INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Unsupervised Deep Auditory Model Using Stack of Convolutional RBMs For Speech Recognition

Hardik B. Sailor and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

{sailor_hardik, hemant_patil}@daiict.ac.in

Abstract

Recently, we have proposed an unsupervised filterbank learning model based on Convolutional RBM (ConvRBM). This model is able to learn auditory-like subband filters using speech signals as an input. In this paper, we propose two-layer Unsupervised Deep Auditory Model (UDAM) by stacking two ConvRBMs. The first layer ConvRBM learns filterbank from speech signals and hence, it represents early auditory processing. The hidden units' responses of the first layer are pooled as short-time spectral representation to train another ConvRBM using greedy layer-wise method. The ConvRBM in second layer trained on spectral representation learns Temporal Receptive Field (TRF) which represent temporal properties of the auditory cortex in human brain. To show the effectiveness of the proposed UDAM, speech recognition experiments were conducted on TIMIT and AURORA 4 databases. We have shown that features extracted from second layer when added to filterbank features of first layer performs better than first layer features alone (and their delta features as well). For both databases, our proposed two-layer deep auditory features improve speech recognition performance over Mel filterbank features. Further improvements can be achieved by system-level combination of both UDAM features and Mel filterbank features.

Index Terms: Convolutional RBM, filterbank, Temporal Receptive Field (TRF), speech recognition.

1. Introduction

Representation learning is a type of deep learning where features from the raw data can be learned by the underlying model with several layers of nonlinearities [1]. Unsupervised representation learning is the most important form of learning since many of the learning tasks in humans is unsupervised in nature such as language acquisition by the infants [2]. Features for various cognitive tasks such as vision and hearing in human are not present from infant stage instead they are learned from the experience [3]. Features based on human auditory processing perform better for various speech processing applications including speech recognition in clean and noisy conditions [4], [5]. Auditory processing includes the modeling of various processing stages in human ear (also called as early auditory processing [6]) and processing the auditory nerve signals in auditory cortex [4]. The methods for auditory modeling are based on computational models and data-driven approaches. There are several data-driven approaches for early auditory modeling [7–10] including our recently proposed work in [11]. We have proposed single layer unsupervised learning model Convolutional Restricted Boltzmann Machine (ConvRBM) to learn filterbanks directly from the speech signals. The computational auditory models for early auditory and auditory cortex are discussed in [12].

Several supervised deep learning methods were applied on speech signals to learn features and acoustic model jointly [13–17]. Earlier stacks of ConvRBM (called as a Convolutional Deep Belief Network) with sigmoid units as an unsupervised learning model was applied to spectrograms to learn higher-level temporal modulation features [18]. We have proposed single layer ConvRBM with rectified linear units (ReLU) to learn temporal modulation features from Mel spectrograms [19]. Our both the works using ConvRBM reported in [11], [19] are single layer models.

In this paper, we propose to use our recent work of filterbank learning in [11] and receptive field learning in [19] by stacking two ConvRBM as a deep auditory model. In addition, proposed ConvRBM-based approach is also shown to be stable under additive noise which aids its robustness under signal degradation conditions. Our work has strong similarity with recently proposed acoustic modeling framework from the raw speech signals using Convolutional Neural Networks (CNN) [16]. Compared to the work in [16], our generative deep model is unsupervised in nature and can be scalable to variable length inputs. It can be used to learn features as a front-end for supervised deep models as back-end. Our deep model with two convolution stages is related to computational auditory model [20] and deep scattering spectrum [21] where the filters are not learned from data. The ASR experiments on TIMIT and AURORA 4 databases shows that features from our proposed deep model perform better than handcrafted Mel filterbanks.

