedings of the International Conference on Learning Representations (ICLR-16)*, 2016.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proceedings of the International Conference on Learning Representations (ICLR-14)*, 2014.

[4] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1278–1286, arXiv:1401.4082.

[5] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in Neural Information Processing Systems 28 (NIPS-15)*, 2015, pp. 2980–2988.

[6] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proceedings of the International Conference on Learning Representations (ICLR-16)*, 2016.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[8] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2129–2139, Oct 2013.

[9] L. H. Chen, Z. H. Ling, L. J. Liu, and L. R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, Dec 2014.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27 (NIPS-14)*, 2014, pp. 2672–2680.

[11] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," in *Proceedings of the International Conference on Learning Representations (ICLR-16)*, 2016.

[12] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 1530–1538.

[13] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 1462–1471.

[14] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams, "Composing graphical models with neural networks for structured representations and fast inference," 2016, arXiv:1603.06277.

[15] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems 28 (NIPS-15)*, 2015, pp. 3483–3491.

[16] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-08)*, 2008, pp. 3933–3936.

[17] K. Ridgeway, J. Snell, B. D. Roads, R. S. Zemel, and M. C. Mozer, "Learning to generate images with perceptual similarity metrics," 2015, arXiv:1511.06409.

[18] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)*, 2016, pp. 1558–1566.

[19] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP-94)*, 1994.

[20] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proceedings of the International Conference on Learning Representations (ICLR-16)*, 2016.

[21] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR-15)*, 2015.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 448–456.

[23] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-00)*, vol. 3, 2000, pp. 1315–1318.

[24] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "How to train deep variational autoencoders and probabilistic ladder networks," 2016, arXiv:1602.02282.

[25] D. J. Im, S. Ahn, R. Memisevic, and Y. Bengio, "Denoising criterion for variational auto-encoding framework," 2015, arXiv:1511.06406.

[26] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, Feb 1975.