Unsupervised modulation filter learning for noise-robust speech recognition

Purvi Agrawal, and Sriram Ganapathy

Citation: The Journal of the Acoustical Society of America **142**, 1686 (2017); doi: 10.1121/1.5001926
View online: http://dx.doi.org/10.1121/1.5001926
View Table of Contents: http://asa.scitation.org/toc/jas/142/3
Published by the Acoustical Society of America

Unsupervised modulation filter learning for noise-robust speech recognition

Purvi Agrawal[a] and Sriram Ganapathy
Indian Institute of Science, Bangalore, India

(Received 16 February 2017; revised 2 June 2017; accepted 24 August 2017; published online 27 September 2017)

The modulation filtering approach to robust automatic speech recognition (ASR) is based on enhancing perceptually relevant regions of the modulation spectrum while suppressing the regions susceptible to noise. In this paper, a data-driven unsupervised modulation filter learning scheme is proposed using convolutional restricted Boltzmann machine. The initial filter is learned using the speech spectrogram while subsequent filters are learned using residual spectrograms. The modulation filtered spectrograms are used for ASR experiments on noisy and reverberant speech where these features provide significant improvements over other robust features. Furthermore, the application of the proposed method for semi-supervised learning is investigated. © *2017 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.5001926]

[DDO] Pages: 1686–1692

I. INTRODUCTION

Even with several advancements in the practical application of speech recognition technology, the performance of the state-of-the-art automatic speech recognition (ASR) systems remain fragile in high levels of noise and reverberation. One of the popular methods for robust feature extraction is the modulation filtering method which involves filtering the input speech spectrogram along the temporal and spectral axis with filters that retain only relevant speech information. It is inspired by human perceptual studies relating to the importance of temporal modulations (rate frequencies measured in Hz) and spectral modulations (scale frequencies measured in cycles/octave) (Elliott and Theunissen, 2009). The evidence of spectro-temporal modulations in the perception of complex sounds was shown with experiments in which systematic degradations of the speech signal were correlated with the gradual loss of intelligibility (Shannon *et al.*, 1995; Drullman *et al.*, 1994).

Several studies have attempted incorporating the knowledge of modulation filters for ASR. One of the earliest use of temporal modulations was the RASTA filtering approach (Hermansky, 1994). There have been several attempts on the use of spectro-temporal modulation filters for feature extraction [for example, Gabor filtering (Kleinschmidt, 2003; Ezzat *et al.*, 2007; Domont *et al.*, 2008)]. For learning the temporal modulation filters in a data-driven manner, the linear discriminant analysis (LDA) has been explored (Van Vuuren and Hermansky, 1997; Hung and Lee, 2006). A data-driven approach for parameter selection of Gabor filter set has been recently studied (Kovacs *et al.*, 2015; Schadler *et al.*, 2012). A recent approach to separable spectro-temporal Gabor filter bank features shows that spectral and temporal processing can be performed independently (Schadler and Kollmeier, 2013). A data-driven approach has also been attempted to learn the acoustic filters from raw audio signal (Sainath *et al.*, 2013; Palaz *et al.*, 2013; Sailor and Patil, 2016).

In this work, we propose a new approach to learn spectral and temporal modulation filters purely from an unsupervised data-driven perspective. In particular, a filter learning method is developed using the speech spectrogram in conjunction with a convolutional restricted Boltzmann machine (CRBM) (Norouzi *et al.*, 2009). The projection of the input spectrogram on the learned filter is removed to obtain a residual spectrogram which is iteratively used in the CRBM framework for learning subsequent filters (Mallat and Zhang, 1993). The filter learning process is performed separately in the spectral and temporal domains. After a set of filters are learned, a filter selection method is used based on the average hidden activations in the CRBM. The selected modulation filters are applied on the input spectrogram to derive features for speech recognition.

The ASR experiments are performed on the Aurora-4 database using a deep neural network (DNN) acoustic model, both with clean and multi condition training setup. We also perform ASR experiments on reverberant speech provided in the REVERB challenge (Kinoshita *et al.*, 2016). The results from these experiments indicate that the features derived from data-driven modulation filters provide significant improvements over other noise robust front-ends. Finally we highlight the application of the proposed features for semi-supervised training where limited amount of labeled training data are available.

The rest of the paper is organized as follows. In Sec. II, we describe the filter learning method using convolutional RBM architecture, followed by the application of modulation filtering for feature extraction. Section III describes the ASR experiments using the proposed front-end. We conclude with a brief discussion in Sec. IV.

II. UNSUPERVISED MODULATION FILTER LEARNING

The proposed feature extraction scheme consists of three stages—learning a modulation filter using CRBM

[a]Electronic mail: purvi_agrawal@ee.iisc.ernet.in

architecture, learning multiple irredundant filters and filter selection, and feature extraction for ASR.

