DEEP VARIATIONAL FILTER LEARNING MODELS FOR SPEECH RECOGNITION

Purvi Agrawal and Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) lab,
Electrical Engineering, Indian Institute of Science, Bangalore, India.

ABSTRACT

We present a novel approach to derive robust speech representations for automatic speech recognition (ASR) systems. The proposed method uses an unsupervised data-driven modulation filter learning approach that preserves the key modulations of speech signal in spectro-temporal domain. This is achieved by a deep generative modeling framework to learn modulation filters using convolutional variational autoencoder (CVAE). A skip connection based CVAE enables the learning of multiple irredundant modulation filters in the time and frequency modulation domain using temporal and spectral trajectories of input spectrograms. The learnt filters are used to process the spectrogram features for ASR training. The ASR experiments are performed on Aurora-4 (additive noise with channel artifact) and CHiME-3 (additive noise with reverberation) databases. The results show significant improvements for the proposed CVAE model over the baseline features as well as other robust front-ends (average relative improvements of 9% in word error rate over baseline features on Aurora-4 database and 23% on CHiME-3 database). In addition, the performance of the proposed features is highly beneficial for semi-supervised training of ASR when reduced amounts of labeled training data are available (average relative improvements of 29% over baseline features on Aurora-4 database with 30% of the labeled training data).

***Index Terms*—** Unsupervised filter learning, Convolutional variational autoencoder, skip connections, modulation filtering, robust speech recognition.

1. INTRODUCTION

The performance degradation of speech applications such as voice search or conversational-bots in noisy and reverberant environment demands the need for improved robustness in automatic speech recognition (ASR) systems. While several advancements have been made in the acoustic modeling for ASR, the presence of extrinsic noise sources and reverberations continue to pose challenge to the ASR system deployment [1]. The noise robustness can be partly addressed by multi-condition training (utilizing noisy training data from multiple environments) [2]. In spite of this training, the performance difference between multi-condition train-test and the clean train-test of ASR is pronounced, which warrants the need for attaining noise robustness either at speech representation stage or the training stage. This work focuses on the robust representation learning using unsupervised generative modeling method.

Modulation filtering is an approach to robust feature extraction that is based on enhancing the key dynamics of the speech signal in the spectro-temporal domain, while suppressing speech modulations that are susceptible to noise/reverberation. This is partly motivated from the remarkable robustness seen in the human auditory system to many of the environmental artifacts [3, 4]. Several works in the past have incorporated the knowledge of spectro-temporal modulation filtering for ASR. One of the earliest use of temporal modulations (rate) was the RASTA filtering approach [5]. The spectro-temporal modulation (rate-scale) filters for feature extraction, for example, Gabor filtering [6, 7], have shown further improvements for ASR. A data-driven approach for parameter selection of Gabor filter set has been studied in [8]. The linear discriminant analysis (LDA) has also been explored to learn the temporal modulation filters in a supervised manner [9, 10]. Recently, we have also analyzed unsupervised rate-scale filtering using generative modeling approaches [11] where a two-stage filter learning approach with restricted Boltzmann architecture is used.

In this work, we propose a new approach to learn temporal and spectral modulation filters from a variational modeling perspective. In particular, convolutional variational autoencoder (CVAE) is used to learn multiple modulation filters from the input spectrogram [12]. To learn multiple irredundant modulation filters, we propose a skip connection based network architecture [13]. The 'skip' architecture in the encoder of CVAE is employed to serve two purposes: to find residual of the input with some filtered representation (which has already been learnt), and to combine the two different filtered representations to be fed to the rest of the network for joint minimization of the loss function. The kernels of the two convolutional layers of the encoder are interpreted as modulation filters, that captures modulations derived from large amount of unsupervised speech spectrogram data, and can provide important cues regarding the useful spectral and temporal modulations of speech. The learnt filters are then used to derive robust spectrogram representations for ASR. The ASR experiments are performed on Aurora-4 (additive noise with channel artifact) and CHiME-3 challenge (additive noise with reverberation) databases. The proposed filter learning approach provides significant improvements in terms of WER over various other noise robust front-ends.

The rest of the paper is organized as follows. The theory of variational modeling in autoencoder networks is described in Sec. 2. The use of convolutional variational autoencoder for filter learning from speech signal is discussed in Sec. 3. Sec. 4 describes various ASR experiments and results, followed by summary in Sec. 5.

2. VARIATIONAL AUTOENCODER (VAE)

The VAE differs from a standard AE where the VAE model assumes that the samples of latent representation can be drawn from a standard normal distribution, i.e $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [12]. The encoder estimates the parameters of the latent data distribution that approximates the posterior distribution of the latent vector given the data. The decoder then samples from the approximate distribution and attempts

This work was partly funded by grants from the Department of Science and Technology (DST), Early career award (ECR01341) and the Pratiksha Trust Young Investigator Award.

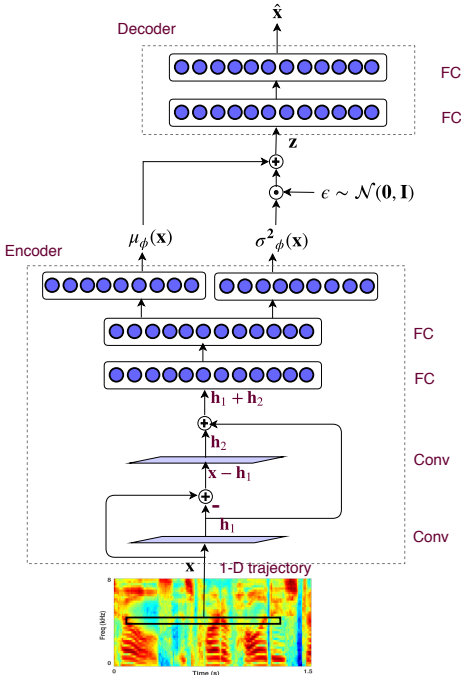

Fig. 1. Block schematic (bottom-up) of rate filter learning with CVAE using skip connections in Encoder for residual learning. Here FC denotes fully connected layer, Conv denotes convolution layer.

