GAUSSIAN-BERNOULLI RESTRICTED BOLTZMANN MACHINES AND AUTOMATIC FEATURE EXTRACTION FOR NOISE ROBUST MISSING DATA MASK ESTIMATION

Sami Keronen *KyungHyun Cho* *Tapani Raiko* *Alexander Ilin* *Kalle Palomäki*

Aalto University School of Science
Department of Information and Computer Science
PO Box 15400, FI-00076 Aalto, Finland

ABSTRACT

A missing data mask estimation method based on Gaussian-Bernoulli restricted Boltzmann machine (GRBM) trained on cross-correlation representation of the audio signal is presented in the study. The automatically learned features by the GRBM are utilized in dividing the time-frequency units of the spectrographic mask into noise and speech dominant. The system is evaluated against two baseline mask estimation methods in a reverberant multisource environment speech recognition task. The proposed system is shown to provide a performance improvement in the speech recognition accuracy over the previous multifeature approaches.

Index Terms— Noise robust, speech recognition, mask estimation, GRBM, deep learning

1. INTRODUCTION

Missing data methods, one of the many approaches for reducing the gap between human listeners and automatic speech recognition (ASR) in noisy environments, are based on studies motivated by the human auditory system [1]. In missing data methods, the noise corrupted speech is divided into reliable, speech-dominated, and unreliable, noise-dominated, components. The unreliable components can be discarded, as in the marginalization approach, used as an upper bound to the missing clean speech values [2], or they can be reconstructed by the respective clean speech estimates [2, 3].

Estimating the reliable and unreliable spectro-temporal regions of the speech signal, i.e. mask estimation, can be challenging in varying noise environments. Some of the previous work on the field have considered mask estimation a binary classification problem by training machine learning based classifiers such as Gaussian mixture models (GMMs) [4, 5] or support vector machines (SVMs) [6] with several acoustic features in conjunction. These multifeature approaches counteract the adverse environmental factors with their comprehensive set of features – cues discriminating between speech and non-speech are effective in non-speech noisy environments [4], whereas directional cues provide information on competing speakers [5, 7].

As an alternative to basing the multifeature approach on a set of "design" features, a GRBM [8] can be trained to learn the acoustical patterns for an arguably better performing set of features. Due to the contrastive divergence algorithm [9] and the recent advances in graphics processing units, GRBMs and deep belief networks (DBNs) are displacing the traditional combination of hidden Markov models and GMMs as the basis of the state of the art ASR systems. GRBMs and DBNs are capable of learning the acoustical patterns efficiently, which has been shown in many speech related tasks such as phone and large vocabulary speech recognition [10, 11, 12], speech separation [13], and likability classification [14].

Ultimately, the confrontation between design and automatically learned features reduces to quantity versus quality; the discrimination power of a single automatically learned feature may be small but the number of them can be made arbitrarily large, whereas a single design feature such as interaural time difference (ITD) or interaural level difference (ILD) [5, 7] may be effective alone but the overall number of them is usually much smaller. Additionally, the multilayer DBNs may provide higher level information on the audio signal [12], which may further improve the system performance.

In this paper, we use GRBMs to learn the cross-correlation representation of a dual channel multisource reverberant CHiME corpus and apply it in a missing data reconstruction-based automatic speech recognition task. The speech recognition performance of the proposed system is evaluated against systems based on 14 design features and on the unprocessed channel-wise cross-correlation values.

2. METHODS

In this section, we describe the proposed method starting with an introduction to missing data mask estimation, followed by descriptions of GRBM, feature extraction and GRBM training, classifier, and reconstruction of missing data.

2.1. Missing data mask estimation

Missing data techniques are based on the assumption that the spectro-temporal units of the noisy speech can be divided into

speech and noise dominated regions [2]. Typically a set of log-mel features $\boldsymbol{Y}$ is computed for each time-frequency (TF) unit of the speech signal and a so called spectrographic mask labels the feature observations as speech or noise dominated. A TF unit $Y_r(\tau, d)$ is considered reliable if $Y(\tau, d) \approx S(\tau, d)$, where τ denotes the time frame, d the frequency channel, and $S(\tau, d)$ the clean speech signal without corrupting noise. Units $Y_u(\tau, d)$ are considered unreliable if $Y(\tau, d) \geq S(\tau, d)$.

In this work, binary valued masks are used for labeling and cluster-based imputation [3] (see Sec. 2.5), which has been shown to perform well on various speech recognition tasks, is used to reconstruct the missing values.

