INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Speech Representation Learning Using Unsupervised Data-Driven Modulation Filtering for Robust ASR

Purvi Agrawal and Sriram Ganapathy

Learning and Extraction of Acoustic Patterns Lab, Dept. of Electrical Engg.,
Indian Institute of Science, Bengaluru-560012, India.
`(purvi_agrawal, sriram)@ee.iisc.ernet.in`

Abstract

The performance of an automatic speech recognition (ASR) system degrades severely in noisy and reverberant environments in part due to the lack of robustness in the underlying representations used in the ASR system. On the other hand, the auditory processing studies have shown the importance of modulation filtered spectrogram representations in robust human speech recognition. Inspired by these evidences, we propose a speech representation learning paradigm using data-driven 2-D spectro-temporal modulation filter learning. In particular, multiple representations are derived using the convolutional restricted Boltzmann machine (CRBM) model in an unsupervised manner from the input speech spectrogram. A filter selection criteria based on average number of active hidden units is also employed to select the representations for ASR. The experiments are performed on Wall Street Journal (WSJ) Aurora-4 database with clean and multi condition training setup. In these experiments, the ASR results obtained from the proposed modulation filtering approach shows significant robustness to noise and channel distortions compared to other feature extraction methods (average relative improvements of 19% over baseline features in clean training). Furthermore, the ASR experiments performed on reverberant speech data from the REVERB challenge corpus highlight the benefits of the proposed representation learning scheme for far field speech recognition.

Index Terms: unsupervised learning, data-driven modulation filtering, convolutional restricted Boltzmann machine, speech recognition.

1. Introduction

Inspite of recent advances in deep learning, the development of speech recognition systems in noisy and reverberant environments continues to be a challenging task. However, the human auditory system exhibits remarkable robustness to many of these environmental artifacts. While the early processing stages in the auditory periphery are mimicked in speech feature extraction for automatic speech recognition (ASR) systems [1, 2, 3], the recent evidence from human auditory system reveal that the inherent robustness may be primarily attributed to the spectro-temporal filtering performed by cortical neurons [4, 5, 6].

For ASR, several studies have attempted incorporating the knowledge of spectro-temporal filters (for example, Gabor filtering [7, 8]). In general, these approaches define a series of spectral (scale), temporal (rate), and spectro-temporal modulation filters that can be seen as modeling the tuning of cortical neurons to different spectro-temporal patterns. For the ASR application, use of temporal modulations such as RASTA filtering [9], TRAPS [10] and HATS [11] have been well studied. A supervised data driven approach for deriving temporal modulation filters using the linear discriminant analysis (LDA) is explored in [12]. Also, a recent approach to separable spectro-temporal Gabor filter bank features is proposed in [13].

In this paper, we propose to use the speech spectrogram to learn the two dimensional (2-D) spectro-temporal modulation filters in an unsupervised data-driven paradigm. While a data-driven approach was previously attempted for learning the peripheral auditory filter bank [14, 15, 16], this work represents the first attempt for designing modulation filters in an unsupervised data-driven fashion to the best of our knowledge.

The proposed filter learning method is developed using convolutional restricted Boltzmann machine (CRBM) [17]. In particular, 2-D filters characterizing the distribution of spectro-temporal modulations are derived from large amount of unsupervised speech spectrogram data. In this approach, we do not apply any prior knowledge of the perceptual studies of auditory processing and allow the data to learn the key spectro-temporal modulation content present in the data. We also propose a rank-1 contraint on learning the 2-D filters using contrastive divergence in CRBM in order to have separable 2-D filters. After learning a filter, the projection of the input spectrogram on the learnt filter is removed and the residual spectrogram is then used in the CRBM framework for learning subsequent filters. Once a set of filters are derived, an unsupervised filter selection criterion is used and the input spectrogram is filtered using the selected modulation filters to derive features for ASR.

The ASR experiments are performed on the Wall street Journal (WSJ) Aurora-4 database with clean and multi condition training set up using a deep neural network (DNN) acoustic model. Further, the ASR experiments are performed on reverberant speech provided in the REVERB challenge [18]. The results from these experiments indicate that the features derived from proposed filters provide significant improvements over other noise robust front-ends. We also investigate the performance of the proposed features in a semi-supervised setting where availability of labeled data is limited.

