

On Representation Learning for Artificial Bandwidth Extension

Matthias Zöhrer¹, Robert Peharz² and Franz Pernkopf¹

Signal Processing and Speech Communication Lab, Graz University of Technology¹,
BEE-PRI BioTechMed-Graz, iDN – Inst. of Physiology, Medical University of Graz²,

matthias.zoehrer@tugraz.at, robert.peharz@tugraz.at, pernkopf@tugraz.at

Abstract

Recently, sum-product networks (SPNs) showed convincing results on the ill-posed task of artificial bandwidth extension (ABE). However, SPNs are just one type of many architectures which can be summarized as representational models. In this paper, using ABE as benchmark task, we perform a comparative study of Gauss Bernoulli restricted Boltzmann machines, conditional restricted Boltzmann machines, higher order contractive autoencoders, SPNs and generative stochastic networks (GSNs). Especially the latter ones are promising architectures in terms of its reconstruction capabilities. Our experiments show impressive results of GSNs, achieving on average an improvement of 3.90dB and 4.08dB in segmental SNR on a speaker dependent (SD) and speaker independent (SI) scenario compared to SPNs, respectively.

Index Terms: Bandwidth extension, representation learning, general stochastic network, sum-product network

1. Introduction

In recent years, deep learning techniques have been enjoying great success in machine learning, signal processing and speech technology [1, 2, 3]. A recently proposed deep architecture are sum-product networks (SPNs) [4, 5, 6], which represent highly structured probability distributions while still allowing exact and tractable inference. They show convincing performance on the task of image completion, i.e. recovering missing (covered) parts of an image given the observed parts [4, 5, 7]. Motivated by this success, SPNs have been recently used to model speech and were applied to the task of artificial bandwidth extension (ABE) [6] for enhancing telephone speech signals. They outperformed state-of-the-art systems in terms of log-spectral distortion (LSD) and in informal listening tests.

In [6], SPNs were only compared to classical signal processing methods, namely an HMM system based on a vocal tract filter model using linear predictions [8] and a Gaussian mixture model. In this paper, we apply many models from representation learning on the ABE task, including Gauss Bernoulli restricted Boltzmann machines (GBRBMs) [9], conditional restricted Boltzmann machines (CGBRBMs) [10], and higher order contractive autoencoders (HCAEs) [11]. Furthermore, we evaluated the recently proposed general stochastic networks (GSNs) [12, 13] generalizing denoising autoencoders. In [14] it was shown that GSNs, under mild conditions, indirectly capture the data distribution as the stationary distribution of a Markov chain, defined by a corruption/denoising process. In [12], GSNs showed convincing reconstruction results, which,

similar as argued in [6] for SPNs, motivates their usage for ABE. In [13], GSNs have been extended to a hybrid generative-discriminative learning objective. They outperformed many state-of-the-art models in classification, confirming their representational power.

In this paper, we systematically compare these deep learning approaches for ABE. In particular, we advocate for GBRBMs, CGBRBMs, HCAEs and GSNs a *filter* approach. In the training case we feed telephone spectrograms into the model in a frame-wise fashion and map it to the full-band signal. In the test scenario we infer the missing frequencies given the telephone band. Our experiments show that HMM-SPNs achieve the best log-spectral distortion (LSD) – slightly better than GSNs, while GSNs outperform HMM-SPNs in terms of segmental SNR (sSNR), resulting in an average relative improvement of 3.90dB and 4.08dB on both speaker dependent (SD) and speaker independent (SI) tasks, respectively.

The paper is organized as follows: In Section 2 we discuss the used representation models. Section 3 describes our experimental setup for ABE using these representational models. Section 3.3 presents experimental results and Section 4 concludes the paper.

