NOISE ROBUST MISSING DATA MASK ESTIMATION BASED ON AUTOMATICALLY LEARNED FEATURES

Sami Keronen *Ulpu Remes* *Heikki Kallasjoki* *Kalle Palomäki*

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

ABSTRACT

In this work, we present a missing feature reconstruction based automatic speech recognition (ASR) system in which masks are estimated by binary classification of features generated by Gaussian-Bernoulli restricted Boltzmann machines (GRBMs). The system is evaluated on Track 1 of the 2nd CHiME challenge data. Overall, the best performance is achieved when the reconstructed speech features are recognized with a discriminatively trained matched condition model.

Index Terms— Noise robust, speech recognition, mask estimation, GRBM, discriminative training

1. INTRODUCTION

Missing data methods for noise robust ASR assume that the noise corrupted speech can be divided into reliable, speech-dominated, and unreliable, noise-dominated, components which can be indicated with so called spectrographic masks. A common way to estimate masks, and the way used in this work, is to train a binary classifier on a set of features computed for each time-frequency (TF) unit of the noisy speech signal. Given the mask, the unreliable components can then be, for example, reconstructed by the respective clean speech estimates. Some of the previous studies have solved the mask estimation problem by training the classifier with features designed by the authors [1] or by generating them automatically with GRBMs [2]. GRBMs and their respective multilayer versions, deep belief networks, have recently made a break through in state of the art ASR systems since their superior capabilities to learn acoustical patterns (see e.g. [3]). Here, GRBMs are used to learn the acoustical patterns for generating a quality set of features. This work advances our recent study (see [2] for more detailed description of the method and evaluation against other mask estimation methods) by investigating improved ASR training and missing feature imputation, and applying them to the Track 1 of the 2nd CHiME challenge data.

2. MISSING DATA MASK ESTIMATION

2.1. 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 two layers, the first of Gaussian visible units that correspond to components of data vectors, and the second of binary hidden units. Each unit of one layer is connected to all units in the other layer.

The work was financially supported by Langnet and Hecse graduate schools, Tekes under the FUNESOMO project, and by the Academy of Finland under the grants no 135003, 136209 and 251170 Finnish Centre of Excellence in Computational Inference Research.

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

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

where n_v and n_h are the numbers of visible and hidden 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 [4]. The probability of each visible unit can be defined through Eq. 1 and the Boltzmann distribution as $p(v_i = v|\mathbf{h}) = \mathcal{N}(v \mid b_i + \sum_j h_j w_{ij}, \sigma^2)$, where $\mathcal{N}(\cdot \mid \mu, \sigma^2)$ denotes the Gaussian p.d.f. with mean μ and a variance σ^2 shared by all visible units v_i. The input to the j^{th} noisy rectified linear hidden unit [3] is given by $a_j = \sum_i w_{ij} \frac{v_i}{\sigma^2} + c_j$. Therefore, the approximate mean activation of the hidden unit becomes

$$h_j = \max\big(0, a_j\big). \quad (2)$$

Each weight parameter w_{ij} is updated with enhanced gradient method, whose learning rate is automatically adjusted by the adaptive learning rate [4].

2.2. Feature Extraction and GRBM training

In this work, cross-correlation vectors of bandpass filtered (denoted here as BPF) speech signals were used as input to multiple GRBMs to generate a set of features. To generate the input, the left-ear $x_l(n)$ and right-ear $x_r(n)$ signals, where n is the time domain sample index, were filtered into 21 BPF signals $X_l(n, d)$ and $X_r(n, d)$, where d is the frequency channel. The center frequencies of the filters conformed to the audio-MFCC conversion used in our baseline systems. After that, the cross-correlation values between the windowed BPF 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

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

where $N = 256$ denotes the window length and $()^*$ the complex conjugate. Thus, the cross-correlation vector for a frame starting at sample n on channel d is obtained by $\boldsymbol{x}_{corr}(n, d) = [R(n, -50, d), \ldots, R(n, 49, d)]$.

For each channel d, separate GRBMs with 50 hidden units were trained with 2,000 sample vectors in 100 epochs and a mini-batch size of 64. The sample vectors, $\boldsymbol{x}_{corr}(n, d)$ with random n values, were arbitrarily selected from the noisy training set so that the training corpus for each GRBM contained approx. equal amount of respective channel data from all the signal-to-noise ratios (SNRs). In evaluation, the inputs to the GRBMs were computed from speech signals converted into a series of 256 samples long frames with a shift of 128 samples.

2.3. Classifier

For classifying the TF units into reliable and unreliable, separate SVMs with radial basis function (RBF) kernels were trained for each frequency channel. The activations of the GRBMs given in Eq. (2) were taken as input features and oracle masks were used as targets. The binary oracle masks were constructed using the noisy and reverberated training data to compute the exact SNR of each TF unit with a reliability threshold of 0 dB. A single RBF kernel width was used for all classifiers. TF regions that contained less than 20 connected reliable elements were removed from the estimated masks.

