FACTORIAL HIDDEN RESTRICTED BOLTZMANN MACHINES FOR NOISE ROBUST SPEECH RECOGNITION

Steven J. Rennie[⋆] *Petr Fousek*[†] *Pierre L. Dognin*[⋆]

[⋆] IBM T.J. Watson Research Center, Yorktown Heights, N.Y., U.S.A.
[†] IBM Czech Republic, Prague 4, Czech Republic

ABSTRACT

We present the Factorial Hidden Restricted Boltzmann Machine (FHRBM) for robust speech recognition. Speech and noise are modeled as independent RBMs, and the interaction between them is explicitly modeled to capture how speech and noise combine to generate observed noisy speech features. In contrast with RBMs, where the bottom layer of random variables is observed, inference in the FHRBM is intractable, scaling exponentially with the number of hidden units. We introduce variational algorithms for efficient approximate inference that scale linearly with the number of hidden units. Compared to traditional factorial models of noisy speech, which are based on GMMs, the FHRBM has the advantage that the representations of both speech and noise are highly distributed, allowing the model to learn a parts-based representation of noisy speech data that can generalize better to previously unseen noise compositions. Preliminary results suggest that the approach is promising.

Index Terms— Robust Speech Recognition, Source Separation, Deep Belief Networks, Restricted Boltzmann Machines, Variational Methods.

1. INTRODUCTION

Restricted Boltzmann machines (RBMs) have recently been applied to several well established problems in machine learning and signal processing, with great success. In Automatic Speech Recognition (ASR), Deep Belief Networks (DBNs) of RBMs have been applied to large vocabulary speech recognition (LVCSR) and phone recognition, outperforming (or improving) systems that utilize the *de facto* Gaussian Mixture Model (GMM) of speech acoustics. DBNs of RBMs represent phenomena using a *distributed* state representation, which has extraordinary modeling power, but yet facilitates efficient inference. As such, RBMs [1] are a general tool for modeling, and can extract and represent highly non-linear and complex phenomena. However, when the relationships between real-world variables are well understood, this knowledge can and should be leveraged.

[1]RBMs will be used to refer to DBNs of RBMs often from this point forward, for simplicity.

Model-based approaches to robust ASR that utilize explicit models of noise, channel distortion, and their interaction with speech are a well established and continually evolving research paradigm in robust ASR. Many interesting and effective approximate modeling and inference techniques have been developed to represent these acoustic entities, and the reasonably well understood but complicated interactions between them [1, 2, 3, 4]. In this paper, we consider the use of RBMs as models of (clean) speech and noise, and their integration with explicit models of how speech and noise interact to define Factorial Hidden Restricted Boltzmann Machines (FHRBMs) for robust speech recognition.

2. RESTRICTED BOLTZMANN MACHINES

An RBM is a Markov Random Field (MRF) with one layer of hidden random variables $\mathbf{h}$, and one layer of visible random variables $\mathbf{v}$ [5, 6]. RBMs are distinguished by the characteristic that the hidden variables are not connected to one another. An RBM with Gaussian visible variables and Bernoulli hidden variables has log probability:

$$\log p(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{V}\frac{(v_i-b_i)^2}{2\sigma_i^2} + \sum_{j=1}^{H} a_j h_j + \sum_{i=1}^{V}\sum_{j=1}^{H}\omega_{ij}v_i h_j - Z \qquad (1)$$

where V and H are the number of visible and hidden random variables, respectively. Here Z is a normalization constant (the log partition function), and b_i, σ_i^2, w_{ij}, and a_j $\forall i,j$ are the parameters of the model.

The posterior probability that a given hidden unit is on is:

$$\begin{aligned} p(h_j=1|\mathbf{v}) &= \frac{\sum_{\mathbf{h}\neq h_j}\exp(\log p(\mathbf{v},\mathbf{h}))}{\sum_{\mathbf{h}}\exp(\log p(\mathbf{v},\mathbf{h}))} \\ &= sig(a_j + \sum_{i=1}^{V}\omega_{ij}v_i) \end{aligned} \qquad (2)$$

where $sig(x) = \frac{1}{1+\exp(-x)}$. This corresponds to the activity of a node in a feed-forward neural net.