A. Convolutional restricted Boltzmann machine

The restricted Boltzmann machine (RBM) (Salakhutdinov *et al.*, 2007) is a two-layer, undirected graphical model with a set of binary hidden units **h**, a set of (binary or real-valued) visible units **v**, and symmetric connections between these two layers represented by a weight matrix **W**. In the forward pass, the RBM uses inputs to make predictions about node activations, or the probability of output given the input **v**: $p(\mathbf{h}|\mathbf{v};\mathbf{W})$. In the backward pass, the model attempts to estimate the probability of inputs **v** given activations **h**, expressed as $p(\mathbf{v}|\mathbf{h};\mathbf{W})$. We use the contrastive divergence (CD) learning algorithm for RBM training using gradient ascent based optimization procedure (Hinton, 2002). The one-step contrastive divergence approximation (one-step Gibbs sampler) is given as

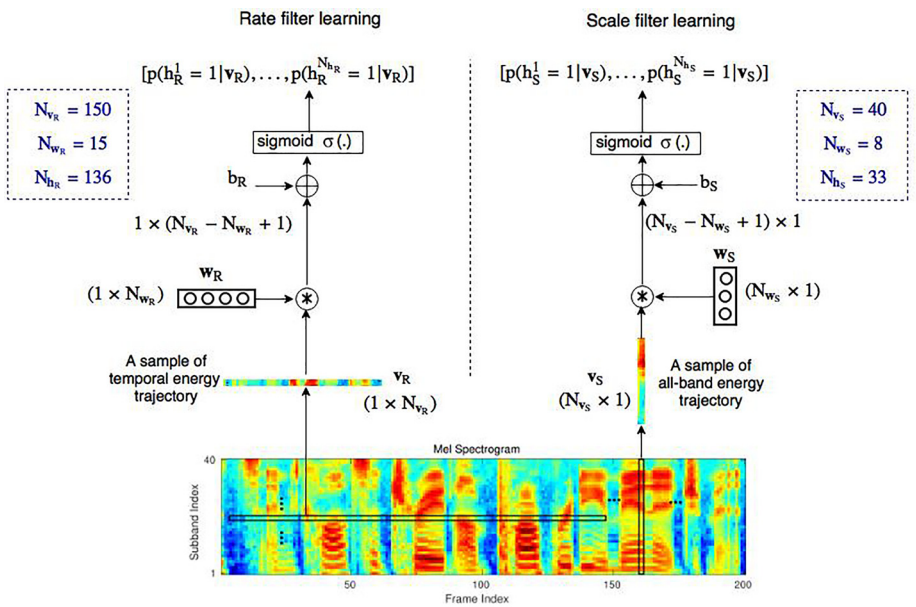
$$\triangle_{w_{ji}} J(\mathbf{W}, a, b, ; \mathbf{v}) = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}, \tag{1}$$

where J is the log likelihood, $\langle\cdot\rangle$ denotes the expectation under the distribution specified by the subscript, v_i and h_j are the ith and jth, visible and hidden node values, respectively. A convolutional operation can be added to RBM learning by weight sharing, reconstructing and identifying the features of the signal locally (Norouzi *et al.*, 2009; Lee *et al.*, 2009).

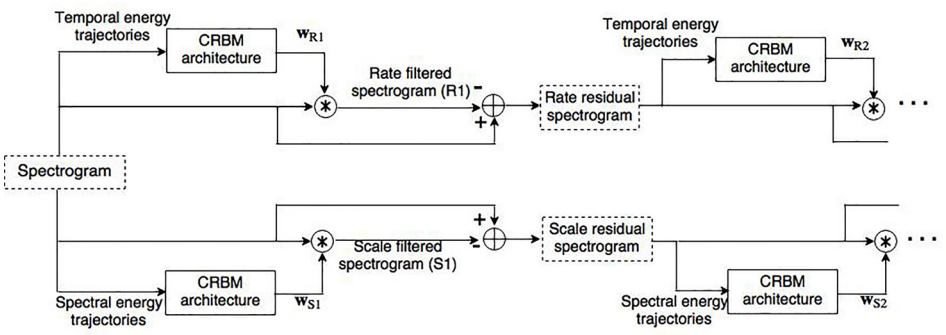
The block schematic of the proposed modulation filter learning scheme from speech spectrogram is shown in Fig. 1(b). Here, the speech spectrogram is processed with a CRBM architecture [shown in Fig. 1(a) for deriving one rate and scale filter]. For a 1-D rate filter ($\mathbf{w}_R$) learning, the input ($\mathbf{v}_R$) consists of temporal energy trajectories of individual subbands for 1.5 s of speech from training dataset (each of dimension $1 \times N_{V_R}$, with $N_{V_R} = 150$). For 1-D scale filter ($\mathbf{w}_S$) learning, the input ($\mathbf{v}_S$) consists of all-band energy trajectories of individual speech frames (each of dimension $N_{V_S} \times 1$, $N_{V_S} = 40$ for mel spectrogram). The visible layer and the hidden layer have bias a_R and b_R (a_S and b_S) for rate (scale) filter learning, respectively. The conditional distributions used to perform block Gibbs sampling for rate filtering (similar relations hold for scale filtering also) are

