SPARSE AUTOENCODERS TO ENHANCE SPEECH RECOGNITION

Selen Hande Kabil Hervé Bourlard

Idiap-RR-10-2022

AUGUST 2022

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Sparse Autoencoders to Enhance Speech Recognition

Selen Hande Kabil[1,2] *and Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

selen.kabil@idiap.ch, bourlard@idiap.ch

Abstract

Starting from a state-of-the-art Lattice-Free Maximum Mutual Information (LF-MMI) speech recognition system as baseline, we investigate the use of shallow sparse autoencoders to further process LF-MMI acoustic model outputs (i.e., LF-MMI senone likelihoods) to produce more reliable phone or senone likelihoods. The sparse and overcomplete autoencoders investigated here are shown to project the LF-MMI senone likelihoods to a high-dimensional sparse space. Detailed analysis of this representation space shows that the encodings obtained from the autoencoder are indeed sparse and capable of improving classification performance. Motivated by this, we then exploit the resulting high-dimensional sparse representations to train a new acoustic model. Different combinations of the available acoustic models are explored. In particular, it is shown that combining the senone likelihoods from the acoustic model trained on the high-dimensional sparse encodings and from the acoustic model in the baseline system leads to promising improvements in the recognition performance on Augmented Multiparty Interaction (AMI) data set for both Individual Head-mounted Microphone (IHM) and Single-Distant Microphone (SDM) tasks.

Index Terms: speech recognition, deep neural network, sparse autoencoder, high-dimensional sparse representations

1. Introduction

Automatic Speech Recognition (ASR) is a fast evolving research field where significant effort has been made to produce more robust systems. In particular, for acoustic modeling, different architectures such as Convolutional Neural Networks (CNN) [1], Time-Delay Neural Network (TDNN) [2], Long Short-Term Memory Recurrent Neural Network (LSTM) [3] are utilized alongside to different training procedures such as Lattice-Free Maximum Mutual Information (LF-MMI) [4]. Several studies have been conducted to improve the recognition performance by processing the input and output features of the DNN acoustic models. For instance, it is shown in [5] that Linear Discriminant Analysis (LDA) can be used to project the input features and improve the ASR performance. Similarly, in [6], the modeling of DNN-based posteriors using low-rank and sparse modeling approaches is studied. For low-rank modeling, posterior features belonging to a particular senone class (based on ground truth alignments) are stacked together to form a senone-specific posterior matrix. The principal components learned using Principal Component Analysis (PCA) on each of these matrices acts as senone-specific dictionaries. They cover global patterns which emphasize the senone characteristics while ignoring the impact of local distortions such as noise. When a posterior sample is projected over its corresponding senone-specific dictionary, low-dimensional, dense, senone-wise representations are extracted as shown in Fig. 1 (low-rank modeling). Whereas, for sparse modeling, the senone-specific dictionaries from low-rank modeling are concatenated to form an initialization for the overcomplete dictionary, which is later trained by online dictionary learning algorithm [7]. The overcomplete dictionary is capable of modeling non-linear speech manifold as a union of low-dimensional spaces. Hence, when posterior features are projected over the overcomplete dictionary, senone-specific sparse representations manifest themselves on different dimensions in the common high dimensional sparse space, as shown in Fig. 1 (sparse modeling). Finally, these intermediate representations extracted either by low rank (PCA based) or sparse (dictionary learning based) modeling are projected back onto the original dimensions of the DNN posterior space. These reconstructed enhanced DNN posteriors are shown to be better targets for training more robust acoustic models, leading to improvements on recognition performance.

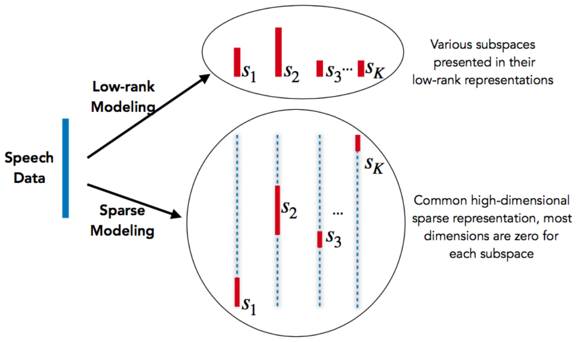
Speech Data
Low-rank Modeling
Sparse Modeling
s_1
s_2
s_3
s_K
Various subspaces presented in their low-rank representations
Common high-dimensional sparse representation, most dimensions are zero for each subspace