2. Representational models

The first class of representational learning architectures are *restricted Boltzmann machines* (RBMs) [15, 16, 17, 9, 18]. RBMs are a particular form of log-linear Markov random fields, i.e. the energy function is linear in its free parameters. Learning in RBMs can be performed via contrastive divergence training, i.e. a kind of block Gibbs sampling applied to the RBM Markov chain for k -steps. RBMs can be used as generative models capable of learning a representation of the underlying data. RBMs can also be extended to learn a real valued representation of the data i.e. Gauss Bernoulli RBMs (GBRBMs) [9], or to learn temporal relations, i.e. Conditional RBMs (CGBRBMs) [10]. They also form the basis of more complex and powerful neural networks, i.e. deep belief networks [18] if stacked and trained in a greedy manner. Therefore they are widely used in many applications [19, 20].

The second class are (*deep*) *autoencoders* (AE) [21, 22, 23, 11, 24, 14]. AEs map an input to a hidden representation and transfer the latent representation back into a reconstruction of the input. AEs are mainly used as filters, feature extractors [23] or data generators [14] optimized via back-propagation and capable of learning a data distribution. An interesting variant is the higher order contractive autoencoder (HCAE) [11]. HCAEs regularize the norm of the Jacobian (*analytically*) and the Hessian (*stochastically*) to obtain a more robust representation of the underlying data.

The third class of representational models are *sum-product*

This work was supported by the Austrian Science Fund (FWF) under the project number P25244-N15. Furthermore, we acknowledge NVIDIA for providing GPU computing resources.

networks. SPNs can be interpreted as deep neural networks containing sum and product nodes, where the sums perform a non-negatively weighted sum over their inputs, while product nodes compute unweighted products over their inputs. The inputs to the SPNs are distributions over model variables. SPNs represent a joint distribution over the model variables if certain structural constraints called *completeness* and *decomposability* are fulfilled [4, 25, 7]. SPNs allow *exact* and *efficient* inference, i.e. marginalization over an arbitrary subset of the model variables can be performed in time linear to the network size (i.e. the number of edges). SPNs are able to represent distributions with high degree of dependency among the model variables.

The fourth class of representational models are *general stochastic networks* [12, 13], extending the class of AEs to multiple hidden layers, which are jointly optimized during training.

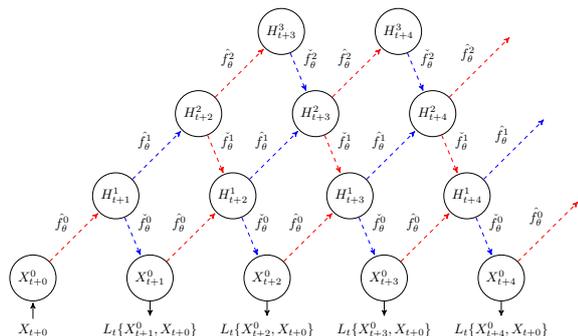


Figure 1: *Unfolded multi-layer GSN with backprop-able stochastic units* [13].

Figure 1 visualizes an unfolded multi-layer GSN for $k = 4$ walkback steps, described in detail in [13]. GSNs indirectly capture the data distribution as the stationary distribution of the Markov chain, defined by a corruption/denoising process, under mild conditions. Due to *walkback* training and their hierarchical structure which is jointly optimized they form a powerful model class, especially when used for input reconstruction [12].

