On Representation Learning for Artificial Bandwidth Extension

Matthias Zöhrer[1]*, Robert Peharz*[2] *and Franz Pernkopf*[1]

Signal Processing and Speech Communication Lab, Graz University of Technology[1],
BEE-PRI BioTechMed-Graz, iDN – Inst. of Physiology, Medical University of Graz[2],
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 a relative improvement of 18.65% and 8.73% in log-spectral distortion on a speaker dependent (SD) and speaker independent (SI) scenario compared to the second best candidate – 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, 7, 8, 9], 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, 10]. Motivated by this success, SPNs have been recently used to model speech and were applied to the task of artificial bandwidth extension (ABE) [8] 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 [8], SPNs were only compared to classical signal processing methods, namely an HMM system based on a vocal-tract filter model using linear predictions [11] 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) [12], conditional restricted Boltzmann machines (CGBRBMs) [13], and higher order contractive autoencoders (HCAEs) [14]. Furthermore, we evaluated the recently proposed general stochastic networks (GSNs) [15, 16] generalizing denoising autoencoders are particularly interesting. In [17] 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 [**?**], GSNs showed convincing reconstruction results on MNIST, which, similar as argued in [8] for SPNs, motivates their usage for ABE. In [16], 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 consistently outperform frame-wise GBRBMs, CGBRBMs and HCAEs. GSNs exceed the performance of HMM-SPNs, achieving an average relative improvement in log-spectral distortion (LSD) on both speaker dependent (SD) and speaker independent (SI) tasks, of 18.65% and 8.73% 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) [18, 19, 20, 12, 21]. 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 corresponds to modifying this energy function to obtain desirable properties. This can be achieved 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) [12], or to learn temporal relations, i.e. Conditional RBMs (CGBRBMs) [13]. They also form the basis of more complex and powerful neural networks, i.e. deep belief networks [21] if stacked and trained in a greedy manner. Therefore they are widely used in many applications [22, 23].

The second class are *(deep) autoencoders* (AE) [24, 25, 26, 14, 27, 17]. AEs map an input to a hidden representation and transfer the latent representation back into a reconstruction the input. AEs are mainly used as filters, feature extractors [26] or data generators [17] optimized via back-propagation and capable of learning a data distribution. An interesting variant is the higher order contractive autoencoder (HCAE) [14]. HCAEs regularize the norm of the Jacobian (*analytically*) and the Hessian (*stochastically*) to obtain a more robust representation of

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.

the underlying data.

The third class of representational models are *sum-product 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, 7, 10]. 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* [15, 16], extending the class of AEs to joint layer training. GSNs are multi-layer network architectures using backprop-able stochastic neurons. These neurons are modeled with the help of *deterministic* functions of random variables $f_\theta^i \supseteq \{\hat{f}_\theta^i, \check{f}_\theta^i\}$ expressing 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 $\hat{f}_\theta^i$ 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 $\hat{f}_\theta^i = \eta_{out}^i + g(\eta_{in}^i + \hat{a}^i)$ 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 $\theta \supseteq \{W^0, ..., W^d, b^0, ..., b^d\}$ using multiple *deterministic* functions of random variables $f_\theta \in \{\hat{f}_\theta^0, ..., \hat{f}_\theta^d, \check{f}_\theta^0, ...\check{f}_\theta^d\}$, where d is the number of hidden layers.

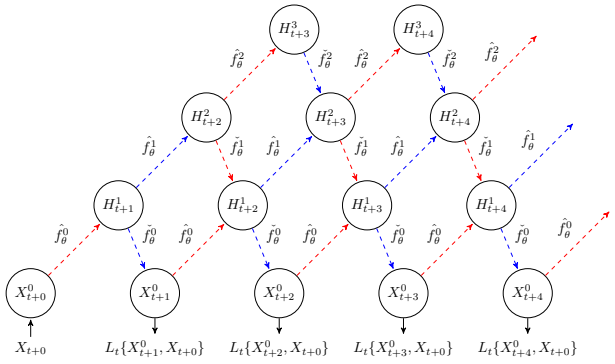

Figure 1: *Unfolded multi-layer GSN with backprop-able stochastic units [16].*

Figure 1 visualizes an unfolded multi-layer GSN described in detail in [16]. 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 *walk-back* training and their hierarchical structure which is jointly optimized they form a powerful model class, especially when used for input reconstruction [15].