Figure 1: *Sparse and low-rank modeling of acoustic model output frame. s_i represents the i-th senone. Figure adapted from [6].*

The most significant drawback in this approach is that the enhancement of the DNN posteriors are done in supervised manner. For both low-rank and sparse modeling, ground truth based senone alignments are needed. Hence, in [6], for unseen test data without transcriptions and senone alignments, an additional computation procedure is introduced. Inspired from the work on knowledge distillation [8], a new network (i.e., a student DNN) is trained using the training set acoustic features as input and corresponding enhanced DNN posteriors as soft targets. After training, unseen test data is forwarded through this new student network to get posteriors for ASR decoding. Apart from this additional training procedure, this approach is not easily scalable for large vocabulary speech recognition task, which usually involves thousands of senones.

In this paper, we propose a generic way of sparse modeling of speech that does not require prior knowledge about the senone classes in the dataset. We use an acoustic model trained using LF-MMI [4] to generate senone likelihoods. Then, we train a shallow sparse autoencoder on these senone likelihoods to reach their corresponding high-dimensional sparse representations. The analysis of these representations reveals that they are indeed sparse. As shown in Fig. 1 (sparse modeling), we expect the senone subspaces (shown in red) to lie in the common high dimensional sparse representation space in a scattered manner. *Our hypothesis is that this behaviour can induce the separability of the senones (and even phones), which can be exploited in speech recognition. To validate our hypothesis, we train a senone classifier using these high dimensional sparse representations and report higher senone accuracy with respect to the LF-MMI acoustic model.* Finally, we get promising improvements on WER when a weighted combination of the senone likelihoods coming from the senone classifier and the LF-MMI acoustic model is used.

This paper is organized as follows. In Section 2, we present our motivation and methodology with a guideline for the models that we use in the experiments. In Section 3, we give the implementation details for the models introduced in the previous section along with the results and the analysis. Finally, in Section 4, we present our conclusion.

2. Proposed Approach

As discussed in Section 1, dictionary learning based sparse modeling approach, where DNN-based posteriors y_t at time t are projected onto a high dimensional space through an overcomplete dictionary D, is shown to be useful for improving the recognition performance. In this paper, we propose a generic way of sparse modeling by means of sparse autoencoders. The motivaton for employing sparse autoencoder for our proposed approach stems from the relation between the dictionary learning and sparse autoencoder concepts.

The goal of the dictionary learning is to find a dictionary $\mathbf{D} \in \mathbb{R}^{d \times n} : \mathbf{D} = [d_1, \ldots, d_n]$ and a representation $\mathbf{X} = [x_1, \ldots, x, \ldots, x_K], \mathbf{x} \in \mathbb{R}^n$, given the input data $\mathbf{Y} = [y_1, \ldots, y, \ldots, y_K], \mathbf{y} \in \mathbb{R}^d$, such that all $\|\boldsymbol{y} - \mathbf{D}\boldsymbol{x}\|_2^2$ are minimized and the representations $\boldsymbol{x}$ are sparse. This can be formulated as the following optimization problem:

$$\min_{\mathbf{D},\boldsymbol{x_i}} \sum_{i=1}^{K} \|\boldsymbol{y_i} - \mathbf{D}\boldsymbol{x_i}\|_2^2 + \lambda \|\boldsymbol{x_i}\|_0, \|\boldsymbol{d_i}\|_2 \leq 1 \qquad (1)$$

Since (1) is NP-hard [9], the ℓ_0-norm $\|\mathbf{x}_i\|_0$ is usually relaxed to the ℓ_1-norm $\|\mathbf{x}_i\|_1$.

Once learned, the overcomplete dictionary $\mathbf{D}$ is used for sparse coding [10]. That is, it projects the data $\boldsymbol{y_i}$ into a high-dimensional sparse space where the sparse representation $\boldsymbol{x_i}$ resides. The projection leads the various phenomena in the data to disentangle and activate the relevant dimensions (entries) of $\boldsymbol{x_i}$. In addition, the ℓ_1-constraint on $\boldsymbol{x_i}$ allows only the predominant phenomena to be represented. This is useful for speech recognition where variation from different sources constitutes challenges. Hence, inspiring from the dictionary learning theory which is convenient for devising robust recognition systems, we work with sparse autoencoder.

