Voice Activity Detection Using Discriminative Restricted Boltzmann Machines

Rogério G. Borin and Magno T. M. Silva
Escola Politécnica, University of São Paulo, Brazil
{rborin, magno}@lps.usp.br

***Abstract*—Voice Activity Detection (VAD) plays an important role in current technological applications, such as wireless communications and speech recognition. In this paper, we address the VAD task through machine learning by using a discriminative restricted Boltzmann machine (DRBM). We extend the conventional DRBM to deal with continuous-valued data and employ feature vectors based either on mel-frequency cepstral coefficients or on filter-bank energies. The resulting detector slightly outperforms the VAD often used as benchmark for detector comparison. Results also indicate that DRBM is able to deal with strongly correlated feature vectors.**

I. Introduction

Voice activity detection has been a topic of intense research in the signal processing community for many years (see e.g. [1]–[15]). In telephony systems, voice activity detectors (VADs) enable a significant reduction in the bandwidth used for speech communications. They are also used in systems for noise reduction, where the noise spectrum is estimated during the absence of speech [1].

Conventional techniques for voice activity detection use measurements and/or statistical models to emphasize differences between presence and absence of speech [2]–[12]. Some of the most used measurements are: energy, parameters of linear predictive coding [2], zero-crossing rate [3], periodicity [4], cepstral features [5], formant configurations [6], and spectral entropy [7]. Three VADs based on some of these techniques should be emphasized: G.729-B [16], G.729-II which is an improved version of the first [17], and the long-term spectral divergence (LTSD) [1]. G.729-B/II were adopted in industry as part of a speech codec and uses energy measurements, zero-crossing rate, and parameters of linear predictive coding. LTSD compares the long-term spectral envelope of the signal with an estimate of the noise spectrum and is often used as a benchmark for other detectors due to its good performance for a wide range of signal-to-noise ratio (SNR) [1].

Recently, VADs based on machine-learning techniques have attracted attention in the literature (see e.g., [13]–[15] and their references). Some approaches use measurements of the conventional techniques as inputs of the classifier. This is the case of the scheme proposed in [13], where a support vector machine (SVM) was fed with measurements of G.729-B to improve the performance of the conventional VAD. In some other approaches, the learning mechanism is fed with measurements that represent the sound, as mel-frequency cepstral coefficients [14]. In this case, besides taking the final decision, the classifier is also responsible for discovering the features that should be used to decide between presence or absence of speech. Independently of the approach, the use of learning mechanisms enables the addition of new information to the detector. In [15], for instance, feature vectors with a wide range of measurements were used as inputs of a deep belief network – deep neural network (DBN-DNN), which allowed this scheme to merge different types of information and obtain a powerful detector.

In machine learning, restricted Boltzmann machines (RBMs) have been successfully used for reproducing discrete probability distributions [18], [19] and also for classification [20]. Due to the efficient algorithm proposed by Hinton in [21], they have played an important role in the training of deep belief networks (DBNs) [22], [23]. Despite their good performance in nonlinear classification problems (e.g., character recognition) [20], RBMs have been little exploited as a stand-alone solution to classification problems when compared to other standard classifiers as neural networks and SVMs. Furthermore, to the best of our knowledged, RBMs were used for classification only with binary data.

In this paper, we propose a VAD based on a discriminative restricted Boltzmann machine (DRBM). We focus on the discriminative training since this modality of RBM training is more suitable to achieve good classification models [20]. Since DRBM works with binary data in its original form, we propose a variant of its model, named Gauss-Bernoulli DRBM, in order to enable continuous-valued data. By means of simulations, we verify that the proposed Gauss-Bernoulli DRBM with a small number of hidden units is able to obtain a performance slightly superior compared to LTSD in terms of area under the receiver operating characteristic (ROC) curve, accuracy, and computational cost. Furthermore, the proposed detector is compared with G.729-B/II. The paper is organized as follows. Section II revisits the the RBM model that includes a classifier layer and proposed the Gauss-Bernoulli DRBM model that deals with continuous-values data. In Section III, we present experimental configuration issues and the simulation results. Finally, Section IV closes the paper with conclusions and perspectives for future works.