2. Architecture of Convolutional RBM for auditory processing

Convolutional RBM (ConvRBM) is a probabilistic unsupervised learning model with two layers, namely, visible layer and hidden layer [22]. We first describe the ConvRBM to model *1-D* signals such as speech which can easily be extended to ConvRBM with different subbands (i.e., *2-D* input). The input to the visible layer ($\mathbf{v}$) is an entire signal of length n_V. Hidden layer is divided into K number of groups of $n_H - D$ array (where $n_H = n_V - n_W + 1$ length of 'valid' convolution). Visible and hidden layer ($\mathbf{h}$) is connected with *K* number of weights denoted as $\mathbf{W}$ each of n_W-dimensional. Weights $\mathbf{W}$ between visible and hidden units are shared among all the locations. Biases are also shared in the hidden layer and visible layer denoted as b_k and c, respectively. If we denote b_k as the hidden bias for k^{th} group, then response of the convolution layer for the k^{th} group is given as:

$$I_k = (\mathbf{v} * \tilde{\mathbf{W}}^k) + b_k, \quad (1)$$

http://dx.doi.org/10.21437/Interspeech.2016-812

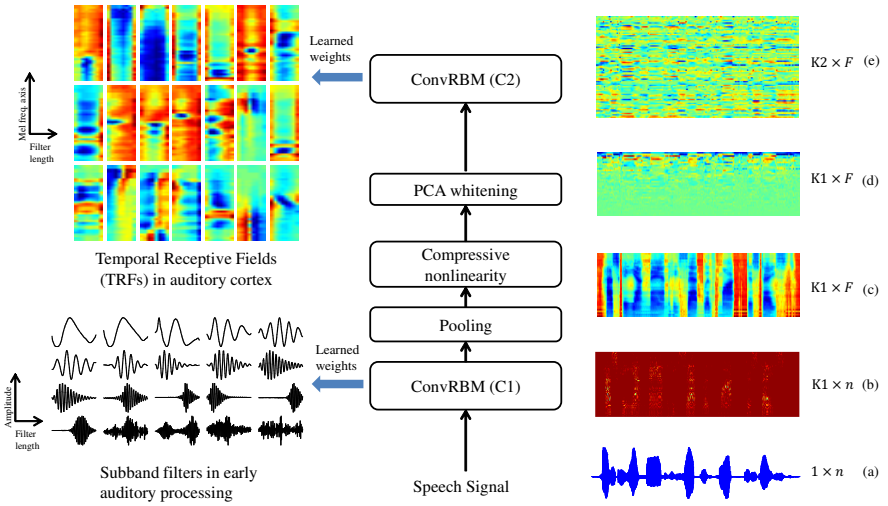

Figure 1: Block diagram of proposed UDAM using ConvRBMs. (a) Speech signal, (b) learned subband features of C1, (c) pooled subband signals followed by compressive nonlinearity, (d) PCA whitening and (e) learned modulation representation.

where $*$ denotes 'valid' length convolution operation and $\tilde{\mathbf{W}}^k$ denotes flipped array (for convolution operation) [22]. The energy function for ConvRBM is defined as [18],

$$E(\mathbf{v},\mathbf{h}) = \frac{1}{2}\sum_{i=1}^{n_V} v_i^2 - \sum_{k=1}^{K}\sum_{j=1}^{n_H}\sum_{r=1}^{n_W} h_j^k W_r^k v_{j+r-1} - \sum_{k=1}^{K} b_k \sum_{j=1}^{n_H} h_j^k - c\sum_{i=1}^{n_V} v_i. \quad (2)$$

The joint probability distribution (PDF) in terms of this energy function is given by [22]:

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp(-E(\mathbf{v},\mathbf{h})). \quad (3)$$

We have used rectified linear units (ReLU) in hidden layer of ConvRBM. The sampling from hidden units is performed using noisy ReLU as done in [23]. Sampling equations for hidden and visible units are given as [11]:

$$\mathbf{h}^k \sim max(0, I_k + N(0,\sigma(I_k))),$$
$$\mathbf{v} \sim \mathcal{N}\left(\sum_{k=1}^{K}(\mathbf{h}^k * \mathbf{W}^k) + c, 1\right), \quad (4)$$

where $N(0,\sigma(I_k))$ is a Gaussian noise with zero-mean and sigmoid of I_k as it's variance. During feature extraction stage (i.e., testing stage), we have used deterministic version of ReLU activation $max(0, I_k)$. Second ConvRBM is stacked on top of first ConvRBM to model *2-D* time-frequency representation (i.e., subband filterbank) obtained from first ConvRBM. The sampling equation for second ConvRBM can similarly be written from eq. (4). Both ConvRBMs are trained using single-step Contrastive Divergence (CD) [24] in greedy layer-wise manner. Let C1 and C2 denote ConvRBMs for first and second layer, respectively. The block diagram of our proposed UDAM architecture is shown in Figure 1.