3. ABE using Representational Models

3.1. Experimental Setup

We follow the same experimental setup as in [8], using speaker 1, 2, 18 and 20 from the GRID corpus [28] as test speakers. We simulate narrow-band telephone speech as in [8] 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 8 ms.

For SPNs we used the same model as in [8], 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 [29]. 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. Thus training and test data are matched. In contrast to SPNs, we perform a frame-wise processing with the GSNs, HCAEs GBRBMs and CGBRBMs.

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 [30] 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 an uniform distribution [31] 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 [8], using 100 iterations of the Griffin&Lim algorithm [32] to synthesize phase for the reconstructed time-frequency bins. For the low frequency band the phase from the initial telephone band is used. In [11] the phase from the telephone band is taken for the upper-band. This is also a good alternative for computationally constraint systems.

3.2. Evaluation Objective

For objective evaluation, we use the log-spectral distortion in the high-band, similar as in [11, 8]. Using 9^{th} order LPC analy-

sis 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 $\mathbf{a} = (a_0, \ldots, a_9)$ are the LPC coefficients. The high-band LSD 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 $\mathbf{a}_\tau$ 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.

In a detailed analysis we also considered wide-band PESQ [33] (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 [34].

3.3. Results

The LSD of the frame-wise and HMM-models for the SD and SI tasks are shown in Table 1 and 2, respectively. Furthermore, we added the LSD for the narrow-band signal. The GSN clearly outperforms the GBRBM and CGBRBM variants, due to its ability to handle multi-modal input data distributions. When comparing the model to the HMM-SPN and HMM-GMM baselines on both the SI and SD tasks, the GSN was also able to outperform the HMM hybrids for most speakers. Most notably, the HMM-SPN consisting of 64 sub-models connected to one HMM achieved the second best overall result, whereas the GSN achieved the best average (avg.) performance. 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.

Model	s1	s2	s18	s20	avg.
narrow-band	6.98	7.58	6.66	6.48	6.93
GBRBM	5.92	8.82	6.55	4.89	6.56
CGBRBM	3.82	4.4	4.99	4.26	4.37
HMM-GMM [8]	3.18	2.93	2.28	2.82	2.80
HMM-SPN [8]	3.12	2.84	**2.15**	2.59	2.68
HCAE	2.72	3.05	3.36	2.96	3.02
GSN	**1.92**	**2.20**	2.17	**2.42**	**2.18**

Table 1: Log spectral distortion for SD frame-wise and HMM-based models. The narrow-band baseline is included. Bold numbers denote best results for each speaker.

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

Model	s1	s2	s18	s20	avg.
narrow-band	6.98	7.58	6.66	6.48	6.96
GBRBM	4.65	6.27	5.70	4.86	5.34
CGBRBM	3.87	4.40	4.99	4.26	4.38
HMM-GMM [8]	3.62	4.46	3.82	3.60	3.88
HMM-SPN [8]	3.42	3.85	**3.05**	3.36	3.32
HCAE	2.87	3.64	4.35	3.36	3.55
GSN	**2.56**	**3.20**	3.48	**2.91**	**3.03**

Table 2: Log spectral distortion for SI frame-wise and HMM-based models. The narrow-band baseline is included. Bold numbers denote best results for each speaker.