The rest of the paper is organized as follows. In Sec. 2, we describe the data driven framework for learning 2-D modulation filters and the filter selection criterion. Sec. 3 describes the ASR experiments with the proposed front-end followed by the results. We conclude with a summary of the proposed front-end.

2. 2-D filter learning

2.1. Restricted Boltzmann machine

The restricted Boltzmann machine (RBM) [19] is a two-layer, undirected graphical model with a set of binary hidden units $\mathbf{h}$ (as output layer), a set of (binary or real-valued) visible units $\mathbf{v}$

This work was supported by Defence Research and Development Organization (DRDO), Government of India.

http://dx.doi.org/10.21437/Interspeech.2017-901

(as input layer), and symmetric connections between these two layers represented by a weight matrix $\mathbf{W}$. The energy function of the Gaussian RBM is given as:

$$E_1(\mathbf{v}, \mathbf{h}, \theta) = -\sum_{i,j} \mathbf{v}_i \mathbf{W}_{ij} \mathbf{h}_j - \sum_i \mathbf{b}_i \mathbf{v}_i - \sum_j \mathbf{c}_j \mathbf{h}_j \quad (1)$$

where i and j are indices that iterate over visible and hidden units, respectively, model parameters are $\theta = (\mathbf{W}, \mathbf{b}, \mathbf{c})$, with $\mathbf{b}$ and $\mathbf{c}$ being the bias at visible and hidden layer, respectively. The conditional probability model is given by:

$$P(\mathbf{h}_j = 1|\mathbf{v}) = \sigma(\mathbf{c}_j + \sum_i \mathbf{v}_i \mathbf{W}_{ij}) \quad (2a)$$

$$P(\mathbf{v}_i = 1|\mathbf{h}) = \sigma(\mathbf{b}_i + \sum_j \mathbf{W}_{ij} \mathbf{h}_j), \quad (2b)$$

where σ is the sigmoid function defined by $\sigma(x) = 1/(1 + exp(-x))$ and $\mathbf{W}_{ij}$ is the ijth element of $\mathbf{W}$. We use the contrastive divergence (CD) learning algorithm for RBM training [20] using gradient ascent based optimization procedure. With regard to visible-hidden weights, the one-step CD (Gibbs sampler) followed by weight update is given as:

$$\triangle_{\mathbf{W}_{ij}} J(\mathbf{W}, \mathbf{b}, \mathbf{c}; \mathbf{v}) =< \mathbf{v}_i \mathbf{h}_j >_{data} - < \mathbf{v}_i \mathbf{h}_j >_{model},$$

$$\mathbf{W}' = \mathbf{W} + \eta(\triangle_{\mathbf{W}} J), \quad (3)$$

where J is the log likelihood defined as the exponential of negative of E_1, $< . >$ denotes the expectation under the distribution specified by the subscript, $\mathbf{v}_i$ and $\mathbf{h}_j$ are the ith and jth elements of visible and hidden layer, respectively, $\mathbf{W}'$ is the updated $\mathbf{W}$ matrix, and η is the learning rate.

2.2. Convolutional RBM

The local characteristic of the signal is ignored by RBMs, so a given feature detected by weights must be learned separately for every location [21]. While RBMs learn to reconstruct and identify the features of each signal as a whole, convolutional neural networks (CNNs) learn the mapping to the targets using feature maps locally [22, 23]. The CNNs require supervised training data and typically operate on smaller contextual windows (11 frames). A convolutional operation can be added to RBM learning by weight sharing, reconstructing and identifying the features of the signal locally [17, 21].