$$P(h_R^j = 1|\mathbf{v}_R) = \sigma((\mathbf{w}_R \star \mathbf{v}_R)_j + b_R), \tag{2}$$

$$P(v_R^i|\mathbf{h}_R) = \mathcal{N}((\mathbf{w}_R \star \mathbf{h}_R)_i + a_R), \tag{3}$$



(a) CRBM architecture to learn a rate and scale filter from mel spectrogram



(b) Learning multiple filters using residual approach

FIG. 1. (Color online) The top panel (a) shows the CRBM architecture used for learning a single rate ($\mathbf{w}_R$) and a scale ($\mathbf{w}_S$) filter separately from the spectrogram (forward pass of CRBM). The bottom panel (b) shows the proposed schematic for learning multiple rate and scale filters.

where σ is the sigmoid function, i and j are the index of the visible and hidden layer unit, $\mathcal{N}$ is the Gaussian distribution function. We begin the CRBM training with random initialization of weight vector $\mathbf{w}_R$ ($\mathbf{w}_S$) and perform several iterative steps to learn the rate (scale) filter.

B. Learning multiple irredundant filters and filter selection

Once a rate (scale) filter is derived, the filtered component of the input spectrogram is removed from the input spectrogram and the residual spectrogram is fed back to CRBM for learning subsequent rate (scale) filters, as shown in Fig. 1(b). This method, similar to the matching pursuit (MP) algorithm (Mallat and Zhang, 1993), allows us to learn irredundant set of filters. In our work, three filters are successively learned from CRBM. The learned rate (scale) filters are denoted by $\mathbf{w}_{R1}$, $\mathbf{w}_{R2}$, $\mathbf{w}_{R3}$ ($\mathbf{w}_{S1}$, $\mathbf{w}_{S2}$, $\mathbf{w}_{S3}$), respectively. The corresponding filtered spectrograms are denoted by R1, R2, R3 (S1, S2, S3), respectively.

The magnitude response of the data-driven filters obtained from clean Aurora-4 database are shown in Fig. 2. In addition, a comparison has been made between the CRBM based rate and scale filters with filters learned from principal component analysis (PCA) (obtained from complex 1-D Fourier representation of the corresponding temporal and spectral energy trajectories) (Jolliffe, 2002), convolutive nonnegative matrix factorization (CNMF) with the spectrogram inputs (Wang and Zhang, 2013) and LDA based filters learned on time trajectory of each subband (Van Vuuren and Hermansky, 1997). As seen here, the proposed CRBM based approach for modulation filter learning provides more smoother filters with broad modulation selectivity similar to those observed in perceptual studies (Chi *et al.*, 2005; Elliott and Theunissen, 2009). While a convolutional neural network (CNN) (when used for acoustic modelling) also performs spectro-temporal filtering in ASR, the CNN filters are learnt in a supervised manner (using labelled data). Also, the CNNs typically employed in speech perform local spectro-temporal filtering of 200–300 ms of temporal context (Huang *et al.*, 2015). In our case, the filters span entire spectral range for a long temporal context of 1.5 s.

TABLE I. Average hidden activation probability obtained from filtering validation dataset with each of the obtained learned filter individually (averaged over all utterances) in Aurora-4 database on the Mel (MFB) and auditory (ASp) spectrogram and in REVERB database on the Mel (MFB) spectrogram.

Spectrogram	Rate filter			Scale filter		
	$\mathbf{w}_{R1}$	$\mathbf{w}_{R2}$	$\mathbf{w}_{R3}$	$\mathbf{w}_{S1}$	$\mathbf{w}_{S2}$	$\mathbf{w}_{S3}$
Aurora-4: Clean training						
Mel	0.32	**0.38**	0.06	**0.34**	0.23	0.27
Auditory	0.26	**0.30**	0.05	**0.31**	0.26	0.09
Aurora-4: Multi condition training						
Mel	0.28	**0.33**	0.09	**0.35**	0.18	0.26
Auditory	0.23	**0.29**	0.06	**0.30**	0.27	0.08
REVERB training						
Mel	0.30	**0.31**	0.09	**0.39**	0.23	0.22

In order to choose the rate and scale filters for ASR, we compute the average hidden activation probability for each filter by a forward pass operation of the input spectrograms through the CRBM. Table I shows the average hidden activation probability value obtained for clean Aurora-4, multi condition Aurora-4 as well as REVERB challenge database. Based on the highest average activation values from the validation data, we select second rate ($\mathbf{w}_{R2}$) and first scale filter ($\mathbf{w}_{S1}$) to derive R2 + S1 features, which is consistent for each case. In ASR using Aurora-4 clean training, we observed that adding R2 + S2 features provided additional improvements along with R2 + S1 features. Hence, we use (R2 + S1, R2 + S2) features for ASR.