2.1. ConvRBM to model speech signal

The input to ConvRBM (C1) is an entire speech signal of length *n*-samples. Weights of C1 with length *m1*-samples in each are also called as *subband* filters with respect to speech perception mechanism in hearing [11]. Convolution with $K = K1$ subband filters decompose the speech signal into different subbands. Subbands are ordered according to center frequencies of subband filters. The output of C1 is pooled according to *25* ms window length and *10* ms window shift followed by compressive nonlinearity as shown in Figure 1 [11]. Let us denote this short-time spectral representation as $\mathbf{y}$ which is of $K1 \times F$ dimensional (where *F* is the number of frames).

2.2. ConvRBM to model subband filterbank

The input $\mathbf{v}$ to the ConvRBM C2 is $\mathbf{y}$ which is a time-frequency representation of speech with $K1$ subbands and $n_V = F$ frames pooled from C1 responses. Before passing the input to C2, Principal Component Analysis (PCA) as a whitening transform is applied on $\mathbf{y}$ (as shown in Figure 1). Whitening the data using PCA gives approximation to sub-cortical processing which was observed in auditory cortex [25]. The weights of C2 are having length $m2$ number of frames. Pooling is not performed after C2 since we want to keep same number of frames to use as features with C1. The hidden layer has $K = K2$ groups which is two times overcomplete (i.e., $K2 = 2K1$). Hence, if $K1 = 40$ subbands then $K2 = 80$ groups in C2 resulting in *120-D* feature representation (we kept this to compare standard *120-D* Mel filterbank with *40* filters and their delta features). The summary of notations and corresponding configurations for the both layers are given in Table 1.

Table 1: *Notations of UDAM architecture*

ConvRBM	input $\mathbf{v}$	n_V	n_W	K
C1	speech	n samples	$m1$ samples	$K1$
C2	filterbank	F frames	$m2$ frames	$K2$

3. Analysis of the proposed model

In this Section, we analyze both ConvRBMs in terms of what it learns from the data.

3.1. Analysis of first layer C1

As we have shown in our recent work [11], the first ConvRBM C1 learns auditory-like filterbank when trained on speech signals. Hence, C1 may represent early auditory processing with subband filters (as shown in Figure 1) and half-wave rectified nonlinearity (i.e., ReLU). The feature representation steps involved in this ordering resemble the simplified form of auditory processing in the human ear [6], [12].

3.2. Analysis of second layer C2

The weights learned in C2 are visualized by applying inverse of PCA whitening on the C2 weights. Since convolution is applied in temporal-domain (for each subbands), patches of subband filters represent Temporal Receptive Fields (TRFs) [18]. Examples of TRFs learned on AURORA 4 database are shown in Figure 1 where each block represents one TRF. ConvRBM subband filters capture temporal modulation information with different *subband modulation frequencies* from first layer filterbank. Detailed analysis of TRFs is presented in [18] and [19]. Each subband filter represent temporal variations in different phonetic units similar as delta features ($\Delta + \Delta\Delta$) of filterbanks.

3.3. Stability of convolutional network to additive noise

Since ConvRBM C1 follows Mel scale as shown in [11], it can be proved that ConvRBM is also stable to deformations in the speech signals [21]. Here, we will discuss the stability w.r.t. additive noise. Let T be the transformation applied on input $\mathbf{v}$. For T to be stable to additive noise $\hat{\mathbf{v}} = \mathbf{v} + \epsilon$, Lipschitz continuity condition needs to be satisfied for constant $\lambda > 0$ which is given as [26],

$$\|T\mathbf{v} - T\hat{\mathbf{v}}\|_2 \leq \lambda \|\mathbf{v} - \hat{\mathbf{v}}\|_2 \tag{5}$$

This condition is proved for scattering convolutional networks [21] and recently for supervised CNN with certain criteria such as max-norm regularization for weights, ReLU nonlinearity and max-pooling [27]. Our model also has convolution and ReLU stages and hence, we can prove that ConvRBM is also stable to the additive noise. The stability conditions are discussed for first layer C1. However, can easily be extended for the second layer.