2.4. Missing-feature reconstruction

Missing-feature reconstruction is applied in the 21-dimensional log-mel-spectral domain. TF units that have been classified as reliable are used as estimates for the corresponding clean speech values whereas units classified as unreliable are substituted with estimates calculated based on a clean speech prior. The clean speech prior used in this work is a 13-component full-covariance GMM trained on 5-frame windows as described in [1]. The model was trained on 1,500 utterances sampled from the reverberated training set. Given the GMM prior and the reliable features, a GMM posterior is calculated for the missing features. The posterior is approximated with a Gaussian and feature estimates are calculated as the bounded mean of the approximate posterior (BCMI) as proposed in [5]. In the experiments reported in [5] and also in our preliminary experiments with the 2nd CHiME challenge data, BCMI outperformed the cluster-based imputation method used in our previous work [1].

3. EXPERIMENTS

3.1. Data and speech recognition systems

The Track 1 of the 2nd CHiME challenge [6] considers the problem of recognizing spoken commands from recordings made in a noisy living room using a binaural dummy head. The data set is divided into three training sets, a development set and an evaluation set. The training sets consist of 17,000 utterances of either clean, reverberated but noise-free, or reverberated and noisy speech. The development and evaluations sets consist of 600 shared speaker utterances mixed with 6 SNRs from -6 to 9 dB at 3 dB intervals.

The baseline system used in this work is a HMM based large vocabulary continuous speech recognizer (LVCSR) adapted to the CHiME recognition task (see [1] for details) with superior performance compared to the reverberated challenge baseline (CBL).

As for the baseline, we trained two systems with standard maximum likelihood criterion; a regular baseline (BL) using the reverberated training set and a matched condition baseline (MBL) using the noisy training set. A third baseline system was trained discriminatively with minimum phone frame error criterion (MBL+MPFE) using the noisy training set. The features reconstructed by the proposed missing data method based system (MM+GRBM) were recognized with the MBL+MPFE system. For the 2nd CHiME challenge, unsupervised maximum likelihood linear regression was applied to the MBL+MPFE and MM+GRBM systems denoted as MM+MLLR and MG+MLLR, respectively. The adaptation data for each speaker was obtained from the first-pass recognition hypotheses of MBL+MPFE and MM+GRBM systems using the entire evaluation data set.

3.2. Results

The keyword accuracies of the systems are collected in Table 1. The highest scores on each evaluation set SNR is shown in bold type. The highest accuracies on the evaluation set are achieved by MG+MLLR in all SNRs and on average (79.4%). The results of MM+GRBM and MG+MLLR are not directly comparable to the official challenge results because the fact that the same utterances are used within the reverberated and noisy training data was exploited in constructing the oracle masks. The pairwise accuracy differences between the system averages are all statistically significant (Wilcoxon signed-rank test with 95% confidence level).

Table 1. Keyword accuracy rates of the 2nd CHiME challenge baseline (CBL) system and our systems for the development and evaluation sets of Track 1.

	Development set						
	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
CBL	83.5	75.1	64.0	50.3	36.3	32.1	56.9
BL	86.1	81.3	70.4	58.4	47.2	40.8	64.0
MBL	86.4	83.2	80.5	70.5	63.8	55.8	73.4
MBL+MPFE	88.4	85.1	82.1	73.6	67.6	56.7	75.6
MM+GRBM	88.2	85.5	82.7	76.3	69.9	64.3	77.8
	Evaluation set						
CBL	83.8	76.1	62.7	52.1	38.2	32.2	57.5
BL	87.3	80.3	70.2	57.3	45.6	42.0	63.8
MBL	86.3	83.3	78.9	71.8	64.1	54.3	73.1
MBL+MPFE	88.6	86.8	80.8	74.6	66.3	57.9	75.8
MM+MLLR	88.2	86.9	81.3	75.7	66.6	59.3	76.3
MM+GRBM	88.0	85.8	82.9	77.4	68.8	63.5	77.7
MG+MLLR	**89.8**	**87.4**	**84.2**	**77.8**	**71.3**	**65.6**	**79.4**

4. CONCLUSIONS

We have presented an ASR system based on automatic feature extraction from cross-correlation representation of binaural speech signal using GRBMs, SVM classifiers and BCMI feature reconstruction (MM+GRBM). The system outperforms the next best system, the discriminatively trained matched condition system (MBL+MPFE), in all but two SNR cases and the other baselines in all SNR cases.

5. REFERENCES

[1] 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.

[2] S. Keronen, K. Cho, T. Raiko, A. Ilin, and K. Palomäki, "Gaussian-Bernoulli restricted Boltzmann machines and automatic feature extraction for noise robust missing data mask estimation," in *Proc. ICASSP*, 2013.

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

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

[5] U. Remes, "Bounded counditional mean imputation with an approximate posterior," submitted.

[6] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. ICASSP*, 2013.