The conditional probability of v_i is:

$$p(v_i|\mathbf{h}) = \frac{\exp(\frac{-(v_i-b_i)^2}{2\sigma_i^2} + \sum_{j=1}^{H}\omega_{ij}v_i h_j)}{\int_{v_i}\exp(-\frac{(v_i-b_i)^2}{2\sigma_i^2} + \sum_{j=1}^{H}\omega_{ij}h_j)} \quad (3)$$

$$= \mathcal{N}(v_i; b_i + \sigma_i^2\sum_{j=1}^{H}\omega_{ij}h_j; \sigma_i^2), \quad (4)$$

which reveals that $p(\mathbf{v}|\mathbf{h})$ is a diagonal covariance Gaussian, with fixed covariance. Thus an RBM as defined above implements a mixture of Gaussians representation of the data $\mathbf{v}$, and has hidden state $\mathbf{h}$, which is the configuration of a collection of binary random variables. This highly *distributed* representation of the data implements a mixture of 2^H diagonal covariance gaussians, but can be evaluated exactly given $\mathbf{v}$ in time linear in the number of hidden units, H, due to its factorial structure. This is clear from the form of the posterior of $\mathbf{h}$, which *factors*, as shown in (2).

3. FACTORIAL HIDDEN RBMS

We take a model-based approach to noise robust speech recognition, and explicitly represent the probability distribution function (PDF) of speech and noise with the RBMs described above. In this model, both the noise features, $\mathbf{v}^n$, and the speech features, $\mathbf{v}^x$, are unobserved, and must be estimated based on observed mixed data $\mathbf{y}$. The relationship between the speech, noise, and mixture, is captured using an *interaction model*, $p(\mathbf{y}|\mathbf{v}^x, \mathbf{v}^n)$.

Exact inference in FHRBMs scales exponentially with the total number of hidden units in the factorial hidden RBM, $H_n + H_x$, since the "visible" units of the speech and noise RBMs are not actually observed, and the interaction model, $p(\mathbf{y}|\mathbf{v}^x, \mathbf{v}^n)$, is generative. Complicating matters further is the fact that the interaction model must typically be approximated on a context-dependent basis to make inference analytically and computationally tractable. For example, the interaction model (8) is highly non-linear, and so is generally approximated uniquely for every combination of speech and noise states. This issue we will return to shortly.

4. EFFICIENT INFERENCE USING VARIATIONAL METHODS

For any given distribution $q(\mathbf{h}, \mathbf{v})$ over the hidden random variables of the speech and noise RBMs ($\mathbf{v} = \{\mathbf{v}^x, \mathbf{v}^n\}, \mathbf{h} = \{\mathbf{h}^x, \mathbf{h}^n\}$), we can define the following lower-bound on the log probability of the observed data $\mathbf{y}$:

$$\log p(\mathbf{y}) = \log\sum_{\mathbf{h},\mathbf{v}} p(\mathbf{h}^x, \mathbf{v}^x)p(\mathbf{h}^n, \mathbf{v}^n)p(\mathbf{y}|\mathbf{v}^x, \mathbf{v}^n) \quad (5)$$

$$\geq \sum_{\mathbf{h},\mathbf{v}} q(\mathbf{h}, \mathbf{v}) \log\frac{p(\mathbf{h}^x, \mathbf{v}^x)p(\mathbf{h}^n, \mathbf{v}^n)p(\mathbf{y}|\mathbf{v})}{q(\mathbf{h}, \mathbf{v})} \quad (6)$$

$$= E_{q(\mathbf{h}^x,\mathbf{v}^x)}[\log p(\mathbf{h}^x, \mathbf{v}^x)] + E_{q(\mathbf{h}^n,\mathbf{v}^n)}[\log p(\mathbf{h}^n, \mathbf{v}^n)] + E_{q(\mathbf{v}^x,\mathbf{v}^n)}[\log p(\mathbf{y}|\mathbf{v})] + H_{q(\mathbf{h},\mathbf{v})} \equiv \mathcal{L} \quad (7)$$

where $E_{q(x)}[f(x)]$ denotes the expected value of $f(x)$ w.r.t the probability distribution $q(x)$, and $H_{q(x)}$ denotes the entropy of $q(x)$. If $q(\mathbf{h}, \mathbf{v}) = p(\mathbf{h}, \mathbf{v}|\mathbf{y})$ the bound is *tight*, but $p(\mathbf{h}, \mathbf{v}|\mathbf{y})$, as discussed previously, is intractable to compute. To make inference tractable, we *factor* the form of q, which makes determining its parameters tractable.