II. Gauss-Bernoulli DRBM

An RBM is a stochastic neural network that is able to generate data according to a probability distribution [18]. Fig. 1 sketches an RBM, where circles represent its units. The rectangle on the top represents the layer of hidden units. The rectangles on the bottom represent layers of visible units: on the right, we have the input layer and on the left, the classification layer that contains the corresponding label [20].

ISBN 978-0-9928626-7-1

The units of an RBM are modeled as random variables with the following joint probability distribution

$$P(y, \mathbf{x}, \mathbf{h}) = \frac{\exp(-E(y, \mathbf{x}, \mathbf{h}))}{Z}, \tag{1}$$

where $\mathbf{x} = [x_1, \ldots, x_{n_d}]^T$ and $\mathbf{h} = [h_1, \ldots, h_{n_h}]^T$ are state vectors of the input and hidden variables, respectively, $y \in \{1, \ldots, n_c\}$ is the label (class) corresponding to the input vector, $E(y, \mathbf{x}, \mathbf{h})$ is known as global energy function, and Z is a scalar used to guarantee that the sum of $P(y, \mathbf{x}, \mathbf{h})$ over its domain is equal to one.

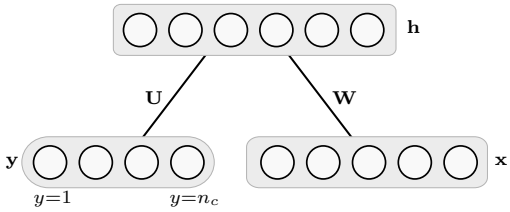

Figure 1. Diagram of an RBM including a classification layer.

Different definitions for $E(y, \mathbf{x}, \mathbf{h})$ lead to different models. In this work, the definition of [20] was changed in order to ensure that the visible variables are conditionally Gaussian, i.e.,

$$E(y, \mathbf{x}, \mathbf{h}) = -\sum_{i=1}^{n_d}\sum_{j=1}^{n_h} h_j w_{ji} \frac{x_i}{\sigma_i^2} - \sum_{j=1}^{n_h} b_j h_j - \sum_{k=1}^{n_c} d_k \delta_{k,y} - \sum_{k=1}^{n_c}\sum_{j=1}^{n_h} h_j u_{jk} \delta_{k,y} + \sum_{i=1}^{n_d} \frac{(x_i - c_i)^2}{2\sigma_i^2}. \tag{2}$$

This definition for $E(y, \mathbf{x}, \mathbf{h})$ was inspired by the one proposed in [24], where RBMs were modeled without a classification layer. In (2), $\delta_{r,s}$ represents the Kronecker delta and w_{ji}, b_j, c_i, σ_i^2, d_k, u_{jk}, with $i{=}1, \ldots, n_d$, $j{=}1, \ldots, n_h$, $k{=}1, \ldots, n_c$, are parameters of the model. We consider that the hidden units can assume individually values in the set $\{0, 1\}$. Given y and $\mathbf{x}$, it can be shown that the vector of hidden variables is jointly independent and its entries present a Bernoulli distribution with success probability

$$P(h_j{=}1|y, \mathbf{x}) = \varphi\left(b_j + u_{jy} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}\right), \tag{3}$$

where $\varphi(z) = 1/(1 + e^{-z})$ is the sigmoid function.

RBMs are commonly trained by using the contrastive divergence (CD) algorithm proposed in [21] to minimize the following generative loss function

$$\mathcal{L}_{gen} = -\sum_{t=1}^{n_t} \log P(y^{(t)}, \mathbf{x}^{(t)}), \tag{4}$$

where n_t is the number of training samples and $(\mathbf{x}^{(t)}, y^{(t)})$ represents the t^{th} training sample, constituted by the input $\mathbf{x}^{(t)}$ and its respective label (class) $y^{(t)}$. The efficient search for the minima of this function envolves the calculus of its gradient with respect to the parameters of the model, which is intractable. This problem was solved by considering certain approximations, which led to the CD algorithm [21].