2.2. Gaussian-Bernoulli Restricted Boltzmann Machines

A GRBM is a neural network that models the probability density of continuous-valued data using binary latent variables. It consists of a layer of Gaussian visible units that correspond to components of data vectors, and a layer of binary hidden units. Each unit is connected to all units in the other layer (i.e. no lateral connections). It has been shown that the latent variables of a learnt GRBM can be used as meaningful unsupervised features (see, e.g., [10, 15]).

The energy given by a GRBM to each state of visible units v_i and hidden units h_j is defined as

$$E(\mathbf{v}, \mathbf{h} \mid \theta) = \sum_{i=1}^{n_v} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{n_v} \sum_{j=1}^{n_h} w_{ij} h_j \frac{v_i}{\sigma_i^2} - \sum_{j=1}^{n_h} c_j h_j,$$

where n_v and n_h are the numbers of hidden and visible units, and the parameters $\boldsymbol{\theta}$ include weights w_{ij} connecting the visible and hidden units, the standard deviation σ_i associated with a Gaussian visible unit v_i, and biases b_i and c_j for each unit [16]. Based on it, one can define a Boltzmann distribution by $p(\mathbf{v}, \mathbf{h} \mid \theta) = \frac{1}{Z(\theta)} \exp\{-E(\mathbf{v}, \mathbf{h} \mid \theta)\}$.

Due to the bipartite structure, the visible units given the the hidden units are conditionally independent, and the probability of each visible unit is

$$p(v_i = v|\mathbf{h}) = \mathcal{N}\left(v \mid b_i + \sum_j h_j w_{ij}, \sigma_i^2\right),$$

where $\mathcal{N}(\cdot \mid \mu, \sigma^2)$ denotes the Gaussian p.d.f. with mean μ and variance σ^2. Similarly, the hidden units are conditionally independent, and their probabilities are given by

$$p(h_j = 1|\mathbf{v}) = \text{sigmoid}\left(c_j + \sum_i w_{ij} \frac{v_i}{\sigma_i^2}\right),$$

which makes it simple to compute the activations of features learned by the GRBM for further use.

Learning parameters of a GRBM is commonly done by maximizing the log-likelihood function with a stochastic gradient. Recently in [17], the enhanced gradient was proposed and shown to outperform the conventional gradient update direction. In the enhanced gradient, each weight parameter w_{ij} is updated by

$$w_{ij} \leftarrow w_{ij} + \eta \left[\text{Cov}_\text{d}\left(v_i, h_j\right) - \text{Cov}_\text{m}\left(v_i, h_j\right)\right], \quad (1)$$

where $\text{Cov}_p(v_i, h_j)$ is a covariance between v_i and h_j under distribution p, and d and m denote the data distribution $p(\mathbf{h} \mid \mathbf{v}, \boldsymbol{\theta})D(\mathbf{v})$ and the model distribution $p(\mathbf{v}, \mathbf{h} \mid \boldsymbol{\theta})$, respectively. Learning rate η can be automatically adjusted by the adaptive learning rate proposed in [16, 17].

In [18], it was shown that a GRBM can learn more discriminative filters when the binary hidden units were replaced by the noisy rectified linear units (NReLU). When NReLU hidden units are used, the approximate mean activation of the hidden unit becomes

$$h_j = \max\left(0, a_j\right), \quad (2)$$

where $a_j = \sum_i w_{ij} \frac{v_i}{\sigma_i^2} + c_j$ is the input to the j^{th} hidden unit.

2.3. Feature Extraction and GRBM training

In this work, cross-correlation vectors from bandpass filtered speech signals were used as input to the GRBM to generate a set of features, according to the following description.

First, the left-ear $x_l(n)$ and right-ear $x_r(n)$ speech signals, where n denotes the sample number, were filtered into 21 bandpass signals $X_l(n, d)$ and $X_r(n, d)$, respectively, where d is the frequency channel. The center frequencies between 171 Hz and 7097 Hz of the bandpass filters were designed to match the centers of the triangular filters used in conventional audio-MFCC conversion.