4.1. Mean-field approximation using the log-sum model

We assume the following approximate interaction between speech and noise in the log Mel power spectral domain [2]:

$$p(\mathbf{y}|\mathbf{v}^x, \mathbf{v}^n) = \prod_f \mathcal{N}(y_f; g(v_f), \psi_f^2), \quad (8)$$

$$g(v_f) = \log(\exp(v_f^x) + \exp(v_f^n)) \quad (9)$$

where $v_f = [v_f^x \; v_f^n]^T$, ψ_f models noise in the representation, which ignores phase interactions, and f is a frequency subscript.

To make inference tractable, we assume a posterior distribution q of the form:

$$q(\mathbf{h}^x, \mathbf{v}^x, \mathbf{h}^n, \mathbf{v}^n) = \prod_f q(v_f^x, v_f^n)\prod_{j=1}^{H_x} q(h_j^x)\prod_{k=1}^{H_n} q(h_k^n)$$

$$= \prod_f \mathcal{N}(v_f; \mu_f, \Phi_f)\prod_{s=x,n}\prod_{j=1}^{H_s}(\gamma_{h_j^s})^{h_j^s}(1-\gamma_{h_j^s})^{1-h_j^s} \quad (10)$$

where $\gamma_{h_j^s} = q(h_j^s = 1)$, and $()^{h_j^s}$ denotes exponentiation by binary random variable h_j^s. In this work, we assume that Φ_f is diagonal.

Substituting this parameterization of q into (6), and assuming $p(\mathbf{y}|\mathbf{v}^x, \mathbf{v}^n)$ as given in (8) (the log-sum approximation), we obtain a lower bound $\mathcal{L}$, that can be optimized to identify the parameters of q. Unfortunately $\mathcal{L}$ does not have a simple form in these parameters. To overcome this, we approximate $p(\mathbf{y}|\mathbf{v}^x, \mathbf{v}^n)$ as follows:

$$p(\mathbf{y}|\mathbf{v}^x, \mathbf{v}^n) \approx \prod_f \mathcal{N}(y_f; g(\mu_f) + (v_f - \mu_f)^T d_f, \psi_f^2), \quad (11)$$

where $d_f = [d_{v_f^x} \; d_{v_f^n}]^T = \frac{\partial g}{\partial v_f}\Big|_{v_f=\mu_f}$ is defined based on the current estimates of the speech and noise features (μ_f). Note that, in contrast with the approximations in [2, 3], which compute a Gaussian estimate of the posterior distribution of the speech and noise features for every combination of speech and noise states, here we approximate this posterior as Gaussian, since an FHRBM has $2^{H_x+H_n}$ unique states. This makes

the former approach intractable for models when the speech or noise RBM has a significant number of hidden units.

Differentiating w.r.t. the variational parameters of q, and also w.r.t. g_f, we arrive at the following set of updates, which may be iterated to maximize $\hat{\mathcal{L}}$, our approximation of the lower bound $\mathcal{L}$:

$$\gamma_{h_j^s} = sig(a_j^s + \sum_{f=1}^{V^s} \omega_{fj}^s \mu_{v_f^s}) \tag{12}$$

$$\phi_{v_f^s}^2 = (\sigma_{v_f^s}^{-2} + d_{v_f^s}^2 (\psi_f')^{-2})^{-1} \tag{13}$$

$$\mu_{v_f^s} = \phi_{v_f^s}^2 (\sigma_{v_f^s}^{-2} (b_{v_f^s} + \sigma_{v_f^s}^2 \sum_{j=1}^{H^s} \omega_{fj}^s \gamma_{h_j^s}) + d_{v_f^s} (\psi_f')^{-2} y_f') \tag{14}$$

$$d_{v_f^s} = sig(\mu_{v_f^s} - \mu_{v_f^{\tilde{s}}}) \tag{15}$$

where $y_f' = y_f - g(\mu_f) - d_{v_f^{\tilde{s}}} \mu_{v_f^{\tilde{s}}}$, $(\psi_f')^{-2} = (\psi_f^2 + d_{v_f^{\tilde{s}}}^2 \sigma_{v_f^{\tilde{s}}}^2)^{-1}$, and $\tilde{s} = x$ when $s = n$, and n when $s = x$.