The model of a DRBM is identical to that of Figure 1. The difference is that a DRBM is trained to minimize the following discriminative loss function

$$\mathcal{L}_{disc} = -\sum_{t=1}^{n_t} \log P(y^{(t)}|\mathbf{x}^{(t)}). \tag{5}$$

Using the definition of (2), we can show that

$$P(y|\mathbf{x}) = \frac{\exp\left(d_y + \sum_{j=1}^{n_h} \zeta\left(b_j + u_{jy} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}\right)\right)}{\sum_{y^*=1}^{n_c} \exp\left(d_{y^*} + \sum_{j=1}^{n_h} \zeta\left(b_j + u_{jy^*} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}\right)\right)}, \tag{6}$$

with $\zeta(z) = \ln(1 + e^z)$. This equation has the same form to that obtained for binary input variables. As observed in [20], $P(y|\mathbf{x})$ takes time $\mathcal{O}(n_h n_d + n_h n_c)$ to be computed. Furthermore, the gradient of $P(y|\mathbf{x})$ can be exactly computed in an efficient manner. Therefore, the gradient of $\mathcal{L}_{disc}$ can also be exactly computed. Using the stochastic gradient method, we obtain the update rules for the DRBM parameters, summarized in Table I. This table does not contain rules for updating neither c_i nor σ_i^2, since c_i are not relevant for the discriminative training and the learning of σ_i^2 can be avoided by making a normalization in the variance of the input data.

In practical terms, DRBMs usually achieve good classification results with small dimension models when compared to RBMs. Furthermore, since the gradient of the discriminative loss function is exact, the training algorithm enables higher learning rates with no divergence.

III. Experimental analysis

All the experiments shown in this section were performed by using MATLAB 7.11, running in *Windows 7*. The used processor was an *Intel Xeon* with 6 physical kernels operating over 2,4 GHz and 32 GB of RAM.

Table I
UPDATE RULES FOR DRBM PARAMETERS.

Definitions:

$$\Delta b_j = P(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) - \sum_{y^*=1}^{n_c} P(y^*|\mathbf{x}^{(t)}) P(h_j{=}1|y^*, \mathbf{x}^{(t)})$$

$$\Delta d_k = \delta_{k,y^{(t)}} - P(y{=}k|\mathbf{x}^{(t)})$$

$$\Delta w_{ji} = \Delta b_j \left(\frac{x_i^{(t)}}{\sigma_i^2}\right)$$

$$\Delta u_{jk} = \Delta d_k P(h_j{=}1|y{=}k, \mathbf{x}^{(t)})$$

Note: Functions $P(h_j{=}1|y, \mathbf{x})$ and $P(y|\mathbf{x})$ employed in above definitions are presented in (3) and (6), respectively.

Update rules: According to the stochastic gradient method, the update of ϕ can be performed as $\phi \leftarrow \phi + \lambda \Delta\phi$, where λ is the learning rate of the algorithm.

Test *corpus*: We used a modified version of the NOIZEUS speech *corpus* [25], which was chosen for three reasons: firstly, this *corpus* is constituted by 30 phrases making a total of 80 segundos, which makes its manual labeling feasible; secondly, the phrases in the *corpus* contain all the phonemes of English; and, finally, it is available for free.

The audio in each file contained in the *corpus* was manually labeled in order to indicate the ranges with speech, considering voiced and unvoiced sounds. According to this

ISBN 978-0-9928626-7-1 © EURASIP 2017

labeling, the percentage of voice activity of each file varies in the interval $[64.9\%,\ 91.2\%]$, which represents an average of $83{,}4\%$ of presence of speech. Therefore, this information base is unbalanced since it presents more positive examples (presence of speech) than negative examples (absence of speech). Due to this unbalance, the detectors G.729-B/II and LTSD could not be properly evaluated in the absence of speech. On the other hand, for machine-learning-based detectors, the training with this unbalanced data would lead to biased detectors. Thus, we modified the audio files with no noise, by introducing 0.8 s of silence before and after the original audio. Then, car noise obtained from AURORA-2 database were added in order to achieve an SNR in the interval $[5\ \text{dB},\ 20\ \text{dB}]$ in steps of 5 dB. For this purpose, we used the same procedure to generate noisy files from the original *corpus* [25]. This balancing ensures that the detectors G.729-B/II and LTSD work properly since they use the beginning of the modified files to estimate the noise features. We should notice that 70% of the modified files were randomly chosen for training and the other 30% for test.