An Autoencoder (AE) [11] has two main components, i.e., the encoder which maps the input to the code and the decoder which maps the code to the input reconstruction. Since its objective is to reconstruct the input, an autoencoder can learn the identity mapping if no constraints are enforced. These constraints can be in the form of limiting the size of encoding space which restricts the modeling capacity and forces the autoencoder to learn compact representations in the encoding layer. On the other hand, if the encoding is not limited and the encoding layer has the same or higher dimensionality than the input, a constraint (i.e. ℓ_1-constraint) can be introduced on the code so that the autoencoder is forced to learn meaningful representations.

Shallow (with one hidden layer), overcomplete (number of hidden units >> input feature dimension) autoencoder with linear activation, tied weights (encoder and decoder weights of the autoencoder are transpose of each other), no bias and ℓ_1-regularization on hidden unit activations exhibit similar mathematical properties with dictionary learning and share the same goal of projecting the data to high-dimensional sparse spaces. In this paper, we use this sparse autoencoder architecture to map senone likelihoods into a sparse representation:

$$\boldsymbol{x} = \boldsymbol{D^T}\boldsymbol{y}, \qquad (2)$$

where $\boldsymbol{D^T} \in \mathbb{R}^{N_x \times N_y}$ is the encoder weights of the autoencoder, $\boldsymbol{x} \in \mathbb{R}^{N_x}$ is the high-dimensional sparse representation of $\boldsymbol{y}$, and $\boldsymbol{y} \in \mathbb{R}^{N_y}$ is the senone likelihoods from the LF-MMI acoustic model per frame, as shown in Fig. 2. The encoder weights are just the transpose of the decoder weights $\boldsymbol{D} \in \mathbb{R}^{N_y \times N_x}$ which is a rectangular matrix ($N_x \gg N_y$). Hence, $\boldsymbol{D}$ here acts like the overcomplete dictionary used for sparse modeling in [6].

We obtain $\boldsymbol{D}$ by solving the following optimization problem

$$\min_{\boldsymbol{D},\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{D^T}\boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_1, \boldsymbol{x} = \boldsymbol{D^T}\boldsymbol{y}, \qquad (3)$$

where λ is an hyperparameter controlling the ℓ_1-regularization which promotes sparsity of $\boldsymbol{x}$. This is analogous to basis pursuit [12] in sparse recovery theory, and to LASSO regression [13] in statistics. When the sparse autoencoder is trained to solve the optimization problem (3), forward pass can be taken as sparse coding [10] step in dictionary learning, since we obtain high-dimensional sparse representation $\boldsymbol{x}$ on the hidden layer. Similarly, backward pass is analogous to the dictionary update step in dictionary learning, as decoder weights $\boldsymbol{D}$ are updated based on the distance between the original input $\boldsymbol{y}$ and the reconstructed input $\tilde{\mathbf{y}}$ ($\boldsymbol{D}\boldsymbol{x}$).

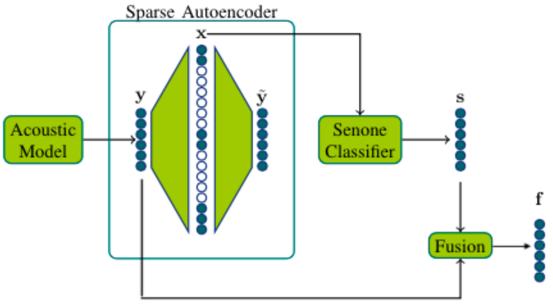

Figure 2: *The illustration of the proposed approach for word recognition, which gives promising improvements on WER as shown in Table 4.*

These insights motivate us to exploit the high-dimensional sparse representations for senone and phone classification. We train senone classifier S and phone classifier P to study how these high-dimensional sparse representations obtained by the sparse autoencoder can improve the senone classification and phone classification accuracy respectively. The senone classifier S can be described as:

$$\boldsymbol{s} = S(\boldsymbol{x}). \tag{4}$$

where $\boldsymbol{s} \in \mathbb{R}^{N_s}$ is the senone posteriors. Similarly, the phone classifier P can be described as:

$$\boldsymbol{p} = P(\boldsymbol{x}). \tag{5}$$

where $\boldsymbol{p} \in R^{N_p}$ is the phone posteriors.

Both S and P are trained using cross entropy (CE) criterion and consist of a simple fully connected layer with linear activation. As later explained in Section 3.3 and Section 3.4, both senone and phone classifier are observed to achieve higher frame-level senone and phone accuracy respectively, when compared to LF-MMI acoustic model. This motivates us to use senone likelihoods from the senone classifier for word recognition task.