Second, the cross-correlation values between windowed (i.e. framed) bandpass filtered signals, starting from sample n, $\boldsymbol{w}_l(n, d) = [X_l(n, d), \ldots, X_l(n+N-1, d)]$ and $\boldsymbol{w}_r(n, d) = [X_r(n, d), \ldots, X_r(n + N - 1, d)]$ with lags l ranging from -50 to 49 are computed as follows

$$R(n, l, d) = \begin{cases} \sum_{t=0}^{N-l-1} w_l(t+l, n, d) w_r(t, n, d)^* & l \geq 0 \\ R^*(n, -l, d) & l < 0 \end{cases}, \quad (3)$$

where $N = 256$ denotes the length of the rectangular window and $[]^*$ the complex conjugate. The cross-correlation vector for a frame starting at sample n on channel d is obtained by

$$\boldsymbol{x}_{corr}(n, d) = [R(n, -50, d), \ldots, R(n, 49, d)]. \quad (4)$$

A single GRBM with 50 hidden units was trained with 20,000 coefficient normalized sample vectors in 2000 epochs and a mini-batch size of 64. Initially, the number of hidden units n_h was varied from 25 to 150 at 25 unit intervals and in small scale testing, 50 units was found optimal. The sample vectors, $\boldsymbol{x}_{corr}(n, d)$ with random n and d values, were arbitrarily selected from the CHiME development set utterances,

described in Section 3.1, in a way that the training corpus contained approximately equal amount of data from all the frequency channels and signal-to-noise ratios (SNRs). NReLU hidden units, CD with the enhanced gradient and adaptive learning rate were used. A single σ was shared and learned for all visible units.

In evaluation, the input vectors $\boldsymbol{x}_{corr}(n, d)$ to the GRBM were computed by converting the bandpass filtered and cross-correlated speech signals into a series of 256 samples long frames with consecutive frames overlapping by 128 samples.

2.4. Classifier

For classifying the TF units into reliable and unreliable, separate SVMs with radial basis function (RBF) kernels were trained for each frequency channel d. Oracle masks were used as targets, while the mean hidden activations of the GRBM given in Eq. (2) were taken as input features. The oracle masks were constructed using the noisy and clean CHiME development data to compute the exact SNR of each TF unit; only the units with SNR over 0 dB were labeled reliable. For each SVM-based system, a single RBF kernel width γ was used. γ values of 2.0 were found optimal for both the baseline mask estimation system (BME+SVM) and for the system taking the $\boldsymbol{x}_{corr}(n, d)$ vectors directly as SVM input (XCOR+SVM). For the proposed GRBM mask estimation system (GME+SVM), the optimal γ value was 2.5. The widths were tuned by using the features computed from a set of 200 randomly selected utterances from all the SNRs of the CHiME development set.

In evaluation, TF regions that contained less than 20 connected reliable elements were removed from the masks.

2.5. Reconstruction of missing data

In this work, cluster-based imputation (CBI) is used to reconstruct the missing data. In CBI, a GMM is created to represent the distribution of feature vectors of clean speech. The model is used to fill the missing values of the observed feature vector with the most probable values. CBI assumes that the reliable components of the observation vector are the real values of a clean speech feature vector and the unreliable components represent an upper bound to the clean speech estimate; this is derived from the additive noise assumption which states that the energy of a signal with additive noise is always higher than the energy of a clean signal. A more detailed description of CBI can be found in [3] and [19].

The missing feature components were reconstructed in 21-dimensional log-compressed mel-spectral domain and the features were processed in 5-frame windows with a window shift of one frame as described in [20]. 1,500 randomly selected utterances from the CHiME training set were used to train a 13-component clean speech GMM with 105-variate component densities and full covariance matrices.

3. EXPERIMENTS

3.1. Data

Here, we use CHiME challenge data [21], where spoken commands are recognized from recordings made in a noisy living room using a binaural dummy head. Target speaker, represented by the binaural impulse responses of the dummy head, was in a fixed 2 meter in front position relative to the head. The data set is divided into training, development and evaluation sets. The training set consists of 17,000 utterances of reverberated but noise-free speech. The development and evaluations sets consist of 600 shared speaker utterances mixed with 6 different SNRs (from -6 to 9 dB at 3 dB intervals) giving 3,600 utterances in total.

3.2. Speech recognition system

The baseline system (BL) used in this work is a hidden Markov model (HMM) based large vocabulary continuous speech recognizer (LVCSR). The acoustic models of the BL system are speaker independent state-tied triphones. The triphone segmentations of the CHiME training data were generated by an LVCSR trained on the WSJ British English corpus [22]. The HMM states are modeled with at most 100 Gaussians (with diagonal covariance matrices) and their durations are modeled with gamma distributions. The speech signal is represented as frames of 12 MFCC and a frame power feature together with their first- and second-order derivatives. Cepstral mean subtraction and maximum likelihood linear transformation are also applied. A language model based on no-backoff bigrams with uniform frequencies for all valid bigrams is used.