Feature extraction: Two configurations of feature vectors stood out for VAD using DRBM. The first is based on mel-frequency cepstral coefficients (MFCCs) and the second, on the filter-bank energies (FBEs), a byproduct of the MFCCs computation. To compute MFCCs and FBEs, the audio of each file was filtered by a pre-emphasis filter (with coefficient equal to 0.97) and the resulting signal was segmented in frames with overlap of 25 ms and shift of 10 ms between frames. From each frame, we extracted the thirteen first MFCCs and the energies (corresponding to the FBEs) of the 23 channels used in filter banks. The mean-square value of the samples of a frame (frame energy – FE) was also included in the feature vector. Details of the considered configurations arw shown in Table II.

Table II
DETAILS OF THE FEATURE VECTORS.

Config.	Contents of the vector	Dimension
C1	13 normalized MFCCs (zero mean and unit variance) + log(FE), in conjunction with their first and second time derivatives.	42
C2	23 normalized FBEs + log(FE) (unity magnitude) in conjunction with their first and second time derivatives.	72

Training: Due to preliminary tests with different number of hidden units (n_h) and learning rates (λ), we chose $n_h = 30$ and $\lambda = 0.005$ for training the DRBM for all SNR values. The training samples, composed by the feature vectors and the corresponding manual labels, were divided in lots of 70 samples for the gradient computation.

Performance evaluation: To evaluate the performance of the detectors, we consider the area under the ROC curve (sensibility *versus* specificity), which is usual in telecommunications field. The VADs G.729-B/II do not have a configurable detection threshold and therefore, only one point of ROC curve can be obtained. This point in conjunction with the theoretical points $(0, 0)$ and $(1, 1)$ enable the calculus of the area. Since the performance of these VADs could be underestimated with this method, we also used the balanced accuracy (average of sensibility and specificity). The performance of LTSD was evaluated by running this VAD with different detection thresholds for the files of the *corpus*. For the detectors based on DRBM, the threshold was chosen by evaluating the probability (Eq. (6)) that a given sample is speech. For both LTSD and DRBM, the accuracies shown in the sequel were obtained with the best detection thresholds (obtained by means of numerous simulations). Besides these performance indicators, we also compare the detectors in terms of computational cost. For this purpose, we use the work rate, defined as the ratio between the processed audio time and the time necessary to complete this task. This measurement indicates how many seconds of audio are precessed in 1 second of processor usage.

Results: Fig. 2 shows the area under ROC curve for the detectors evaluated in different SNR conditions. The DRBMs using configurations C1 and C2 (Table II) are identified as DRBM-C1 and DRBM-C2, respectively. We can observe that the smaller the SNR the worse the performance of G.729-B. On the other hand, G.729-II presents better results, but it is outperformed by the LTSD detector. Finally, DRBM-based VADs slightly outperform LTSD for most values of SNR.

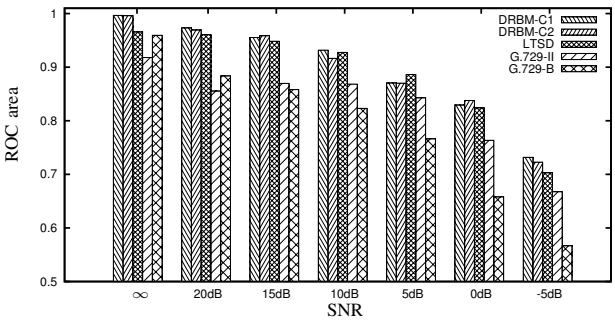

Figure 2. Area under ROC curve for different values of SNR.

Table III shows the accuracy for the evaluated VADs. These results are qualitatively similar to the previous ones. Along the SNR values, we can observe that DRBM-based VADs present better performance in the mean when compared to the others.