The model uses backprop-able stochastic neurons, modeled with the help of functions of random variables $f_{\theta}^i \supseteq \{\hat{f}_{\theta}^i, \check{f}_{\theta}^i\}$. These functions express a Markov chain with additional dependencies between the hidden states, i.e. $H_{t+1} \sim P_{\theta_1}(H|H_{t+0}, X_{t+0})$, $X_{t+1} \sim P_{\theta_2}(X|H_{t+1})$. In particular, the density \hat{f}_{θ}^i models $H_{t+1}^i = \hat{f}_{\theta}^i(X_{t+0}, Z_{t+0}, H_{t+0})$, specified for some independent noise source Z_{t+0} , with the condition that the input X_{t+0} cannot be recovered exactly from H_{t+1} . The function $\check{f}_{\theta}^i = \eta_{out}^i + g(\eta_{in}^i + \hat{a}^i)$ is a backprop-able stochastic non-linearity for layer i , where $Z_t^i \supseteq \{\eta_{in}^i, \eta_{out}^i\}$ are noise processes and g is a non-linear activation function. The term $\hat{a}^i = W^i I_t^i + b^i$ defines the activations for layer i with a weight matrix W^i and bias b^i , representing the parametric distribution P_{θ_1} . In general, $\hat{f}_{\theta}^i(I_t^i)$ specifies an upward path in a GSN, where the input I_t^i is either the realization x_t^i of observed sample X_t^i or the hidden realization h_t^i of H_t^i . In the case of $X_{t+1}^i = \check{f}_{\theta}^i(Z_{t+0}, H_{t+1})$, $\check{f}_{\theta}^i(H_t^i) = \eta_{out}^i + g(\eta_{in}^i + \check{a}^i)$ defines a downward path in the network i.e. $\check{a}^i = (W^i)^T H_t^i + b^i$, using the transpose of the weight matrix W^i . This formulation allows to directly back-propagate the reconstruction log-likelihood $P_{\theta_2}(X|H)$ for all parameters $\theta \supseteq \{W^0, \dots, W^d, b^0, \dots, b^d\}$ using multiple functions of random variables $f_{\theta} \in \{\hat{f}_{\theta}^0, \dots, \hat{f}_{\theta}^d, \check{f}_{\theta}^0, \dots, \check{f}_{\theta}^d\}$, where d is the number of hidden layers.

3. ABE using Representational Models

3.1. Experimental Setup

We follow the same experimental setup as in [6], using 2 male speakers, i.e. 1, 2 and two female speakers, i.e. 18 and 20 from the GRID corpus [26] as test speakers. We simulate narrow-band telephone speech as in [6] by applying a bandpass filter with stop frequencies of 50 Hz and 4000 Hz. For speaker dependent (SD) models we use 10 minutes of speech from the respective speaker. For speaker independent (SI) models we used 10 minutes of speech obtained from the remaining 30 speakers of the corpus, each speaker providing approximately 20 seconds of speech. For testing we reserved 50 utterances not used for training. We use log-magnitude spectral data extracted from frames of 512 samples with 75% overlap. Furthermore, we applied a Hamming window before the Fourier transform. All signals were sampled at 16 kHz which yields a frame length of 32 ms and a frame rate of 125 Hz.

For SPNs we used the same model as in [6], using 64 SPNs as observation models for 64 states in the HMM. The hidden states correspond to a clustering on the training data using the LBG algorithm [27]. This model learns a probability distribution over the clean data of full bandwidth spectrograms. After training, given a telephone band signal, it performs most probable explanation (MPE) inference [4] to complete the missing frequency bands.

For the GSN, HCAE, GBRBM and CGBRBM models we take a different approach; Here we interpret the telephone band as a noisy version of the full band. The telephone band is fed into the network and the reconstruction is compared to the signal of full bandwidth, i.e. a mapping from telephone band to the full band is modeled. In contrast to SPNs, we perform a frame-wise processing with the GSNs, HCAEs GBRBMs and CGBRBMs. For evaluation we replace the predicted telephone band by the original low-frequency band.

In order to find the optimal model size a grid test on SD data using a GSN over $M \times d$ layers, where $M \in \{256, 500, 1000, 2000, 3000\}$ are the neurons per layer and $d \in \{1, 2, 3\}$ is performed. A Gaussian pre- and post activation noise with $\sigma = 0.1$ is used. Sigmoid RBM- and HCAE variants were configured with network size of 2000×1 . The optimal GSN is 2000×2 network with lateral connections in the hidden layers [28] and rectifier activation functions. All models used linear downward activations in the first layer allowing to fully generate the zero-mean and unit variance normalized data. The network weights were initialized with a uniform distribution [29] and trained with early stopping. Stochastic gradient descent (SGD) with a momentum term $\gamma = 0.9$ and an annealed learning rate $\eta_{t=0} = 0.1$ with $\eta_{t+1} = \eta_t \cdot 0.995$ was used for all models.