C. Feature extraction overview

The log-mel spectrogram (MFB) of speech signal is obtained using window length of 25 ms with a shift of 10 ms using 40 mel subband filters between 250–6500 Hz. The auditory spectrogram (ASp) is obtained using an auditory-inspired model of cochlear processing (Chi *et al.*, 2005). It involves stages of affine wavelet transform of the acoustic signal, first derivative with respect to the frequency axis followed by half-wave rectification in time, a short-term integration, and final stage of cubic root compression. The

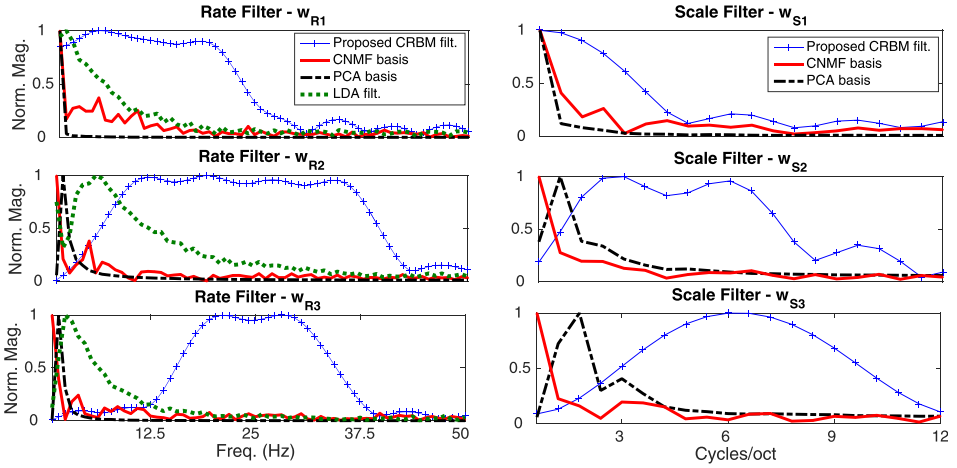


FIG. 2. (Color online) Comparison of magnitude response of the proposed data-driven CRBM filters with the filters obtained from LDA, complex PCA, and CNMF. All the filters are derived for mel spectrogram input extracted from Aurora-4 clean training data.

auditory spectrogram is also sampled at 10 ms window shift, with each frame having 113 spectral bands between 250–6500 Hz. From these two spectrograms, two streams of joint rate-scale filtered spectrograms are derived (R2 + S1 and R2 + S2). These spectrogram streams are concatenated and fed to a deep neural network (DNN) based ASR system. The input features are mean-variance normalized at utterance level before DNN training. Figure 3 shows the rate-scale filtered spectrogram (R2 + S1) on (a) clean test speech file and (b) babble noise test speech file. As seen here, application of modulation filtering can provide more invariant representations compared to conventional mel spectrograms. This may be attributed to the modulation filter characteristics learned from the clean data distribution. For the ASR experiments, we derive data-driven filters separately from mel spectrogram (MFB) input as well as from the auditory spectrogram (ASp) input.[1]

III. EXPERIMENTS

A. Speech recognition system

The WSJ Aurora-4 corpus is used for ASR experiments which consists of continuous read speech recordings, recorded under clean and noisy conditions (street, train, car, babble, restaurant, and airport) at 10–20 dB SNR. The recordings were carried out with two microphones and the training data have two sets of 7138 clean and multi condition recordings, respectively (84 speakers). The validation data have two sets of 1206 clean and multi condition recordings, respectively (14 speakers) and test data have 330 recordings (8 speakers) for 14 clean and noise conditions. The test data are classified into four groups: A—clean data, B—noisy data, C—clean data with channel distortion, and D—noisy data with channel distortion.