We should notice that the results obtained by DRBM-based VADs with the different configurations (C1 and C2) are relatively close. Configuration C2 is based on the FBEs, which lead to a correlated feature vector. On the other hand, Configuration C1 is based on MFCCs, which are computed by applying the discrete cosine transform to the logarithm of the FBEs, producing an uncorrelated vector (which is usually considered beneficial for posterior processing). Therefore, this similar performance indicates that DRBMs can lead to proper results even when the the feature vector is strongly correlated. The same conclusion was pointed out in [23] for DBN-DNNs. The outputs of the detectors for different values of SNR are shown in Fig. 3. We can observe the performance of that all of them gets worse with the decrease of SNR.

To close this section, the computational costs of the detectors were empirically obtained. The work rates of the

ISBN 978-0-9928626-7-1 © EURASIP 2017

Table III
ACCURACY OF THE DETECTORS FOR DIFFERENT SNRS.

SNR(dB)	DRBM-C1	DRBM-C2	LTSD	G729-II	G729-B
∞	**97.76%**	97.72%	93.73%	91.80%	95.95%
20	**93.72%**	91.66%	91.07%	85.55%	88.38%
15	90.93%	**91.40%**	89.70%	86.98%	85.82%
10	**88.27%**	86.81%	86.96%	86.85%	82.32%
5	82.60%	83.14%	82.83%	**84.29%**	76.63%
0	77.31%	**78.52%**	76.88%	76.35%	65.80%
−5	**67.94%**	67.37%	66.62%	66.76%	56.68%
Mean	**85.50%**	85.23%	83.97%	82.65%	78.80%

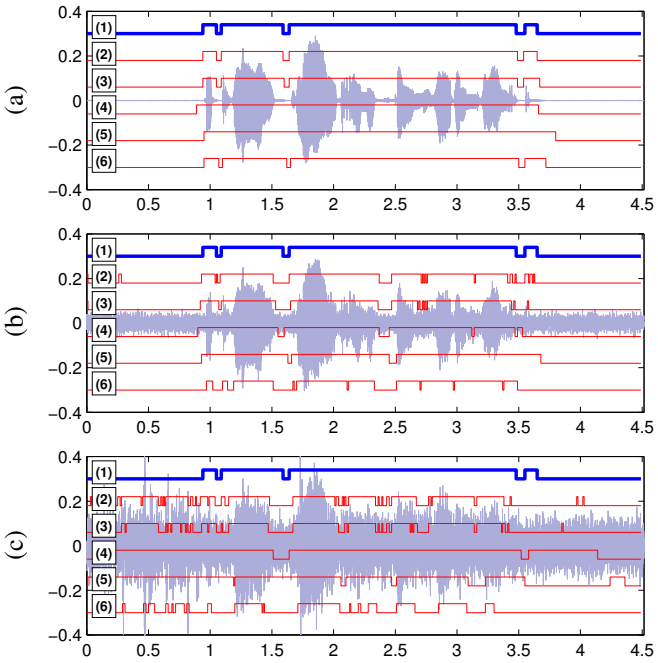

Figure 3. Outputs of the detectores for different SNRs: (a) No noise, (b) 10dB, (c) 0dB. For each SNR, over the audio we can see (1) the manual labeling and the outputs of the detectors: (2) DRBM-C1, (3) DRBM-C2, (4) LTSD, (5) G.729-II e (6) G.729-B.

considered VADs are shown in Table IV. Since this measure depends on the details of the algorithm implementation, the values of Table IV allow us to compare only the orders of magnitude of the costs. Taking this into account, DRBMs present computational costs of the same order of magnitude to that of LTSD. On the other hand, G.729-B/II present higher computational costs since they are implemented in conjunction with speech codification.

IV. CONCLUSION

In this paper, we proposed a Gauss-Bernoulli DRBM and used it for voice activity detection. Simulations were performed on a wide range of signal-to-noise ratios and considering two feature vectors: one based on MFCCs and another on FBEs. From these experiments, we could verify that DRBM-based VADs slightly outperformed the LTSD detector, considered as benchmark for detectors, and considerably outperformed G.729-B and G.729-II, VADs used in industry. Moreover, these behaviors were obtained with a DRBM of small dimensions (30 hidden units) that present a computational cost comparable to that of LTSD. Additionally, simulations algo show that DRBM is able to properly deal with correlated inputs. In a future work, we intend to compare the DRBM-based VAD with another machine-learning VAD (e.g., SVM) and consider different types of noise.