Even though we manage to improve the senone classification performance with the senone classifiers, we only attain similar WER compared to LF-MMI acoustic model. Therefore, as illustrated in Fig. 2, we explore the weighted sum of the the outputs from LF-MMI acoustic model and senone classifier for word recognition:

$$\boldsymbol{f} = (1 - \gamma)\boldsymbol{y} + \gamma \boldsymbol{s}. \tag{6}$$

where $\gamma \in [0, 1]$, we get promising improvement on WER as stated in Section 3.5.

3. Experimental Setup and Results

In this section, we present the experimental setup and results in detail.

3.1. Baseline system

The experiments are conducted on AMI corpus [14] which contains recordings of spontaneous conversations in meeting scenarios in English. The corpus provides audio recordings from close-talk (stated as IHM) and far-field (stated as SDM) microphones. Both close-talk and far-field speech streams have been recorded in parallel. The dataset is available at 16kHz sampling rate with nearly 100 hours of meeting recordings divided approximately as 81 hours train set, 9 hours development and 9 hours evaluation set.

Two acoustic models are trained using IHM and SDM dataset with the LF-MMI criteria [4] using the standard chain model recipe of the Kaldi speech recognition toolkit [15]. The input was 40 dimensional high resolution MFCC features and 100 dimensional i-vectors. The output of the systems was the pseudo-log-likelihoods with N_y = 3776. The recognition performances for IHM and SDM are 19.0% and 40.2% respectively, as shown in Table 4.

3.2. Sparse autoencoder

We use shallow, overcomplete autoencoder with linear activation, tied weights, no bias and ℓ_1 penalty on the hidden unit activations. Here, $\mathbf{y}$ denotes a frame of LF-MMI senone likelihoods with N_y = 3776, $\mathbf{x}$ is the frame of high-dimensional representation with N_x = 11328, $\mathbf{D}$ represents the decoder weights of the autoencoder, and $\mathbf{Dx}$ the frame of reconstructed senone likelihoods $\tilde{\mathbf{y}}$.

The autoencoders are implemented in Pytorch [16] and trained using stochastic gradient descent with learning rate 0.1. For λ, grid search is performed on [10^{-1}, 10^{-6}]. The hyperparameter λ=10^{-4} provides the optimal performance both on IHM and SDM.

To measure the sparsity level of the frame-wise high-dimensional representations, we use the ℓ_0 norm as metric. In Table 1, under IHM column, the sparseness level on IHM evaluation set for the representations obtained from the autoencoders trained on IHM is presented. Same notation also applies for SDM. Overcomplete AE ($\lambda = 0$) has no sparsity constraint on the hidden unit activations. This concludes that sparse autoencoders are indeed capable of producing sparse representations.

Architecture	IHM	SDM
Overcomplete AE ($\lambda = 0$)	11300	11300
Sparse AE ($\lambda = 10^{-4}$)	7000	5000

Table 1: *The mode of the distribution of the number of non-zero elements in the frame-wise high-dimensional representations of size 11328.*

3.3. Senone classifier

We expect the senone subspaces to spread in the common high-dimensional sparse representation space, as shown in Fig. 1. Thus, we expect this behaviour to help separability of different senone classes.

As shown in Fig. 2, we train the senone classifier using the frame-wise high-dimensional representations from the sparse autoencoder as input and get the frame-wise senone posteriors $\mathbf{s}$ with N_s = 3776 on evaluation set. The forced alignments are taken as the ground truth. To make a fair comparison, we train another senone classifier using frame-wise senone likelihoods from LF-MMI acoustic model to assess how much of the improvement comes from neural network training.

Architecture	IHM	SDM
LF-MMI acoustic model	75.3	48.0
Senone classifier (LF-MMI likelihoods)	76.7	52.6
Senone classifier	76.5	52.8

Table 2: *The frame-level senone classification accuracies on IHM and SDM evaluation sets.*

The classifiers are implemented in Pytorch and trained with cross-entropy criterion using stochastic gradient descent with learning rate 0.1. As shown in Table 2, the senone classifier trained on high-dimensional sparse representations from sparse autoencoder achieves the highest accuracy for SDM, which contains more noise and reverberation with respect to the IHM dataset.

3.4. Phone classifier

After observing that high-dimensionality and sparsity is capable of improving the frame-level senone accuracy, we examine their usability for phone classification task. Our hypothesis is that high-dimensional sparse representation space provides even more opportunity for the low-level speech characteristics (i.e., phone related information, articulatory information) to spread while suppressing the impact of noise. This is expected to help classifying phones more easily compared to senones.