The speech recognition Kaldi toolkit (Povey *et al.*, 2011) is used for building the ASR. A deep belief network-deep neural network (DBN-DNN) with four hidden layers having ten 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 (MFB) features, power normalized filter bank energy (PFB) features (Kim and Stern, 2012), advanced ETSI front-end (ETS) (ETSI, 2002), RASTA features (RAS) (Hermansky and Morgan, 1994), LDA based features (Van Vuuren and Hermansky, 1997), spectro-temporal Gabor filters with filter selection based features (GAB) (Kovacs *et al.*, 2015), MHEC features (MHE) (Sadjadi and Hansen, 2015), and auditory spectrogram features (ASp) (Chi *et al.*, 2005). The results for the proposed data-driven modulation filtering obtained from MFB and ASp are also shown here.

From the ASR performance in clean training condition reported in Table II, it can be observed that PFB and ETS features provide better performance compared to all other baseline features. The data-driven modulation filtering approach on MFB and ASp improves the performance of MFB and ASp features in noisy and channel distortion scenarios. The rate filtering (using R2) gives a average relative improvements of 16% over MFB and 18% over ASp and the scale filtering (concatenation of S1, S2 filtered spectrograms) also provides moderate improvements. The joint application of selected rate and scale filtering (R2 + S1, R2 + S2) provides significant robustness to noisy and multi-channel test conditions (average relative improvements of 21% over MFB features and 20% over ASp features).

In the matched multi condition training and test scenario in Table III, the GAB features perform better than all other baseline features. The data-driven modulation filtering approach using rate filter on MFB improves the performance compared to the baseline features (average relative improvements of 9% over MFB). The best performance is provided by the joint application of selected rate and scale filtering for all the clean and noisy test conditions, improving the baseline MFB results on average by about 11% and improving the ASp results by about 16%. Our results with clean and multi condition training also improve over the results obtained from a CNN based ASR system using mel spectrogram (Huang *et al.*, 2015).

B. Reverberant speech recognition

The ASR experiments on reverberated speech data are performed in a single channel scenario using WSJCAM0 corpus, released as a part of REVERB challenge (Kinoshita

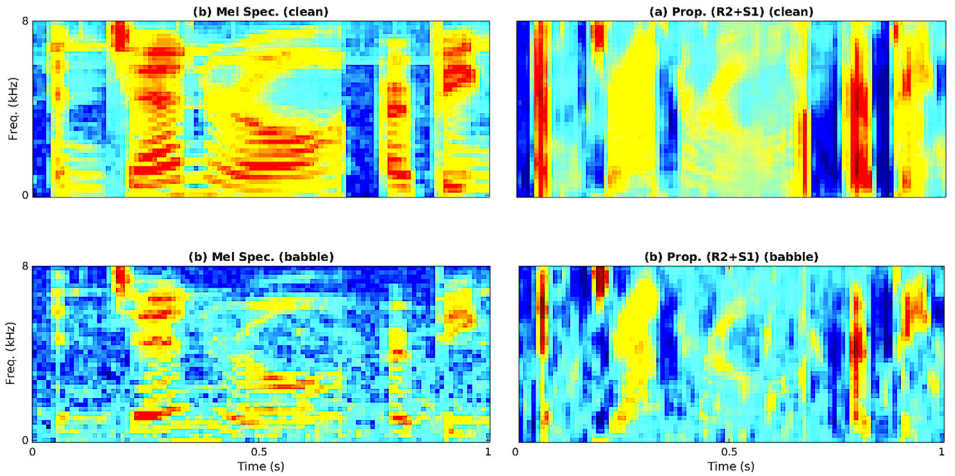


FIG. 3. (Color online) Comparison of mel spectrogram and the data-driven rate-scale filtering of mel spectrogram for (a) clean file and (b) babble noise file (different microphone) recorded from a female speaker in Aurora-4 database. The modulation filters with the highest activation probability ($\mathbf{w}_{R2} + \mathbf{w}_{S1}$) are used in the right side panels to obtain (R2 + S1).

TABLE II. Word error rate (%) in Aurora-4 database for clean training condition with various feature extraction schemes and the proposed (R2 + S1, R2 + S2) modulation filtering approach applied on the auditory (ASp) and the Mel (MFB) spectrogram.