Table IV
WORK RATE OF THE DETECTORS.

DRBM-C1	DRBM-C2	LTSD	G729-II	G729-B
346,4	243,2	174,8	34,7	35,2

REFERENCES

[1] Javier Ramírez et al., "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.

[2] L. Rabiner and M.R. Sambur, "Voiced-unvoiced-silence detection using the itakura lpc distance measure," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process.* IEEE, 1977, vol. 2, pp. 323–326.

[3] Jean-Claude Junqua, Ben Reaves, and Brian Mak, "A study of endpoint detection algorithms in adverse conditions: incidence on a dtw and hmm recognizer," in *2nd European Conf. on Speech Comm. and Technology*, 1991.

[4] R Tucker, "Voice activity detection using a periodicity measure," in *IEE Proc. on Comm., Speech and Vision*. IET, 1992, vol. 139, pp. 377–380.

[5] J.A. Haigh and J.S. Mason, "Robust voice activity detection using cepstral features," in *Proc. IEEE Conf. on Computer, Comm., Control and Power Engineering*, 1993, vol. 3, pp. 321–324.

[6] John D. Hoyt and Harry Wechsler, "Detection of human speech in structured noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, 1994, vol. 2, pp. II–237.

[7] Philippe Renevey and Andrzej Drygajlo, "Entropy based voice activity detection in very noisy conditions," *Proc. Eurospeech*, vol. 5, no. 5.5, pp. 1–4, 2001.

[8] Elias Nemer, Rafik Goubran, and Samy Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 217–231, 2001.

[9] Ke Li, M.N.S. Swamy, and M. Omair Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 965–974, 2005.

[10] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[11] Saeed Gazor and Wei Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498–505, 2003.

[12] Joon-Hyuk Chang, Nam Soo Kim, and Sanjit K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, 2006.

[13] Dong Enqing, Liu Guizhong, Zhou Yatong, and Zhang Xiaodi, "Applying support vector machines to voice activity detection," in *6th International Conf. on Signal Processing*. IEEE, 2002, vol. 2, pp. 1124–1127.

[14] Y.X. Zou, W.Q. Zheng, Wei Shi, and Hong Liu, "Improved voice activity detection based on support vector machine with high separable speech feature vectors," in *19th Int. Conf. on Digital Signal Processing*. IEEE, 2014, pp. 763–767.

[15] Xiao-Lei Zhang and Ji Wu, "Deep belief networks based voice activity detection," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 21, no. 4, pp. 697–710, 2013.

[16] Adit Benyassine et al., "Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 64–73, 1997.

[17] "Appendix II – ITU-T G.729 annex B enhancements in voice-over-IP applications – Option 1," Aug. 2005.

[18] Paul Smolensky, "Parallel distributed processing: Explorations in the microstructure of cognition," in *Information Processing in Dynamical Systems: Foundations of Harmony Theory*, David E. Rumelhart, James L. McClelland, and Corp. PDP Research Group, Eds., vol. 1, chapter 6, pp. 194–281. MIT Press, Cambridge, MA, USA, 1986.

ISBN 978-0-9928626-7-1 © EURASIP 2017

[19] Nicolas Le Roux and Yoshua Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural computation*, vol. 20, no. 6, pp. 1631–1649, 2008.
[20] Hugo Larochelle and Yoshua Bengio, "Classification using discriminative restricted boltzmann machines," in *Proc. 25th Int. Conf. on Machine learning*, 2008, pp. 536–543.
[21] Geoffrey E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
[22] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
[23] Geoffrey E. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.
[24] KyungHyun Cho, Alexander Ilin, and Tapani Raiko, "Improved learning of gaussian-bernoulli restricted boltzmann machines," in *Proc. 21st Int. Conf. on Artificial Neural Networks and Machine Learning*, vol. 6791 LNCS, pp. 10–17. Springer, 2011.
[25] Yi Hu and Philipos C Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

ISBN 978-0-9928626-7-1