For re-synthesizing time signals from the log-magnitude spectrogram reconstructions we use the same approach as in [6], using 100 iterations of the Griffin&Lim algorithm [30] to synthesize phase for the reconstructed time-frequency bins. For the low frequency band the phase and magnitude from the initial telephone band is used. In [8], the phase from the telephone band is taken for the upper-band. This is also a good alternative for computationally constrained systems.

3.2. Evaluation Objectives

All evaluation measures are determined from the re-synthesized time-domain signal. For objective evaluation, we use the log-spectral distortion (LSD) in the high-band, similar as in [8, 6]. Using 9th order LPC analysis of each frame we get the spectral

envelope as

$$E_{\mathbf{a}}(e^{j\Omega}) = \frac{\sigma}{|\sum_{k=0}^9 a_k e^{-jk\Omega}|}, \quad (1)$$

where σ is the square-root of the variance of the LPC-analyzed signal and $\mathbf{a} = (a_0, \dots, a_9)$ are the LPC coefficients. The high-band LSD in [dB] for the τ^{th} frame is computed as

$$\text{LSD}_{\tau} = \sqrt{\frac{\int_{\nu}^{\pi} (20 \log E_{\mathbf{a}_{\tau}}(e^{j\Omega}) - 20 \log E_{\hat{\mathbf{a}}_{\tau}}(e^{j\Omega}))^2 d\Omega}{\pi - \nu}}, \quad (2)$$

where $\nu = \pi \frac{4000}{f_s/2}$, f_s is the sampling frequency, and \mathbf{a}_{τ} and $\hat{\mathbf{a}}_{\tau}$ are the LPC coefficients of the τ^{th} frame of the original and reconstructed signal, respectively. We report the utterance LSD given as the average of LSD_{τ} over all frames. Furthermore, we compute the unweighted segmental SNR (sSNR) in the frequency domain [31], limited in the range of [-10dB, 35dB].

In a detailed analysis we also considered wide-band PESQ [32] (WB-PESQ), which provides an instrumental prediction for the mean opinion score (MOS) to show the improvement obtained by the proposed methods compared to the narrow-band telephone signal and original wide-band signal. It was reported that WB-PESQ correlates well with subjective test results [33].

3.3. Results

The LSD and sSNR values of the frame-wise and HMM-based models [6] for the SD and SI tasks are shown in Table 1 and 2, respectively. Furthermore, we added the LSD and sSNR for the narrow-band signal. When looking at the LSD scores on both SI and SD tasks, the HMM-SPN slightly outperforms the GSN. However a pair-wise T-test with $p = 0.05$ reveals no significance in this difference. Furthermore, as mentioned above we replaced the predicted telephone band by the original one. This results in slightly lower LSD values for GSNs, but removes audible artifacts. In particular, the predicted full bandwidth signals of GSNs achieve an average LSD of 2.15dB and 2.93dB compared to 2.81dB and 3.59dB for the SD and SI task. The GSN achieved an average improvement of 3.90dB in sSNR compared to the HMM-SPN in the SD task. In case of the SI task the average improvement between both models in sSNR is 4.08dB. Both improvements are statistically significant when using a T-test with $p = 0.1$. Most notably, the HMM-SPN consists of 64 sub-models connected to one HMM, whereas the GSN uses a single model and simple frame-wise processing. The GSN also clearly outperforms the GBRBM and CGBRBM variants, due to its ability to handle multi-modal input data distributions. The GSN jointly optimizes multiple network layers at the same time. Higher layers contribute to the modeling process of the lower layers. We conjecture that this is a key reason for the good overall performance of the model when compared to single layer network variants, such as HCAEs and CGBRBMs. On the downside, far more calculations are needed to compute an output reconstruction of a single frame with a GSN. In particular, the model needs $2 \cdot k \cdot d$ matrix-vector computations to pass the input information to higher layers and to compute the output reconstruction of a single frame, whereas a single layer HCAE needs 2 matrix-vector calculations, i.e. one for the upward pass and one for the downward pass. Nevertheless, reducing the bit-width [34] might allow to implement the proposed method on computationally constrained systems.