We train the phone classifier using the frame-wise high-dimensional representations obtained through the sparse autoencoder as input and get frame-wise phone posteriors $\mathbf{p}$ with $N_p = 165$. Similarly, we train another phone classifier using the frame-wise senone likelihoods from LF-MMI acoustic model.

Architecture	IHM	SDM
LF-MMI acoustic model	79.1	51.8
Phone classifier (LF-MMI likelihoods)	82.7	58.0
Phone classifier	82.4	64.0

Table 3: *The frame-level phone classification accuracies on IHM and SDM evaluation sets.*

The phone classifier is trained using the same criteria as the senone classifier. The performance gap between the baseline (i.e, LF-MMI acoustic model) and our phone classifier on SDM confirms that our hypothesis about the representational power of the high-dimensional sparse space is accurate.

3.5. Fusion of likelihoods for word recognition

Even though we obtain improvements in terms of frame-level phone and senone accuracies, we only attain comparable results when the high-dimensional sparse representations are exploited for word recognition. We stipulate that this is due to the decoder being finely tuned for the LF-MMI output, and we plan to further investigate this in future.

However, inspired from CE-smoothing [17], we also explore the weighted combination of the LF-MMI acoustic model and the senone classifier outputs. The senone posteriors from senone classifier are scaled with the priors to obtain the senone likelihoods and projected on the log space. We then pass the weighted combination of the senone likelihoods from the LF-MMI acoustic model and senone classifier to the decoder. Different combinations are examined and only the best performer (with $\gamma = 0.1$) which gives promising improvements on WER is reported in Table 4.

Architecture	IHM	SDM
LF-MMI acoustic model	19.0	40.2
Fusion of likelihoods ($\gamma = 0.1$)	18.9	39.8

Table 4: *The recognition performance (in WER%) for LF-MMI system and the proposed approach on IHM and SDM evaluation sets.*

4. Conclusion

In this paper, we focused on the use of shallow sparse autoencoders to produce more reliable phone or senone likelihoods. The sparse autoencoders were shown to be capable of projecting the senone likelihoods obtained from the baseline LF-MMI system to a high-dimensional sparse space. Our analysis on the high-dimensional sparse representations presented improvement in frame-level phone and senone classification accuracy. In particular, the highest improvements were seen on SDM dataset which is more noisy compared to IHM. To exploit the high-dimensional sparse representations for ASR, we used the frame-wise senone likelihoods obtained from the senone classifier which was trained on the high-dimensional sparse representations. When we took weighted combination of the senone likelihoods from the LF-MMI acoustic model and the senone classifier, we obtained promising improvements on WER for both IHM and SDM.

5. Acknowledgements

This work was funded by the Swiss National Science Foundation under the project Sparse and HIerarchical Structures for Speech Modeling (SHISSM).

6. References

[1] T. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep convolutional neural networks for LVCSR,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8614–8618.

[2] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] H. Sak, A. W. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” 2014.

[4] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for asr based on lattice-free mmi,” in *Proceedings of Interspeech*, 2016, pp. 2751–2755.

[5] M. Katz, H.-G. Meier, H. Dolfing, and D. Klakow, “Robustness of linear discriminant analysis in automatic speech recognition,” in *Object recognition supported by user interaction for service robots*, vol. 3, 2002, pp. 371–374.

[6] P. Dighe, A. Asaei, and H. Bourlard, “Low-rank and sparse soft targets to learn better dnn acoustic models,” in *Proceedings of 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)*, 2017.

[7] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online dictionary learning for sparse coding,” in *Proceedings of the 26th International Conference on Machine Learning*, ACM. ACM Press, 2009, pp. 689–696.

[8] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[9] E. J. Candès and M. B. Wakin, “An introduction to compressive sampling,” *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[10] B. Olshausen and D. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, no. 6583, p. 607, 1996.

[11] M. Kramer, “Autoassociative neural networks,” *Computers & chemical engineering*, vol. 16, no. 4, pp. 313–328, 1992.

[12] S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, 1998.

[13] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[14] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, “The ami meeting corpus,” in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005, p. 100.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” in *Proceedings of the IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[16] A. Paszke, S. Gross, S. Chintala, and G. Chanan, “Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration,” 2017.

[17] S. Wiesler, P. Golik, R. Schlüter, and H. Ney, "Investigations on sequence training of neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4565–4569.