A CRBM is a probabilistic model where hidden units $\mathbf{H}$ (dimension $N_{H_r} \times N_{H_s}$) represent the presence/absence of local features in subwindows of visible units $\mathbf{V}$ ($N_{V_r} \times N_{V_s}$) [24]. The joint energy function of CRBM is given as:

$$E_2(\mathbf{V}, \mathbf{H}, \theta) = -\sum_q \mathbf{H}_q(\mathbf{W} \odot \mathbf{V}_{(q)}) - \sum_i b\mathbf{V}_i - \sum_q c\mathbf{H}_q$$

Here, $\mathbf{W}$ is the weight matrix (filter) of dimension ($N_{W_r} \times N_{W_s} = (N_{V_r} - N_{H_r} + 1) \times (N_{V_s} - N_{H_s} + 1)$), $\mathbf{V}_{(q)}$ is subwindow of patch $\mathbf{V}$ with top left corner at unit q and with the dimensions same as that of $\mathbf{W}$, index q iterates over units of $\mathbf{V}$, $\odot$ denotes the dot product of matrices after linearizing its elements, $\theta = (\mathbf{W}, b, c)$ are the model parameters, $\mathbf{H}_q$ is the element of the matrix $\mathbf{H}$ at location q. The conditional probability model is given by:

$$P(\mathbf{H}_q = 1|\mathbf{V}) = \sigma((\mathbf{W} \odot \mathbf{V}_{(q)}) + c) \quad (4a)$$

$$P(\mathbf{V}_p = 1|\mathbf{H}) = \sigma((\mathbf{W}^\star \odot \mathbf{H}_{(p)}) + b), \quad (4b)$$

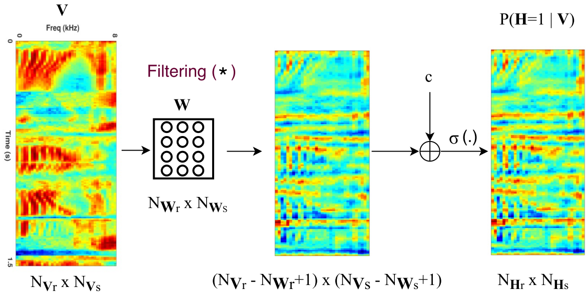
V
Freq (kHz)
0
8
Time (s)
0
1.5
Filtering (⋆)
W
$N_{\mathbf{W}_r}$ x $N_{\mathbf{W}_s}$
c
σ (.)
P(**H**=1 | **V**)
$N_{\mathbf{V}_r}$ x $N_{\mathbf{V}_s}$
($N_{\mathbf{V}_r}$ - $N_{\mathbf{W}_r}$+1) x ($N_{\mathbf{V}_s}$ - $N_{\mathbf{W}_s}$+1)
$N_{\mathbf{H}_r}$ x $N_{\mathbf{H}_s}$

Figure 1: *Block schematic of the proposed CRBM architecture for learning modulation filter* $\mathbf{W}$*) (forward pass of CRBM).*

where σ is the sigmoid function, $\mathbf{W}^\star$ is the horizontally and vertically flipped version of the original filter, $\mathbf{H}_{(p)}$ is subwindow of patch $\mathbf{H}$ with top left corner at unit p and size same as that of $\mathbf{W}$, index p iterates over units of $\mathbf{H}$, $\mathbf{V}_p$ is the element of the matrix $\mathbf{V}$ at location p. The one-step contrastive divergence (Gibbs sampler) approximation for CRBM is given by:

$$\triangle_{\mathbf{W}} J(\mathbf{V}; \theta) =< \mathbf{V} \star \mathbf{H} >_{data} - < \mathbf{V} \star \mathbf{H} >_{model} \quad (5)$$

where $\star$ is the 2-D filtering operation. The weight matrix is updated in an iterative learning process over several steps. The block schematic of the proposed modulation filter learning scheme from speech spectrogram through convolutional RBM (CRBM) is shown in Figure 1. The input layer $\mathbf{V}$ consists of a cell of 2-D patches sampled from speech spectrogram. Each 2-D patch consists of sub-band energy trajectory for 1.5 sec of speech along temporal dimension and an all-band energy trajectory along spectral dimension (40 bands) ($N_{V_r} = 150$, $N_{V_s} = 40$).