3.3.1. Stability of convolution in ConvRBM

The transformation T for convolution operation in ConvRBM for k^{th} group is $T\mathbf{v} = \mathbf{v} * \tilde{\mathbf{W}}^k$. In [27], weights are max-norm regularized to obtain stability criteria. ConvRBM training includes weight decay which penalizes the weights to be small and smooth. For TIMIT and AURORA 4 databases, we have observed that for C1 layer $\|\mathbf{W}\|_1 \leq 3$ and $\|\mathbf{W}\|_1 \leq 2.5$, respectively. Hence, based on derivation in [27] for convolution operation in ConvRBM, following stability condition holds:

$$\left\|\mathbf{v} * \tilde{\mathbf{W}}^k - \hat{\mathbf{v}} * \tilde{\mathbf{W}}^k\right\|_2 \leq \lambda \|\mathbf{v} - \hat{\mathbf{v}}\|_2 , \tag{6}$$

where λ=*3* for TIMIT and λ=*2.5* for AURORA 4 database.

3.3.2. Stability of rectified nonlinearity

As discussed in Section 2, we have used deterministic ReLU for feature extraction. It is proved in [27] that ReLU operation is also stable with λ=1. The stability condition for ConvRBM with response I_k for clean and $\hat{I_k}$ for additive noise is given as,

$$\left\|max(I_k, 0) - max(\hat{I_k}, 0)\right\|_2 \leq \left\|I_k - \hat{I_k}\right\|_2 . \tag{7}$$

The stability of ConvRBM to additive noise resulted in improved performance in AURORA 4 speech recognition task even though the subband filters are learned from data.

4. Experimental setup

The ASR experiments were performed with clean and multicondition training database described as follows:

4.1. Speech databases

4.1.1. TIMIT database

For phone recognition task, TIMIT database was used [28]. In TIMIT database, all SA category sentences (same sentences spoken by all the speakers) were removed as they may bias the speech recognition performance. Training data contains utterances from *462* speakers. Development and test set contains utterances from *50* and *24* speakers, respectively.

4.1.2. AURORA 4 database

We have also used AURORA 4 database created using six different types of additive noises, namely, car, crowd of people (babble), restaurant, street, airport and train station [29]. Multicondition training data consists of *7138* utterances from WSJ0 database with half of them recorded with the Sennheiser microphone and the other half recorded with the second microphone. The *14* test sets each with *330* utterances, are grouped into four categories, namely, A: clean (set 1), B: noisy (set 2 to set 7), C: clean+mismatch (set 8) and D: noisy+mismatch (set 9 to 14).

4.2. Training of ConvRBMs and feature representation

The training parameters for C1 is same as that of used in our recent work in [11]. Training method of C2 is different than the one used in [19]. For C2, the learning rate was chosen to be *0.005* which is fixed for first *20* epochs and decayed later. Compared to work in [18] and [19], we have not used sparsity regularization since our model uses ReLUs which provides sparsity in the hidden units (forcing negative activations to zero). Weights are regularized using weight decay with a factor of *0.0001*. To have a fair comparison with standard *120-D* FBANK features, we restrict ourselves to *40-D* filterbank in C1 and *80-D* features in C2 giving *120-D* feature vector. The notations for different features are given in Table 2.

Table 2: *Notations of different features used in this study*

Description	Notation of features
Mel filterbank with delta features	FBANK (*120-D*)
Filterbank from C1	C1 (*40-D*)
Modulation features from C2	C2 (*80-D*)
Feature fusion of C1 and C2	C1+C2 (*120-D*)

4.3. ASR system building

Monophone GMM-HMM systems were built using *39-D* MFCC features for both the databases to generate force-aligned labels. MFCC features were extracted from windowed speech signal with *25* ms length and *10* ms shift similar as parameters of pooling after C1. For TIMIT database, *48* phones were used for training and mapped to *39* phones during scoring as done in [30]. Language modeling is performed using bi-gram for

TIMIT and tri-gram for AURORA 4. In this paper, all ASR systems were built using KALDI speech recognition toolkit [31]. Hybrid DNN-HMM systems were built using fast implementation of p-norm DNNs with $p = 2$ [32] (different to our recent works [11] and [19] where we have used vanilla DNN). ASR system combination (denoted as $\oplus$ symbol) is performed using the minimum Bayes risk decoding [33].