4.1.1. Extension to deep RBMs

An advantage of the mean-field approach to variational inference is that the algorithms can be extended to deep RBMs in a straightforward manner. For example if an additional layer of hidden variables, $l^s = \{l_1^s, l_k^s, \ldots, l_{L^s}^s\}$, is introduced, along with a corresponding set of additional variational parameters, $q(l^s) = \prod_k q(l_k^s) = \prod_k \gamma_{l_k^s}$, it is straightforward to verify that the update for the posterior estimate of a unit in first hidden layer is given by:

$$\gamma_{h_j^s} = sig(a_j^s + \sum_{i=1}^{V^s} \omega_{ij}^s \mu_{v_i^s} + \alpha_j^s + \sum_{j=1}^{L^s} \varpi_{jk}^s \gamma_{l_k^s}) \tag{16}$$

where ϖ_{jk}^s is the weight between hidden units h_j^s and l_k^s, and α_j^s represents additional bias on h_j^s. The activation of hidden unit h_j^s naturally depends on activation of units in the layers immediately above and below. The current estimates of these units under the approximate posterior q are used as surrogate observations during mean-field inference.

5. EXPERIMENTS

5.1. Data, Acoustic Model, and Recognizer

Experiments were conducted on real data recorded to characterize in-car recognition scenarios. The proprietary database described in [7] was used for all experiments. Audio data is US English in-car speech recorded in various noise conditions (0, 30 and 60 mph), and sampled at 16kHz. The training set is composed of 786 hours of speech, with 10k speakers for a total of 800k utterances. The test set contains a total of 206k words in 39k utterances from 128 held-out speakers. There are 47 tasks covering four domains (navigation, command & control, digits & dialing, radio) in 7 US regional accents.

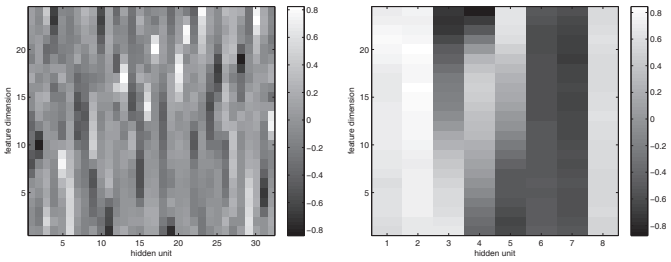

Fig. 1. Learned weight matrices for the bottom layer of one set of clean speech/noise RBMs. The columns correspond to elemental feature vectors, which combine linearly to define the conditional probability distribution of clean speech/noise features, given the current posterior estimate of the hidden units directly above, as shown in (3). Additional hidden layers encode additional constraints on the activity of these feature vectors. The FHRBM combines these representations using an *interaction model*, which describes how speech and noise combine in the modeled feature domain (log Mel depicted) to generate noisy speech.

Our reference acoustic model is a state-of-the-art quantized [8] 10k Gaussian with 865 context-dependent (CD) states. We use a set of 91 phonemes modeled by three-state hidden Markov models (HMM). fMMI uses a secondary acoustic model with 512 Gaussians, with an inner and outer context of 17 and 9 frames, respectively.

5.2. Front-end models

All front-end acoustic models were trained on a small subset of the training data. All speech models were trained on 400k frames of randomized speech from clean conditions ($\geq$25dB); the silence model is trained on 400k frames of unsorted noise from the non-speech parts of utterances. The speech/noise segmentation was based on a forced-alignment. The FHRBMs for speech and noise are both deep RBMs with two RBM layers inside, operating in a 24-dimensional log Mel spectral domain. Their respective topologies are $24-32-8$ and $24-8-3$ ($input-hidden-top$) neurons. Each RBM layer is fully connected; thus there are 1120 trainable parameters for the speech RBM and 259 for the noise RBM. Competitive band-quantized GMM models have similar sizes ($\sim$272 and $\sim$5 states, respectively). In this setup, the state of the front-end is reset at each utterance.

In this paper, the speech and noise RBMs were trained independently via forced-alignments of speech and silence, respectively, using contrastive divergence. Figure 1 depicts learned feature layer weight matrices for clean speech and noise RBM representations, respectively.