The performances of the CHiME challenge baseline system (CBL) and a baseline mask estimation system based on 14 design features [5] and an SVM classifier (BME+SVM) are also presented in the current study. CBL and BL systems differ in that CBL is trained speaker dependently and whole-word HMMs are used.

For comparison, results of the BME system coupled with a GMM classifier (BME+GMM), trained with nine times more data than the SVM classifiers applied here, are presented from our previous paper [5]. The classifier of the BME+SVM system was trained with the 14 design features computed from the same set of 200 randomly selected utterances used to train the other SVM classifiers. The acoustic features used for mask estimation in the BME systems includes modulation-filtered spectrogram, mean-to-peak-ratio and gradient of the temporal envelope, harmonic and inharmonic energies, noise estimates from long-term inharmonic energy and channel difference, noise gain, spectral flatness, subband energy to subband noise floor ratio, ITD, ILD, peak ITD and interaural coherence.

Table 1. Keyword accuracy rates of CHiME baseline (CBL) system, our baseline system (BL), baseline mask estimation methods based on a 14-component design feature set with GMM (BME+GMM) and SVM (BME+SVM) classifiers, mask estimation method based on direct use of cross-correlation representation (XCOR+SVM), and the proposed GRBM mask estimation system (GME+SVM) for the CHiME development and evaluation sets.

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
Development set							
CBL	83.1	73.8	64.0	49.1	36.8	31.1	56.3
BL	83.3	80.0	69.8	55.2	46.0	40.6	62.5
BME + GMM	88.6	85.3	78.1	68.6	60.6	55.1	72.7
BME + SVM	89.7	87.4	78.7	67.7	58.3	53.6	72.5
XCOR + SVM	89.1	87.7	80.6	71.8	63.5	58.2	75.2
GME + SVM	90.0	87.1	82.2	73.1	64.1	59.4	76.0
Evaluation set							
CBL	82.5	75.0	62.9	49.5	35.4	30.3	55.9
BL	86.3	78.3	68.5	53.9	44.3	41.9	62.2
BME + GMM	90.3	84.3	76.9	68.2	58.2	56.3	72.3
BME + SVM	**91.0**	85.3	79.4	68.8	56.2	53.7	72.4
XCOR + SVM	90.5	**86.1**	80.0	69.2	57.4	55.8	73.1
GME + SVM	90.7	85.8	**81.0**	**69.8**	**61.4**	**58.9**	**74.6**

Table 2. Statistical significances of pairwise system comparisons of the evaluation set presented in Table 1. "+" denotes a statistically significant and "-" a non-significant difference.

Pair	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
BME+GMM - BME+SVM	-	-	+	-	-	+	-
BME+GMM - GME+SVM	-	-	+	-	+	-	+
BME+GMM - XCOR+SVM	-	-	-	-	-	-	+
BME+SVM - XCOR+SVM	+	-	-	-	-	-	+
BME+SVM - GME+SVM	-	-	-	-	+	+	+
XCOR+SVM - GME+SVM	-	-	-	-	+	+	+

3.3. Results

The keyword accuracies of the systems are gathered in Table 1. The highest scores on each evaluation set SNR is shown in bold type. On evaluation set, the lowest accuracy rates at every SNR is obtained by CBL (55.9% on average) followed by BL (62.2% on average). BME+GMM and BME+SVM offer similar average performance (72.3% and 72.4%, respectively) but BME+GMM achieves higher accuracy rates in below zero SNR cases, whereas BME+SVM outperforms the BME+GMM at the higher SNR cases. The highest score at 9 dB is obtained by BME+SVM (91.0%). XCOR+SVM exceeds the accuracy rates of both BME systems on average (73.1%) and achieves the highest score on 6 dB case (86.1%). The proposed GME+SVM is the best performing system on average (74.6%) and at SNRs from 3 dB to -6 dB (81.0%, 69.8%, 58.9%, 74.6%, respectively).

Statistical significance of the keyword accuracy difference between each system pair on the evaluation set was computed by the Wilcoxon signed-rank test with a 95% confidence level and the results of the analysis are presented in Table 2.