Condition	Baseline Feature Type								Rate (R2)		Scale (S1, S2)		(R2 + S1,R2 + S2)	
	MFB	ASp	PFB	ETS	RAS	LDA	MHE	GAB	MFB	ASp	MFB	ASp	MFB	ASp
(A) Clean with same microphone														
Clean	3.4	3.2	3.3	3.2	3.5	3.7	3.5	**2.6**	2.9	3.2	3.2	3.2	3.3	3.3
(B) Noisy with same microphone														
Airport	21.9	21.2	18.3	15.0	19.3	23.2	19.5	19.4	14.9	14.6	21.6	22.6	13.1	14.2
Babble	19.6	18.5	16.0	15.5	19.9	21.0	17.7	19.2	15.4	14.2	19.5	18.3	13.9	13.4
Car	8.0	8.2	6.2	9.8	7.9	8.7	7.9	6.9	5.8	6.0	9.3	8.2	5.7	5.7
Restaurant	24.9	26.0	22.9	20.5	23.0	27.0	23.2	23.5	18.2	17.2	23.7	24.7	16.9	16.7
Street	19.5	19.0	17.8	19.5	18.7	20.8	18.1	19.2	15.4	14.2	19.3	19.1	14.9	13.6
Train	19.8	19.0	16.3	17.4	19.4	20.1	17.9	20.1	16.9	13.9	20.4	18.7	16.9	14.6
Average	18.9	18.7	16.2	16.3	18.0	20.1	17.4	18.0	14.5	13.4	19.0	18.6	13.6	**13.0**
(C) Clean with different microphone														
Clean	15.3	14.4	**11.7**	14.5	16.0	15.9	14.6	**11.7**	12.7	12.8	14.8	13.8	13.1	13.1
(D) Noisy with different microphone														
Airport	40.1	37.9	36.4	31.4	39.2	40.4	38.7	36.1	32.9	32.8	39.7	38.3	30.0	31.1
Babble	37.3	36.3	34.2	32.1	38.5	36.8	36.8	37.2	33.8	33.2	37.7	36.1	32.1	30.8
Car	24.9	23.2	21.5	24.9	24.8	25.9	25.9	22.2	20.8	20.8	26.0	21.5	19.0	20.0
Restaurant	39.6	39.5	39.0	35.4	39.1	41.0	39.3	38.2	34.8	33.2	38.6	38.3	31.5	32.8
Street	35.7	35.3	34.1	35.0	35.8	37.0	35.8	37.9	32.9	30.4	36.5	35.1	30.0	28.5
Train	35.6	32.5	31.8	33.2	36.4	36.7	35.9	39.0	34.1	31.4	37.2	34.2	31.5	30.0
Average	35.2	32.4	32.8	32.0	35.6	36.3	35.4	35.0	31.6	30.3	36.0	33.9	29.0	**28.9**
Average of all conditions														
Average	24.7	24.0	22.1	21.9	24.4	25.6	23.9	23.8	20.8	19.8	24.8	23.7	19.4	**19.1**

TABLE III. Word error rate (%) in Aurora-4 database for multi condition training condition with various feature extraction schemes and the proposed (R2 + S1, R2 + S2) modulation filtering approach applied on the auditory (ASp) and the Mel (MFB) spectrogram.

Condition	Baseline feature type								Rate (R2)		Scale (S1, S2)		(R2 + S1, R2 +S2)	
	MFB	ASp	PFB	ETS	RAS	LDA	MHE	GAB	MFB	ASp	MFB	ASp	MFB	ASp
(A) Clean with same microphone														
Clean	4.2	4.6	4.1	4.5	4.6	4.7	4.0	**3.3**	3.5	4.1	3.9	4.4	3.7	4.0
(B) Noisy with same microphone														
Airport	7.5	9.9	7.9	8.0	8.1	10.1	8.2	7.2	6.8	7.7	7.4	9.1	6.8	7.6
Babble	7.7	9.4	7.9	7.9	8.7	9.9	8.6	7.2	6.4	8.4	7.8	9.7	6.8	7.4
Car	4.7	5.7	4.9	5.6	5.0	5.8	4.9	3.8	4.1	4.9	4.5	5.3	4.2	4.7
Restaurant	9.8	12.4	10.2	11.0	11.0	12.6	11.1	9.5	8.8	9.7	9.9	11.1	9.5	9.7
Street	8.6	10.6	8.8	10.0	9.0	10.6	8.8	8.1	7.7	8.3	8.8	9.4	8.1	8.3
Train	8.7	10.3	8.3	9.3	9.1	10.6	8.4	8.6	8.1	9.0	8.7	10.2	8.6	8.9
Average	7.8	9.7	8.0	8.6	8.5	9.9	8.3	7.4	**7.0**	8.0	7.9	9.1	7.3	7.8
(C) Clean with different microphone														
Clean	8.4	10.1	7.8	8.0	9.7	10.0	8.1	**6.1**	7.7	8.6	7.2	9.0	7.1	7.8
(D) Noisy with different microphone														
Airport	19.7	21.8	20.9	18.5	20.1	22.3	20.8	17.5	17.2	18.8	18.1	21.0	16.2	18.0
Babble	20.3	22.3	20.9	19.3	20.0	22.5	21.3	18.4	18.5	20.5	19.2	21.7	17.8	19.6
Car	11.8	12.9	13.1	14.1	12.5	14.5	12.8	8.6	10.6	11.9	10.2	12.4	10.1	10.8
Restaurant	21.7	24.2	23.7	21.8	23.1	25.2	23.1	20.8	19.5	21.2	20.8	23.2	18.7	20.5
Street	19.1	21.9	20.0	19.4	18.9	21.2	20.5	18.6	17.6	18.8	17.9	20.2	17.0	17.8
Train	18.3	20.2	19.6	19.6	19.9	21.6	18.9	19.8	17.9	19.0	18.7	20.2	17.4	18.2
Average	18.5	20.6	19.7	18.8	19.1	21.2	19.6	17.3	16.9	18.4	17.5	19.8	**16.2**	17.5
Average of all conditions														
Average	12.1	14.0	12.7	12.6	12.8	14.4	12.8	11.2	11.0	12.2	11.6	13.3	**10.8**	11.7