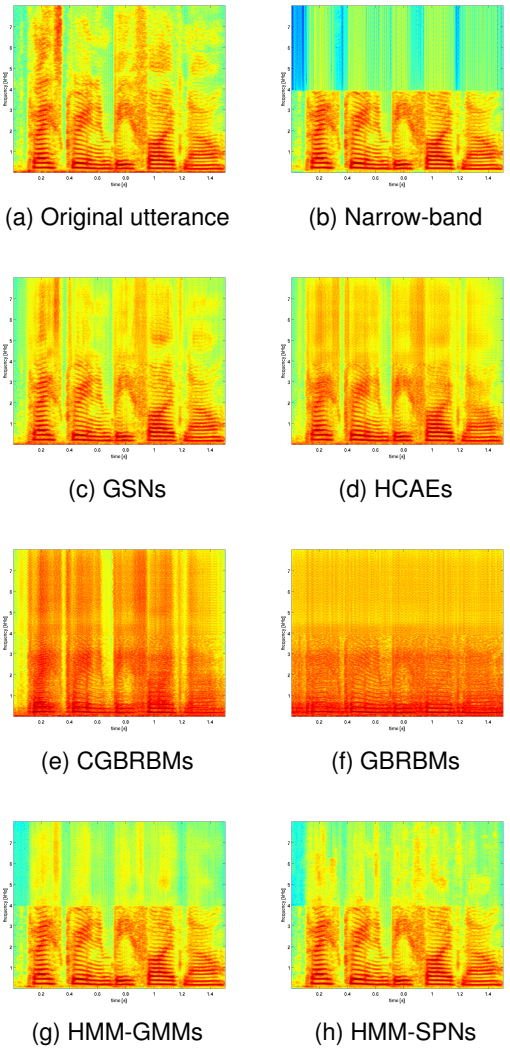

Figure 2: Log-spectrogram of the utterance "Place green in b 5 now", spoken by s20 recovered by various frame-wise SD deep representation models and hybrid HMM models: (a) original full bandwidth signal; narrow bandwidth signal (b); GSNs (c), HCAEs (d), CGBRBMs (e), GBRBMs (f), HMM-GMMs (g) and HMM-SPNs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

BRBM fails to produce a meaningful reconstruction of the high frequency band. This has also impact on the telephone band during recovery of the time-domain signal. The reconstruction by HMM-GMM is blurry and does not recover the harmonic structure of the original signal well, but partly recovers high-frequency content related to consonants. 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 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 stayed 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 improvement of 18.65% and 8.73% in log-spectral distortion on both SD and SI tasks compared to the HMM-SPN baseline [8]. 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. Gens and P. Domingos, "Discriminative learning of sum-product networks," in *Neural Information Processing Systems (NIPS)*, 2012, pp. 3248–3256.

[7] R. Gens. and P. Domingos, "Learning the structure of sum-product networks," in *International Conference on Machine learning (ICML)*, 2013, pp. 873–880.

[8] R. Peharz, G. Kapeller, P. Mowlaee, and F. Pernkopf, "Modeling speech with sum-product networks: Application to bandwidth extension," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[9] A. Rooshenas and D. Lowd, "Learning sum-product networks with direct and indirect variable interactions," *International Conference on Machine learning (ICML)*, pp. 710–718, 2014.

[10] 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.

[11] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, pp. 1707–1719, 2003.

[12] 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.

[13] G. Taylor and G. Hinton, "Factored conditional restricted Boltzmann machines for modeling motion style," in *International Conference on Machine Learning (ICML)*, 2009.

[14] 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 and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2011.

[15] Y. Bengio, E. Thibodeau-Laufer, and J. Yosinski, "Deep generative stochastic networks trainable by backprop," in *International Conference on Machine Learning (ICML)*, 2014.

[16] M. Zöhrer and F. Pernkopf, "General stochastic networks for classification," in *Neural Information Processing Systems (NIPS)*, 2014.

[17] 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.

[18] D. Ackley, G. Hinton, and T. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science*, vol. 9, no. 2, pp. 147–169, 1985.

[19] P. Smolensky, *Information processing in dynamical systems: Foundations of harmony theory*. MIT Press, 1986, vol. 1, pp. 194–281.

[20] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *International Conference on Machine learning (ICML)*, 2008, pp. 1064–1071.

[21] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks." *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[22] 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.

[23] 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.

[24] Y. Bengio and P. Lamblin, "Greedy layer-wise training of deep networks," in *Neural Information Processing Systems (NIPS)*, 2007.

[25] 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.

[26] S. Rifai and X. Muller, "Contractive auto-encoders: Explicit invariance during feature extraction," in *International Conference on Machine learning (ICML)*, 2011, pp. 833–840.

[27] G. Alain, Y. Bengio, and S. Rifai, "Regularized auto-encoders estimate local statistics," *arXiv preprint arXiv:1211.4246*, pp. 1–17, 2012.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," 2007.

[34] 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, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 3439–3443.