2.3. Rank-1 constraint on weight learning

To constrain the weight matrix $\mathbf{W}$ as a separable rank-1 matrix, we define $\mathbf{W}$ as the outer product of 1-D rate filter $\mathbf{r}$ and 1-D scale filter $\mathbf{s}$, i.e., $\mathbf{W} = \mathbf{r}\mathbf{s}^\top$. The gradient of J is computed with respect to $\mathbf{r}$ and $\mathbf{s}$ separately (unlike with respect to each element of $\mathbf{W}$). Let $\tilde{\mathbf{V}} = \mathbf{V} \star \mathbf{s}^\top$. The gradient ascent equation with respect to rate filter ($\mathbf{r}$) gives:

$$\triangle_{\mathbf{r}} J(\mathbf{V}; \theta) =< \tilde{\mathbf{V}} \star \mathbf{H} >_{data} - < \tilde{\mathbf{V}} \star \mathbf{H} >_{model} \quad (6)$$

where $\mathbf{s}$ is the scale filter obtained from previous iteration, $\mathbf{V}$ and $\mathbf{H}$ being the input 2-D patch and hidden activation patch, respectively. Let $\tilde{\mathbf{H}} = \mathbf{H} \star \mathbf{r}$. The gradient ascent equation with respect to scale filter ($\mathbf{s}$) gives:

$$\triangle_{\mathbf{s}} J(\mathbf{V}; \theta) =< \mathbf{V} \star \tilde{\mathbf{H}} >_{data} - < \mathbf{V} \star \tilde{\mathbf{H}} >_{model} \quad (7)$$

Hence, the filter update equations become:

$$\mathbf{r}' = \mathbf{r} + \eta(\triangle_{\mathbf{r}} J); \quad \mathbf{s}' = \mathbf{s} + \eta(\triangle_{\mathbf{s}} J) \quad (8)$$

where η is the learning rate. Subsequently, the 2-D filter $\mathbf{W}$ is updated as $\mathbf{W}' = \mathbf{r}'\mathbf{s}'^\top$. We perform several iterative steps to learn the 2-D filters and the filters are thus learnt purely from a generative modeling perspective.

2.4. Multiple filter learning and selection

In all our analysis, we find that the first 2-D filter learnt from the input mel spectrogram is invariably a low-pass in both rate and scale domain (Figure 2 (a) and 3 (a)). For learning multiple 2-D filters that are less redundant [25], we use the following approach. After an initial 2-D filter is learnt (we name it R1-S1), we remove the contribution of learnt rate component (R1) from the original spectrogram by subtracting the original

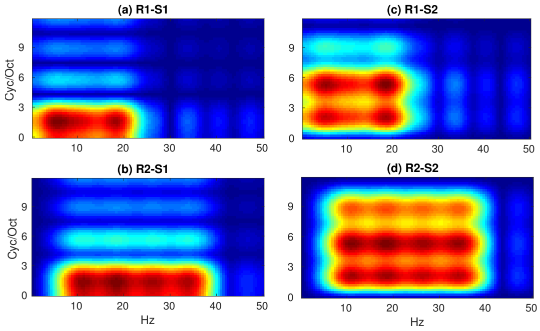

Figure 2: *The magnitude response of the proposed 2-D data driven filters (rank-1) obtained from clean mel spectrogram.*

spectrogram from the rate filtered spectrogram. This residual (containing the high rate and full scale information) is fed back to CRBM for learning next filter (R2-S1). Similarly, we remove the contribution of learnt scale component (S1) from the original spectrogram and the residual (containing the full rate and high scale information) is fed to CRBM for learning next filter (R1-S2). We also remove the contribution of both (R1) and (S1) from the original spectrogram for learning filter (R2-S2) from the residual. This method, similar to matching pursuit (MP) algorithm [26], allows us to learn irredundant set of filters. For the CRBM learning, the 2-D weight matrix $\mathbf{W}$ is initialized as the outer product of the 1-D rate and scale filters learnt from CRBM using corresponding 1-D inputs sampled from spectrogram. Figure 2 shows the magnitude response of the learnt 2-D rank-1 filters obtained from mel spectrogram of clean speech data of Aurora-4 corpus. Similarly, we learn the 2-D rank-1 filters from mel spectrogram of multi condition training data, shown in Figure 3. As seen here, deriving the filters using MP style algorithm provides irredundant 2-D filters.