4. DISCUSSION

We have presented a mask estimation method based on automatic feature extraction from cross-correlation representation of binaural speech signal using a GRBM and an SVM classifier (GME+SVM). The proposed method is able to learn features exceeding the performance of advanced design features.

In some of the previous studies, the common approach for time-frequency unit classification has been to develop descriptive heuristic measures, or design features, some of which are processed through a rather complex model [4, 5]. However, relevant information may be lost when data is described with just a few features. With the help of modern machine learning methods such as GRBM feature extraction coupled with an SVM classifier, we can overcome the problem by using input signals in a less refined format. Even without GRBM feature extraction, we achieved better results with SVM classifier utilizing raw cross-correlation data as input than with the design features. Similarly, a recent study by Wang et al. [23] suggested combining a number of standard ASR features that were less processed than design features for missing data mask estimation.

Initially, two alternative approaches were also considered for the feature extraction. First, learning the discriminating patterns directly from the noisy signal provided only a small improvement over the BL system in the current task. Second, making the net "deeper" showed no improvement in performance over the proposed method. These findings suggest that there may be room for an improvement in GRBM training algorithms. The next step could also be to investigate whether GRBMs were useful in reconstructing the missing data.

The recent work by Wang and Wang [13] boosted speech separation by modeling temporal dynamics with DBNs on monaural audio signal, and the work by Dahl et al. [11] and Hinton et al. [12] successfully adapted DBNs on various LVCSR tasks. Inspired by the recent advances in combining neural networks and ASR, we have continued our previous study applying multiple design features to missing data mask estimation [5] by using automatically generated features taking advantage of the modeling capabilities of GRBMs.

5. ACKNOWLEDGMENTS

The work was financially supported by Langnet (Keronen) and FICS (Cho) graduate schools, TEKES under the FuNeSoMo project (Palomäki), and by the Academy of Finland under the grants no 133145 (Raiko), 134935 (Ilin), 136209 (Palomäki), and 251170 Finnish Centre of Excellence Program (2012-2017).

6. REFERENCES

[1] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *Proc. ICSLP*, Yokohama, Japan, September 1994, pp. 1555–1558.

[2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[3] B. Raj, M. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, 2004.

[4] M. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[5] S. Keronen, H. Kallasjoki, U. Remes, G. J. Brown, J. F. Gemmeke, and K. J. Palomäki, "Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment," *Computer Speech & Language*, vol. 27, no. 3, pp. 798–819, 2013.

[6] J. F. Gemmeke, Y. Wang, M. Van Segbroeck, B. Cranen, and H. Van hamme, "Application of noise robust MDT speech recognition on the SPEECON and SpeechDat-Car databases," in *Proc. INTERSPEECH*, Brighton, UK, September 2009, pp. 1227–1230.

[7] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 58–67, 2006.

[8] G. Hinton and R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.

[9] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, August 2002.

[10] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted Boltzmann machines," in *Proc. ICASSP 2011*, Prague, Czech Republic, May 2011, pp. 5884–5887.

[11] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[12] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 28, no. 6, 2012.

[13] Y. Wang and D. Wang, "Boosting classification based speech separation using temporal dynamics," in *Proc. INTERSPEECH*, Portland, OR, USA, September 2012.

[14] R. Brueckner and B. Schuller, "Likability classification a not so deep neural network approach," in *Speaker Trait Challenge*, Portland, Oregon, United States, 2012.

[15] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., Computer Science Department, University of Toronto, 2009.

[16] K. Cho, A. Ilin, and T. Raiko, "Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines," in *Proc. ICANN*, Espoo, Finland, 2011, pp. 10–17.

[17] K. Cho, T. Raiko, and A. Ilin, "Enhanced Gradient and Adaptive Learning Rate for Training Restricted Boltzmann Machines," in *Proc. ICML*, Bellevue, Washington, USA, 2011.

[18] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *ICML*, Haifa, Israel, 2010, pp. 807–814.

[19] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[20] U. Remes, Y. Nankaku, and K. Tokuda, "GMM-based missing feature reconstruction on multi-frame windows," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 2407–2410.

[21] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[22] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, Detroit, MI, USA, 1995, pp. 81–84.

[23] Y. Wang, K. Han, and D. Wang, "Acoustic features for classification based speech separation," in *Proc. INTERSPEECH*, Portland, OR, USA, September 2012.