5.3. Results

Table 1 depicts the Word Error Rate (WER) and Sentence Error Rate (SER) performance of our embedded recognizer as a function of system configuration and front-end system. All results were obtained with an acoustic model built in a discriminatively trained fMMI feature space. The four baseline configurations (B1-B4) are defined based on their use (or lack thereof) of commercial-grade spectral subtraction (SS) and stochastic feature-space Maximum Likelihood Linear Regression (fMLLR) during decoding. The Static Noise Model (SNM), Dynamic Noise Adaptation (DNA), and DNA with condition detection (DNA-CD) were initialized based on the first 10 frames of each utterance, and utilize a band-quantized (8 level) 272 component Gaussian for the speech model, and a Gaussian model for noise [9]. The FHRBM results were obtained with three layer speech (24-32-8) and noise (24-8-3) RBMs, using the variational algorithm described in this paper. During inference, the noise state posterior estimates from the previous frame were used to initialize those of the current frame in lieu of an explicit dynamical model of noise. Looking at the results, we can see that the FHRBM system outperforms the DNA system, presumably because it is utilizing prior information, can better handle any non-stationary noise, and can explain new combinations of noise that are encountered only at test time, due to its distributed state representation. However, the performance of the FHRBM is surpassed by DNA-CD, which does online Bayesian model averaging to shut off explicit noise modeling when it is not beneficial. Nevertheless, these preliminary results with FHRBMs are remarkable, considering that the FHRBM inference algorithm used here maintains a *single* estimate of SNR per frequency band during inference, whereas the other algorithms compute an SNR estimate for every speech state. Moreover, for the FHRBM system, only its state was initialized on the speech free data at the beginning of each utterance. The other algorithms, in contrast, adapted their parameters on speech-free test data. Both CD and parameter adaptation could be applied to enhance the performance of the FHRBM system.

Algorithm	WER/SER (%)
fMMI (B1)	1.34/3.77
B1 + SNM	1.70/5.06
B1 + DNA	1.27/4.04
B1 + FHRBM	1.20/3.51
B1 + DNA-CD	1.09/3.19
fMMI+SS (B2)	1.18/3.41
B2 + SNM	1.76/5.27
B2 + DNA	1.34/4.24
B2 + FHRBM	1.18/3.48
B2 + DNA-CD	1.10/3.17
fMMI+fMLLR (B3)	1.08/3.00
B3 + SNM	1.25/3.59
B3 + DNA	1.06/3.04
B3 + FHRBM	1.03/2.95
B3 + DNA-CD	0.93/2.59
fMMI+fMLLR+SS (B4)	1.00/2.79
B4 + SNM	1.26/3.56
B4 + DNA	1.02/3.03
B4 + FHRBM	0.99/2.82
B4 + DNA-CD	0.95/2.67

Table 1. Word Error Rates and Sentence Error Rates of our embedded recognizer as a function of system configuration and front-end system. All results were obtained using an acoustic model trained in a discriminatively trained fMMI feature space. Please refer to the text for a full description of the algorithms and results.

6. DISCUSSION

The preliminary results presented in this paper are promising. However, several important experiments and research directions remain. Network topology/size of the speech/noise RBMs, more thorough comparisons against GMM-based front-ends, and alternative, more accurate inference methods such as structured variational approaches still need to be investigated. Re-training the back end recognizer on FHRBM processed outputs or incorporating condition detection [9] should further improve performance, as should state-specific variance modeling, and adaptation of the parameters of the FHRBM during inference. FHRBMs seem ideal for and need to be investigated in the context of more structured/non-stationary background noise robustness. Ultimately, FHRBMs should be jointly trained, and directly used by the decoder as an acoustic model.

7. REFERENCES

[1] A. Nádas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1495–1503, 1989.

[2] Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *ICASSP*, 1996.

[3] B. Frey et al., "Algonquin - learning dynamic noise models from noisy speech for robust speech recognition," *NIPS*, 2001.

[4] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. on Speech and Audio Processing*, vol. 12:2, pp. 133–143, 2004.

[5] M. Welling, M. Rosen-Zvi, and G. E. Hinton, "Exponential family harmoniums with an application to information retrieval," in *NIPS*, 2005.

[6] A. Mohamed and G. E. Hinton, "Phone recognition using restricted boltzmann machines.," in *ICASSP*, 2010.

[7] S. Rennie, P. Dognin, and P. Fousek, "Robust speech recognition using dynamic noise adaptation," in *ICASSP*, May 2011.

[8] Raimo Bakis, David Nahamoo, Michael A. Picheny, and Jan Sedivy, "Hierarchical labeler in a speech recognition system," U.S. Patent 6023673. filed June 4, 1997, and issued February 8, 2000.

[9] Steven J. Rennie, Pierre L. Dognin, and Petr Fousek, "Matched-condition robust dynamic noise adaptation," in *ASRU*, 2011.