5. Experimental Results

5.1. Results on TIMIT database

The parameters of C1 layer is same as tuned in [11] with filter length $m1$=*128* samples. To analyze the significance of second layer C2, we have compared the performance of single layer C1 filterbank with feature fusion of C1 and C2. The results of these experiments are reported in Table 3 in % Phone Error Rate (PER) using hybrid p-norm DNN with parameters: *2000* hidden units, group size of *5*, *2* hidden layers and *9* frame context window. From Table 3, we can see that by adding delta features in filterbank features extracted from C1, we obtained small relative improvement of *0.9 %* compared to C1 features. The second layer features C2 alone perform better or comparable to C1 along with their deltas. The filter length of *8* frames in C2 works better then *6* and *10* frames when added to C1 features. It gives relative improvement of *3.6 %* compared to only C1 features and *2.73 %* compared to the C1 along with their delta features. The FBANK features are compared with deep features C1+C2 using the same hybrid p-norm DNN of *3* hidden layers in Table 4. C1+C2 features gives relative improvement of *5.36 %* (*1.2 %* absolute) on development set and *2.56 %* on test set compared to FBANK features. System combination improve performance on test set only which is *5.13 %* relative to FBANK.

Table 3: *% PER for Comparison of filter length in C2 and comparison with first layer features on TIMIT development set*

ConvRBM features	Filter length in C2 (m2)	Dev
C1 (40-D)	-	22.2
C1+Δ + $\Delta\Delta$ (120-D)	-	22.0
C2 (80-D)	6	22.0
C2 (80-D)	8	21.8
C1+C2 (120-D)	6	22.1
C1+C2 (120-D)	8	**21.4**
C1+C2 (120-D)	10	21.8

Table 4: *Results on TIMIT database in % PER.*

Features	Dev	Test
A: FBANK (120-D)	22.4	23.4
B: C1+C2 (120-D)	21.2	22.8
A $\oplus$ B	**21.2**	**22.2**

5.2. Results on AURORA 4 database

Similarly to that of TIMIT database, we have shown the results of varying the length of second layer filter and compared it with first layer features C1. The % Word Error Rate (WER) of these experiments are reported in Table 5 using hybrid p-norm DNN with parameters: *2000* hidden units, group size of *5*, *2* hidden layers and *9* frame context window. For C2, filter length of *10* frames perform better compared to *8* frames in TIMIT. Use of second layer features C2 improves performance compared to C1 features alone as well as addition of delta features in C1. Specifically, for multicondition test sets (i.e., C and D), using C1+C2 features, there is an absolute reduction of *1.12-1.42* in % WER over C1 and *0.63-1.22 %* absolute reduction in WER over C1 along with delta features. Finally, the C1+C2 features are compared to FBANK features in Table 6 with *3* layer p-norm DNN. Absolute reduction of *1-2 %* in WER is obtained using C1+C2 features compared to FBANK features. The fusion of both C1 and C2 features perform better than C1 and C2 features alone. System combination further reduce WER (except in test set A) with significant reduction for test sets C and D compared to FBANK and C1+C2 features. Hence, both features contain complementary information. The improvements in AURORA 4 task can be justified based on stability of ConvRBM to additive noise as discussed in Section 3.3.

Table 5: *Comparison of filter length in C2 for AURORA 4 database in % WER. Dimensionality of features are same as denoted in Table 2*

Features	A	B	C	D	Avg
C1	9.36	17.90	22.64	34.66	21.14
C1+Δ + $\Delta\Delta$	9	17.05	22.44	33.19	20.42
C1+C2, m2=6	9.25	17.18	22.08	33.29	20.45
C1+C2, m2=8	**8.91**	17.25	22.17	33.47	20.45
C1+C2, m2=10	9.1	**16.97**	**21.22**	**32.54**	**19.96**
C2, m2=10	8.87	18.37	23.48	34.4	21.28

Table 6: *Results on AURORA 4 database in % WER*

Features(120-D)	A	B	C	D	Avg
A: FBANK	10.41	18.16	22.45	34.09	21.28
B: C1+C2	**8.37**	16.89	20.96	33.04	19.82
A $\oplus$ B	8.56	**16.14**	**19.73**	**32.07**	**19.12**