In order to select 2-D filters for ASR (4 learnt from rank-1 and 4 learnt from full rank), the average number of active hidden units of the CRBM (with a total of 4488 hidden units for $N_{\mathbf{W}_r} \times N_{\mathbf{W}_s} = 15 \times 8$) is computed for each 2-D filter by a forward pass operation of the set of input spectrograms through the CRBM. The average active count is computed using $P(\mathbf{H}_q = 1|\mathbf{V})$ summed over all q units (count of active units for a given input) and averaged over a number of input patches from the validation data. Based on the highest average active count, the (R2-S1) and (R2-S2) filter with rank-1 constraint gives maximum average active units amongst all filters, as shown in Figure 4. Similar trend is observed for 2-D filters for multi condition training data. This criterion represents a data driven unsupervised approach to filter selection.

The features for ASR are derived using two streams of filtered spectrograms using the rank-1 filters (R2-S1 and R2-S2). These spectrogram streams are concatenated and fed to a DNN based ASR system. The input features are mean-variance normalized at utterance level before DNN training.

3. Experiments and results

The WSJ Aurora-4 corpus is used for conducting ASR experiments. This database consists of continuous read speech recordings of 5000 words corpus, recorded under clean and noisy conditions (street, train, car, babble, restaurant, and airport) at $10 - 20$ dB SNR. The training data has two sets of 7138 clean and multi condition recordings (84 speakers). The validation data has two sets of 1206 recordings (14 speakers) for clean and multi condition and test data has 330 recordings (8 speakers), each of the 14 clean and noise conditions. The test data is classified into group A - clean data, B - noisy data, C - clean data with channel distortion, and D - noisy data with channel distortion. The speech recognition Kaldi toolkit [27] is used for building the ASR. A deep belief network- deep neural network (DBN-DNN) with 4 hidden layers having 21 frames of input temporal context and a sigmoid nonlinearity is discriminatively trained using the training data and a tri-gram language model is used in the ASR decoding. We compare the ASR performance of the proposed modulation filtering approach with traditional mel filter bank energy (MF) features, power normalized filter bank energy (PN) features [28], advanced ETSI front-end (ET) [29] and RASTA features (RAS) [9].

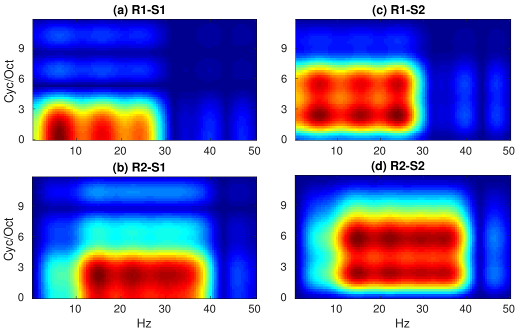

Figure 3: *The magnitude response of the proposed 2-D data driven filters (rank-1) obtained from mel spectrogram of multi condition training data.*

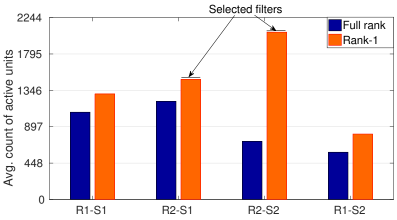

Figure 4: *The average count of active hidden units of CRBM model for full rank and rank-1 filters for clean training.*

The ASR performance in clean training condition is reported in Table 1. From this table, it can be observed that PN and ET features provide better performance compared to the MF and RAS features. The data driven modulation filtering approach on mel spectrogram provides significant improvement in noisy and channel distortion scenarios (average relative improvements of 19 % over MF features).

In the multi condition training and test scenario (reported in Table 2), the MF features perform better than all other baseline features. The proposed feature extraction improves the performance of ASR compared to the baseline features by average relative improvements of 9% over MF.