The performance of different models is also reflected in the spectrogram reconstructions in Figure 2. It shows reconstruc-

Model	s1	s2	s18	s20	avg.
LSD [dB]					
narrow-band	6.98	7.58	6.66	6.48	6.93
GBRBM	16.34	10.66	6.50	6.11	9.90
CGBRBM	3.72	3.91	3.74	3.50	3.71
HMM-GMM [6]	3.18	2.93	2.28	2.82	2.80
HMM-SPN [6]	3.12	2.84	2.15	2.59	2.68
HCAE	3.07	3.70	3.43	3.38	3.39
GSN	3.50	2.88	2.12	2.75	2.81
sSNR [dB]					
narrow-band	1.62	1.64	8.23	6.77	4.56
GBRBM	-4.02	-2.75	1.12	10.07	1.10
CGBRBM	3.82	3.94	10.53	9.95	7.06
HMM-GMM	3.95	4.14	10.33	9.81	7.06
HMM-SPN	3.52	3.86	9.73	9.67	6.95
HCAE	3.82	3.95	10.45	9.97	7.05
GSN	7.95	8.43	13.72	13.35	10.85

Table 1: Log spectral distortion (LSD) and segmental signal to noise ratio (sSNR) for SD frame-wise and HMM-based models. The narrow-band baseline is included. Bold numbers denote best results for each speaker.

Model	s1	s2	s18	s20	avg.
LSD [dB]					
narrow-band	6.98	7.58	6.66	6.48	6.96
GBRBM	6.67	7.29	6.50	6.15	6.65
CGBRBM	4.18	4.87	4.26	6.15	4.87
HMM-GMM [6]	3.62	4.46	3.82	3.60	3.88
HMM-SPN [6]	3.42	3.85	3.05	3.36	3.32
HCAE	3.81	4.05	3.95	3.36	3.79
GSN	3.88	3.77	3.17	3.55	3.59
sSNR [dB]					
narrow-band	-0.32	0.57	3.91	5.19	2.34
GBRBM	3.92	4.13	10.73	10.07	7.21
CGBRBM	3.89	4.08	10.59	10.00	7.14
HMM-GMM	3.69	3.88	10.23	9.81	6.90
HMM-SPN	3.54	3.80	9.94	9.65	6.73
HCAE	3.85	3.97	10.55	9.96	7.08
GSN	7.97	8.27	13.73	13.31	10.81

Table 2: Log spectral distortion (LSD) and segmental signal to noise ratio (sSNR) for SI frame-wise and HMM-based models. The narrow-band baseline is included. Bold numbers denote best results for each speaker.

tions of single frame-wise GSNs (c), HCAEs (d), CGBRBMs (e), GBRBMs (f), HMM-GMMs (g) and HMM-SPNs (h) for a specific telephone band (b) and the original full bandwidth signal (a).

The frame-wise GSN model is able to reproduce the missing high frequency components in a better way than HCAE and CGBRBM. The HCAE produces a strongly smoothed spectrogram of the high frequency bands. The GBRBM and CGBRBM fails to produce a meaningful reconstruction of the high frequency band. The HMM-SPN approach is obtaining a more natural high frequency reconstruction compared to HMM-GMMs. Interestingly, in this example the frame-wise GSN recovers the most similar spectrogram without explicit temporal modeling. According to informal listening tests, the visual impression corresponds to the listening experience; the signals delivered by HMM-SPNs and GSNs clearly enhance the high-frequency