TABLE IV. Word error rate (%) in REVERB Challenge database for clean and multi-condition training with test data from simulated and real reverb environments.

Condition	MFB	PFB	RAS	MHE	R2 + S1, R2 + S2
Clean training					
Sim_dt	37.2	36.3	32.5	34.5	**28.2**
Sim_et	35.8	35.2	30.4	33.4	**27.2**
Real_dt	70	73.3	67.4	69.0	**63.3**
Real_et	73.1	77	71.0	71.1	**68.9**
Multi condition reverb training					
Sim_dt	11.9	11.3	13.5	11.3	**11.2**
Sim_et	12.2	11.5	12.9	11.6	**11.1**
Real_dt	25.9	25.7	30.7	**25.2**	25.3
Real_et	30.9	30.7	33.6	30.3	**29.4**

et al., 2016). 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 head-mounted microphones as well as a desk microphone positioned about a 0.5 m from the speaker's head. The database consists of three subsets: training data set (Train) for both clean and multi condition reverb training using simulated reverb data, a simulated test dataset (Sim), and a naturally reverberant recording of the test dataset (Real). The rate and scale filters are learnt from mel spectrogram of Train dataset—separately for both clean and multi conditions. Table IV shows the ASR performance for clean and multi-condition training conditions using MFB, PFB, RAS, MHE and the proposed modulation filtering (R2 + S1, R2 + S2) applied on MFB.

It can be observed that the proposed features perform better than all the other baseline features under all test conditions with clean and reverb training data. For the clean training, there is an average relative improvement of 24% over MFB features on Sim test data and about 8% with Real test data. For the multi condition reverb training (simulated), there is average relative improvement of 7% over MFB features on the Sim test data and about 4% with Real test data. Furthermore, the results with the proposed front-end are better than the previously published results in theREVERB Challenge (Kinoshita *et al.,* 2016).

C. Semi-supervised learning

For semi-supervised ASR training, we use the Aurora-4 clean condition training set up with 70%, 50%, and 30% of the labeled training data. In this case, the modulation filters were learned using unsupervised training data available in the clean training set with mel spectrogram input. Figure 4 shows the performance comparison of ASR with semi-supervised training using MFB and the proposed feature scheme for clean test data condition (A), as well as the average of all for test data conditions (average of 14 conditions).

It can be observed that the proposed features are more resilient to reduced amounts of labeled training data as compared to the baseline system, even for the matched clean test condition. The proposed features also perform significantly better than MFB on the average of all test conditions (average relative improvement of 29% with use of 30% training data over MFB features).

IV. DISCUSSION AND SUMMARY

The results presented in Sec. III indicate that the paradigm of learning data-driven modulation filters provides significant robustness in noisy speech recognition in both clean and multi condition training cases. Furthermore, the improvements are consistent with semi-supervised training of ASR and for reverberant speech recognition. The improved performance may be attributed to the enhancement of key modulations in the temporal and spectral domain learned from the training data distribution using the CRBM architecture.

The following are the major contributions form this work.

- Proposing an unsupervised data-driven approach to learn spectral and temporal modulation filters with a random initialization.
- Obtaining multiple irredundant data-driven filters with the CRBM and residual spectrograms. A filter selection criterion using average hidden activation probability in CRBM.
- Robustness in noisy and reverberant conditions using the proposed modulation filtering scheme.
- Improved resilience to reduced amounts of labeled training data for the proposed features.