6. Summary and conclusions

Unsupervised deep auditory model (UDAM) is proposed to model human auditory processing by stacking two ConvRBMs. First layer ConvRBM learns filterbank from speech signals and hence, represent early auditory processing. Second layer ConvRBM learns temporal receptive fields and hence, represent model of auditory cortex. We have shown that ConvRBM is stable to additive noise using Lipschitz continuity condition. Significance of features of both layers is verified using ASR experiments on clean and multicondition databases. The second layer modulation features perform better when added to filterbank compared to delta features. Feature fusion of both layers perform better compared to Mel filterbanks for both TIMIT and AURORA 4 databases. Both learned features and handcrafted features contain complementary information resulting in further reduction of error rates using system-level combination. Our future work is to extend the second layer ConvRBM to learn Spectro-Temporal Receptive Fields (STRFs). We would also like to perform detailed mathematical analysis of UDAM including stability to deformations.

7. Acknowledgments

The authors would like to thank Dept. of Electronics and Information Technology (DeitY), Govt. of India for sponsoring two consortium projects, (1) TTS Phase II (2) ASR Phase II and authorities of DA-IICT, Gandhinagar.

8. References

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[3] G. Hinton, "Where do features come from?" *Cognitive Science*, vol. 38, no. 6, pp. 1078–1101, 2014.

[4] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 34–43, Nov 2012.

[5] H. Hermansky, J. Cohen, and R. Stern, "Perceptual properties of current speech recognition technology," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1968–1985, Sept 2013.

[6] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, March 1992.

[7] J. Lee, H. Jung, T. Lee, and S. Lee, "Speech feature extraction using independent component analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1631–1634.

[8] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[9] A. Bertrand, K. Demuynck, V. Stouten, and H. V. hamme, "Unsupervised learning of auditory filter banks using non-negative matrix factorisation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Las Vegas, Nevada*, 2008, pp. 4713–4716.

[10] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted Boltzmann machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5884–5887.

[11] H. B. Sailor and H. A. Patil, "Filterbank learning using convolutional restricted Boltzmann machine for speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016*, Shanghai, China, March 2016, pp. 5895–5899.

[12] R. M. Stern and N. Morgan, "Features based on auditory physiology and perception," in *Techniques for Noise Robustness in Automatic Speech Recognition*. T. Virtanen, B. Raj, and R. Singh, (Eds.) John Wiley and Sons, Ltd, New York, NY, USA, 2012, pp. 193–227.

[13] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2013, pp. 297–302.

[14] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *INTERSPEECH*, Singapore, Sept. 2014, pp. 890–894.

[15] D. Palaz, M. Magimai-Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *40^{th} International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD*, 2015, pp. 4295–4299.

[16] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *INTERSPEECH*, Dresden, Germany, 6-10 Sept. 2015, pp. 26–30.

[17] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, Dresden, Germany, 6–10 Sept 2015, pp. 1–5.

[18] H. Lee, P. T. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *23^{rd} Annual Conference on Neural Information Processing Systems, Canada, 7-10 December*, 2009, pp. 1096–1104.

[19] H. B. Sailor and H. A. Patil, "Unsupervised learning of temporal receptive fields using convolutional RBM for ASR task," accepted in European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 29 Aug - 2 Sept. 2016.

[20] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America (JASA)*, vol. 118, no. 2, pp. 887–906, 2005.

[21] J. Anden and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.

[22] H. Lee, R. B. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26^{th} Annual International Conference on Machine Learning, (ICML), Canada, June 14-18*, 2009, pp. 609–616.

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[24] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[25] A. M. Saxe *et al.*, "Unsupervised learning models of primary cortical receptive fields and receptive field plasticity," in *25th Annual Conference on Neural Information Processing Systems (NIPS), 12-14 December, Granada, Spain.*, 2011, pp. 1971–1979.

[26] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, Aug. 2013.

[27] R. Yeh, M. H. Johnson, and M. N. Do, "Stable and symmetric filter convolutional neural network," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016*, Shanghai, China, March 2016, pp. 2652–2656.

[28] Garofolo *et al.*, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993.

[29] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Ph.D. dissertation, Mississippi State University, 2002.

[30] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. on Acousti., Speech and Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[31] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU), Hawaii*, 11-15 Dec. 2011.

[32] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 215–219.

[33] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802 – 828, 2011.