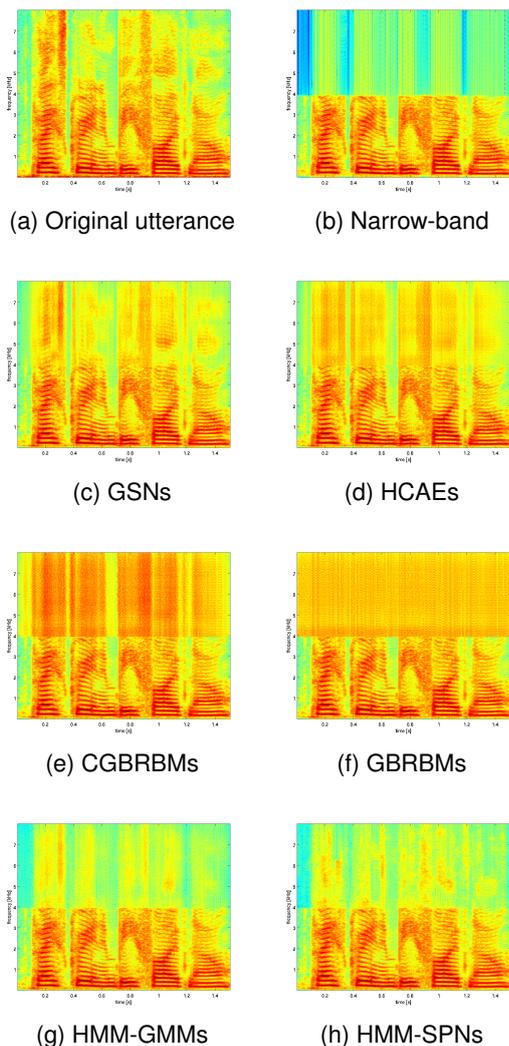


Figure 2: Log-spectrogram of the utterance “Place green in b 5 now”, spoken by s20 recovered by various frame-wise SD deep representation models and hybrid HMM models: (a) original full bandwidth signal; narrow bandwidth signal (b); GSNs (c), HCAEs (d), CGBRBMs (e), GBRBMs (f), HMM-GMMs (g) and HMM-SPNs (h).

content and sound more natural than the signals delivered by HMM-GMMs, HCAEs, and RBM variants. The SPN and GSN models produce a more realistic extension for fricative and plosive sounds.

The average WB-PESQ is above 4.36 for the SPN-HMM, HCAE and GSN models, with no statistical significant differences. The narrow-band signal achieved an average WB-PESQ of 4.35 and the full-band signal 4.5. Therefore, we do not report detailed WB-PESQ scores as the improvements in the high frequencies are not well covered in the score and differences are neglectable.

4. Conclusion

We presented a comparison of representation models applied to ABE. The best model, i.e. GSNs, achieved a relative im-

provement of 3.90dB and 4.08dB in segmental SNR on both SD and SI tasks compared to the HMM-SPN baseline [6]. In general GSNs and auto-encoder variants seem to be the method of choice when used for ABE due to adequate reconstruction results compared to baseline hybrid HMM systems. When analyzing generative representation models for ABE in detail, we also showed that deep generative models outperform single layer networks. Higher layers contribute to the modeling process of the lower layers and therefore lead to better reconstructions in the end. Future work includes the realization of GSNs on hardware to enable real-time ABE and formal listening tests.