3.1. Reverberant speech recognition

The ASR experiments on reverberant speech data are performed using WSJCAM0 corpus in a single channel scenario, released as a part of REVERB challenge [18]. This database consists of 7861 recordings from 92 training speakers, 1488 recordings from 20 development test (dt) speakers and 2178 recordings from two sets of 14 evaluation test (et) speakers, with each speaker providing about 90 utterances. These recordings were carried out with two sets of microphone- head mounted as well as desk microphone positioned about half meter from the

Table 1: *Word error rate (%) in Aurora-4 database for clean training condition with various feature extraction schemes.*

Cond	MF	PF	ET	RA	Prop
A. Clean with same Mic					
Clean	3.4	3.3	**3.2**	3.5	**3.2**
B: Noisy with same Mic					
Airport	21.9	18.3	15.0	19.3	13.2
Babble	19.6	16.0	15.5	19.9	13.8
Car	8.0	6.2	9.8	7.9	5.7
Rest.	24.9	22.9	20.5	23.0	17.3
Street	19.5	17.8	19.5	18.7	15.6
Train	19.8	16.3	17.4	19.4	17.2
Avg.	18.9	16.2	16.3	18.0	**13.8**
C: Clean with diff. Mic					
Clean	15.3	**11.7**	14.5	16.0	13
D: Noisy with diff. Mic					
Airport	40.1	36.4	31.4	39.2	30.4
Babble	37.3	34.2	32.1	38.5	33
Car	24.9	21.5	24.9	24.8	19.3
Rest.	39.6	39.0	35.4	39.1	31.6
Street	35.7	34.1	35.0	35.8	31.8
Train	35.6	31.8	33.2	36.4	33.3
Avg.	35.2	32.8	32.0	35.6	**29.9**
Avg. of all conditions					
Avg.	24.7	22.1	21.9	24.4	**19.9**

speaker's head. The database consists of three subsets: training data set (Train) - for both clean and multi condition training using simulated reverb data, a simulated test dataset (Sim) and a naturally reverberant recording of the test dataset (Real). The 2-D rank-1 filters are learnt from mel spectrogram of Train dataset - separately for both clean and multi condition. Table 3 shows the ASR performance for clean and multi-condition training conditions using MF, PN and the proposed modulation filtering (R2-S1+R2-S2) applied on MF.

It can be observed that the proposed features perform better than MF and PN under almost all test conditions with clean and reverb training data. For the clean training, there is an average relative improvement of 29 % over MF features on Sim test data and about 8 % with Real test data. The results with the proposed front-end are better than the best published results in REVERB Challenge [18]. For the multi condition reverb training (simulated), there is an average relative improvement of 4 % over MF features on the Sim test data and the performance is similar to MF with Real test data.

3.2. Semi-supervised training

For semi-supervised ASR training, the Aurora-4 clean and multi-condition training set up is used with 70, 50 and 30 % of the labeled training data. The modulation filters are learnt using full unsupervised clean and multi-condition training data, respectively, available in the training set with mel spectrogram input. The performance comparison of ASR with semi-supervised training is shown in Table 4 for MF and the proposed feature scheme for the average of all test data conditions (14 conditions). These results indicate that the proposed features are more resilient to reduced amounts of labeled training data as compared to the baseline system (especially for clean training condition). The proposed features perform significantly better than MF features for the average of all test conditions (average relative improvement of 26 % for clean training and average relative improvement of 8 % for multi-condition training with use of 30 % labeled training data).

Table 2: *Word error rate (%) in Aurora-4 database for multi condition training with various feature extraction schemes.*

Cond	MF	PF	ET	RA	Prop
A. Clean with same Mic					
Clean	4.2	4.1	4.5	4.6	**4**
B: Noisy with same Mic					
Airport	7.5	7.9	8.0	8.1	7
Babble	7.7	7.9	7.9	8.7	7.3
Car	4.7	4.9	5.6	5.0	4.4
Rest.	9.8	10.2	11.0	11.0	8.7
Street	8.6	8.8	10.0	9.0	8
Train	8.7	8.3	9.3	9.1	8.4
Avg.	7.8	8.0	8.6	8.5	**7.3**
C: Clean with diff. Mic					
Clean	8.4	7.8	8.0	9.7	**7.6**
D: Noisy with diff. Mic					
Airport	19.7	20.9	18.5	20.1	17
Babble	20.3	20.9	19.3	20.0	19.2
Car	11.8	13.1	14.1	12.5	10.1
Rest.	21.7	23.7	21.8	23.1	18.6
Street	19.1	20.0	19.4	18.9	17.3
Train	18.3	19.6	19.6	19.9	18.1
Avg.	18.5	19.7	18.8	19.1	**16.7**
Avg. of all conditions					
Avg.	12.1	12.7	12.6	12.8	**11.1**