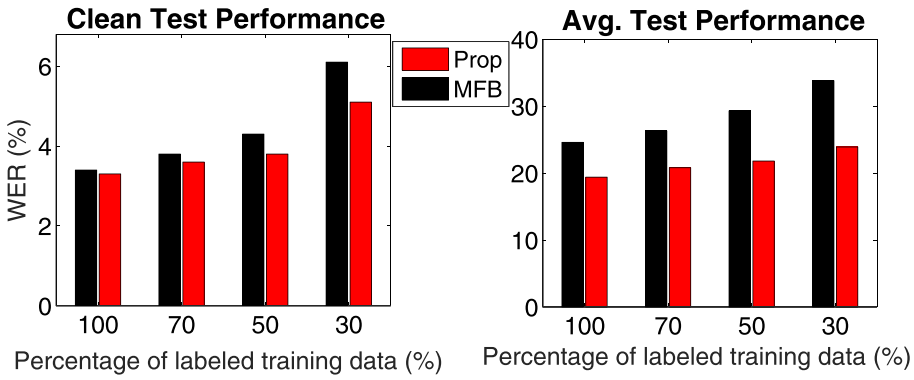


FIG. 4. (Color online) Performance of ASR (WER) versus amount of clean labeled training data. Comparison between MFB and proposed modulation filtering (R2 + S1, R2 + S2) applied on MFB using cleaning training condition on Aurora-4. Results split for clean test condition (A) and average of all 14 test conditions.

ACKNOWLEDGMENTS

This work was supported by academic grant from NVIDIA Corp. We would like to acknowledge Gyorgy Kovacs for providing the Gabor based features on Aurora-4 database.

[1]In order to encourage reproducible research, we provide the implementation of the proposed approach at https://github.com/PurviAgrawal/Unsupervised_modFilt_CRBM-master.

Chi, T., Ru, P., and Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**(2), 887–906.

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

Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**(2), 1053–1064.

Elliott, T. M., and Theunissen, F. E. (**2009**). "The modulation transfer function for speech intelligibility," PLoS Comput. Biol. **5**(3), e100302.

ETSI, E. (**2002**). "202 050 v1. 1.1 STQ; Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms," ETSI ES **202**(050), v1, available at http://www.etsi.org/deliver/etsi_es/202000_202099/202050/01.01.05_60/es_202050v010105p.pdf.

Ezzat, T., Bouvrie, J. V., and Poggio, T. A. (**2007**). "Spectro-temporal analysis of speech using 2-D Gabor filters," Proc. Interspeech 506–509.

Hermansky, H., and Morgan, N. (**1994**). "RASTA processing of speech," IEEE Trans. Speech Audio Process. **2**(4), 578–589.

Hinton, G. E. (**2002**). "Training products of experts by minimizing contrastive divergence," Neural Comput. **14**(8), 1771–1800.

Huang, J. T., Li, J., and Gong, Y. (**2015**). "An analysis of convolutional neural networks for speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4989–4993.

Hung, J. W., and Lee, L. S. (**2006**). "Optimization of temporal filters for constructing robust features in speech recognition," IEEE Trans. Audio Speech Lang. Process. **14**(3), 808–832.

Jolliffe, I. (**2002**). *Principal Component Analysis* (Wiley, New York).

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

Kinoshita, K., Delcroix, M., Gannot, S., Habets, E. A., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., and Sehr, A. (**2016**). "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," EURASIP J. Adv. Sign. Process. **2016**(1), 1–19.

Kleinschmidt, M. (**2003**). "Localized spectro-temporal features for automatic speech recognition," in *Proceedings of Eurospeech*, 2003, pp. 2573–2576.

Kovacs, G., Toth, L., and Van Compernolle, D. (**2015**). "Selection and enhancement of Gabor filters for automatic speech recognition," Int. J. Speech Technol. **18**(1), 1–16.

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

Mallat, S. G., and Zhang, Z. (**1993**). "Matching pursuits with time-frequency dictionaries," IEEE Trans. Sign. Process. **41**(12), 3397–3415.

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

Palaz, D., Collobert, R., and Doss, M. M. (**2013**). "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proceedings of Interspeech*, pp. 1766–1770.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., and Silovsky, J. (**2011**). "The Kaldi speech recognition toolkit," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Sadjadi, S. O., and Hansen, J. H. (**2015**). "Mean Hilbert envelope coefficients (MHEC) for robust speaker with CNN as the ASR training system and language identification," Speech Commun. **72**, 138–148.

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

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

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

Schadler, M. R., and Kollmeier, B. (**2013**). "Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition," J. Acoust. Soc. Am. **137**(4), 2047–2059.

Schadler, M. R., Meyer, B. T., and Kollmeier, B. (**2012**). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," J. Acoust. Soc. Am. **131**(5), 4134–4151.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.

Van Vuuren, S., and Hermansky, H. (**1997**). "Data-driven design of RASTA-like filters," in *Proceedings of Eurospeech*, pp. 1607–1610.

Wang, Y. X., and Zhang, Y. J. (**2013**). "Nonnegative matrix factorization: A comprehensive review," IEEE Trans. Knowledge Data Eng. **25**(6), 1336–1353.