5. References

- [1] Y. Bengio, “Learning deep architectures for AI,” *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
- [2] D. Yu and L. Deng, “Deep learning and its applications to signal and information processing,” *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145–154, 2011.
- [3] G. Hinton, L. Deng, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [4] H. Poon and P. Domingos, “Sum-product networks: A new deep architecture,” in *Uncertainty in Artificial Intelligence (UAI)*, 2011, pp. 337–346.
- [5] A. Dennis and D. Ventura, “Learning the architecture of sum-product networks using clustering on variables,” in *Neural Information Processing Systems (NIPS)*, 2012, pp. 2042–2050.
- [6] R. Peharz, G. Kapeller, P. Mowlae, and F. Pernkopf, “Modeling speech with sum-product networks: Application to bandwidth extension,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.
- [7] R. Peharz, B. Geiger, and F. Pernkopf, “Greedy part-wise learning of sum-product networks,” in *European Conference on Machine Learning (ECML)*, 2013, pp. 612–627.
- [8] P. Jax and P. Vary, “On artificial bandwidth extension of telephone speech,” *Signal Processing*, vol. 83, pp. 1707–1719, 2003.
- [9] K. Cho, A. Ilin, and T. Raiko, “Improved learning of Gaussian-Bernoulli restricted Boltzmann machines,” in *International Conference in Artificial Neural Networks ICANN*, vol. 6791 LNCS, 2011, pp. 10–17.
- [10] G. Taylor and G. Hinton, “Factored conditional restricted Boltzmann machines for modeling motion style,” in *International Conference on Machine Learning (ICML)*, 2009.
- [11] S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, and X. Glorot, “Higher order contractive auto-encoder,” in *European Conference on Machine Learning (ECML)*, 2011.
- [12] Y. Bengio, E. Thibodeau-Laufer, and J. Yosinski, “Deep generative stochastic networks trainable by backprop,” in *International Conference on Machine Learning (ICML)*, 2014.
- [13] M. Zöhrer and F. Pernkopf, “General stochastic networks for classification,” in *Neural Information Processing Systems (NIPS)*, 2014.

- [14] Y. Bengio, L. Yao, G. Alain, and P. Vincent, "Generalized denoising auto-encoders as generative models," in *Neural Information Processing Systems (NIPS)*, 2013, pp. 899–907.
- [15] D. Ackley, G. Hinton, and T. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science*, vol. 9, no. 2, pp. 147–169, 1985.
- [16] P. Smolensky, *Information processing in dynamical systems: Foundations of harmony theory*. MIT Press, 1986, vol. 1, pp. 194–281.
- [17] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *International Conference on Machine Learning (ICML)*, 2008, pp. 1064–1071.
- [18] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks." *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [19] G. E. Dahl, M. Ranzato, A. Mohamed, and G. Hinton, "Phone recognition with the mean-covariance restricted Boltzmann machine," in *Neural Information Processing Systems (NIPS)*, 2010.
- [20] R. Sarikaya, G. Hinton, and A. Deoras, "Application of deep belief networks for natural language understanding," *IEEE Transactions on Audio Speech and Language Processing*, vol. 22, no. 4, pp. 778–784, 2014.
- [21] Y. Bengio and P. Lamblin, "Greedy layer-wise training of deep networks," in *Neural Information Processing Systems (NIPS)*, 2007.
- [22] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *International Conference on Machine Learning (ICML)*, 2008, pp. 1096–1103.
- [23] S. Rifai and X. Muller, "Contractive auto-encoders: Explicit invariance during feature extraction," in *International Conference on Machine Learning (ICML)*, 2011, pp. 833–840.
- [24] G. Alain, Y. Bengio, and S. Rifai, "Regularized auto-encoders estimate local statistics," *arXiv preprint arXiv:1211.4246*, pp. 1–17, 2012.
- [25] R. Gens. and P. Domingos, "Learning the structure of sum-product networks," in *International Conference on Machine Learning (ICML)*, 2013, pp. 873–880.
- [26] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition." *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.
- [27] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transaction on Communication*, vol. 28, no. 1, pp. 84–95, 1980.
- [28] S. Osindero and G. Hinton, "Modeling image patches with a directed hierarchy of markov random fields," in *Neural Information Processing Systems (NIPS)*, 2007, pp. 1–8.
- [29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 249–256.
- [30] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
- [31] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed., 2013.
- [32] "ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2000.
- [33] S. Möller, E. Kelaidi, F. Köster, N. Côté, P. Bauer, T. Fingscheidt, T. Schlien, H. Pulakka, and P. Alku, "Speech quality prediction for artificial bandwidth extension algorithms," in *Interspeech*, 2013, pp. 3439–3443.
- [34] S. Tschitschek and F. Pernkopf, "On Reduced Precision Bayesian Network Classifiers." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 37, no. 4, pp. 774–785.