Table 3: *Word error rate (%) in REVERB Challenge database for clean and multi-condition training.*

Cond.	MF	PF	Prop	MF	PF	Prop
	Clean training			Multi training		
Sim_dt	37.2	36.3	**28.2**	11.9	**11.3**	11.7
Sim_et	35.8	35.2	**23.6**	12.2	**11.5**	**11.5**
Real_dt	70	73.3	**63.6**	25.9	**25.7**	26.5
Real_et	73.1	77	**68.9**	30.9	30.7	**30.6**

Table 4: *Word error rate (%) in Aurora-4 database for clean and multi condition training using lesser amount of labeled training data (70 %, 50 %, 30 %).*

Training data	100 %		70 %		50 %		30 %	
	MF	Prop	MF	Prop	MF	Prop	MF	Prop
Clean	24.6	**19.9**	26.3	**21.1**	29.3	**22.5**	33.8	**24.9**
Multi cond.	12.1	**11.1**	15.8	**14.4**	17.6	**16.3**	21	**19.3**

4. Summary

The various ASR results presented in the previous section indicate that learning data-driven 2-D modulation filters provides useful information for ASR tasks in noisy and reverberant environments. This may be attributed to identification of the relevant spectro-temporal modulations learnt from the data. The major contributions of this work are:

- Proposing an unsupervised data-driven approach to learn spectro-temporal modulation filters.
- Using the rank-1 constraint in gradient ascent method to obtain separable 2-D filters in the CRBM framework.
- Obtaining multiple irredundant filters using residual spectrograms in rate and scale domain.
- Unsupervised filter selection using average number of active hidden units in the CRBM.
- Robustness in noisy and reverberant conditions using the proposed modulation filtering scheme.
- Resilience to semi-supervised training of ASR (with limited labelled data) using proposed features.

5. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

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

[4] N. Mesgarani and S. Shamma, "Speech processing with a cortical representation of audio," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5872–5875.

[5] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS comput biol*, vol. 5, no. 3, p. e1000302, 2009.

[6] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[7] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition." in *Proc. of Eurospeech*, 2003, pp. 2573–2576.

[8] X. Domont, M. Heckmann, F. Joublin, and C. Goerick, "Hierarchical spectro-temporal features for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2008, pp. 4417–4420.

[9] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[10] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in asr of noisy speech," in *International Conference on Acoustics, Speech, and Signal Processing, Proceedings.*, vol. 1. IEEE, 1999, pp. 289–292.

[11] B. Chen, Q. Zhu, and N. Morgan, "Long-term temporal features for conversational speech recognition," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 232–242.

[12] J.-W. Hung and L.-S. Lee, "Optimization of temporal filters for constructing robust features in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 808–832, 2006.

[13] M. R. Schädler and B. Kollmeier, "Separable spectro-temporal gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2047–2059, 2015.

[14] 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)*, 2013, pp. 297–302.

[15] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *Proc. Interspeech*, pp. 1766–1770, 2013.

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

[17] M. Norouzi, M. Ranjbar, and G. Mori, "Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning," in *Conference on Computer Vision and Pattern Recognition, 2009. (CVPR) 2009.* IEEE, 2009, pp. 2735–2742.

[18] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016.

[19] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted Boltzmann machines for collaborative filtering," in *Proceedings of the 24th International Conference on Machine Learning*. ACM, 2007, pp. 791–798.

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

[21] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 609–616.

[22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[23] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[24] M. Norouzi, "Convolutional restricted boltzmann machines for feature learning," Ph.D. dissertation, School of Computing Science-Simon Fraser University, 2009.

[25] S. K. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 416–426, 2013.

[26] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[27] 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 *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[28] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2012, pp. 4101–4104.

[29] E. ETSI, "202 050 v1. 1.1 stq; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES*, vol. 202, no. 050, p. v1, 2002.