to reconstruct the original data back. Both the encoder and decoder parameters can be trained using a deep learning framework. If we assume an observation vector $\mathbf{x}$, a latent vector $\mathbf{z}$ and a set of parameters θ for the decoder network, the aim of the VAE network is to maximize the probability of each $\mathbf{x}$ in the training set under the generative process, according to

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \quad (1)$$

The model involves two steps: (1) $\mathbf{z}$ is generated from prior distribution $p(\mathbf{z})$; (2) a value $\mathbf{x}$ is generated from conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$. However, in the assumed generative model with a decoder neural network, the function $p_\theta(\mathbf{x})$ is not always differentiable w.r.t. θ due to the intractable integral in Eq. 1; therefore θ cannot be optimized directly. The VAE framework resolves this problem based on a variational lower bound method [12]. A new function $q_\phi(\mathbf{z}|\mathbf{x})$ (probabilistic encoder with encoder parameters ϕ) is introduced that can take value of $\mathbf{x}$ and give a distribution over $\mathbf{z}$ values. In other words, the function $q_\phi(\mathbf{z}|\mathbf{x})$ approximates the true posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$. The encoder and decoder parameters, ϕ and θ, respectively, are trained by maximizing the 'variational' lower bound $\mathcal{L}(\theta, \phi; \mathbf{x})$ of the marginal likelihood $\log p_\theta(\mathbf{x})$, given as

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = -D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] + \mathbb{E}_{\mathbf{z}|\mathbf{x}\sim q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})] \quad (2)$$

Thus, maximizing $\mathcal{L}$ inherently maximizes the data likelihood. The negative of the first term in Eq. 2 is termed as 'latent loss' which is the KL divergence of $q_\phi(\mathbf{z}|\mathbf{x})$ with unit Gaussian distribution $p(\mathbf{z})$. The second term in Eq. 2 with Gaussian assumptions on $p_\theta(\mathbf{x}|\mathbf{z})$, reduces to the negative of mean square error (MSE) loss. In the implementation of VAE, the distribution $p_\theta(\mathbf{x}|\mathbf{z})$ is assumed to be Gaussian. In addition, the distribution $q_\phi(\mathbf{z}|\mathbf{x})$ is also assumed to be Gaussian. However, the error through a layer that samples $\mathbf{z}$ from $q_\phi(\mathbf{z}|\mathbf{x})$ needs to be back-propagated, which is a non-continuous operation and has no gradient. Hence, a "reparameterization trick" solution is used to move the sampling to an input layer. Given the parameters of the encoder network, $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\sigma}^2_\phi(\mathbf{x})$ which are the mean and variance parameters of $q_\phi(\mathbf{z}|\mathbf{x})$ - we can sample from $\mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), diag(\boldsymbol{\sigma}^2{}_\phi(\mathbf{x})))$ by first sampling $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then computing $\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + diag(\boldsymbol{\sigma}_\phi(\mathbf{x}))\boldsymbol{\epsilon}$. This operation is shown schematically in Figure 1. A similar architecture is used on spectral slices of the spectrogram for scale filter learning.

Table 1. The architecture of the CVAE model used for rate and scale filter learning.

Number of layers - encoder	Conv: 2, FC: 2
Number of layers - decoder	FC: 2
Number of kernels, kernel size in Conv	1, 1×5
Activation function	tanh
Mini-batch size	30000
Learning rate, Optimization	0.0001, Adam [14]
Number of nodes in FC - rate / scale	150 / 40
Latent Vector $\mathbf{z}$ Dimension - rate / scale	120 / 28

3. CONVOLUTIONAL VAE AND FILTER LEARNING

The block diagram of the VAE model used for filter learning is shown in Figure 1. The use of CVAE is motivated by the goal of learning modulation filters in an unsupervised manner. The kernels (convolutional filters of the two 'Conv' layers in Encoder) of the deep CVAE trained using spectrogram input are interpreted as the modulation filters learned from the data that characterize the key modulations required to generate speech. We train the CVAE in multi-condition fashion with a small number of filters (we use one filter in each convolutional layer). This way, the model is constrained to primarily learn the speech distribution while ignoring the noise distribution.

As outlined in Figure 1, the input to CVAE are the 1-D temporal (spectral) trajectories of log mel spectrograms for rate (scale) filter learning. The mel spectrogram is computed using short-time Fourier transform of speech signal with 25 ms frame length shifted by 10 ms, and the frequency axis is warped to 40 mel-bands. For rate filter learning, the dimension of the 1-D trajectory as the input to CVAE is 1×150 (equivalent to 1.5 s of speech), and for scale filter learning it is 1×40 (corresponding to all 40 mel bands). Table 1 gives the details of the CVAE architecture used in this work. The first layer of the 'encoder' is a convolutional layer with number of kernels $= 1$ and kernel size as 1×5. Let the output of this layer be $\mathbf{h}_1$, where $\mathbf{h}_1 = \mathbf{x} * \mathbf{r}_1$ for rate filter learning and $\mathbf{h}_1 = \mathbf{x} * \mathbf{s}_1$ for scale filter learning. In order to learn multiple irredundant filters, we remove the contribution of the learnt kernel from the input $\mathbf{x}$ using skip connection and feed the $tanh$ of the residual $(\mathbf{x} - \mathbf{h}_1)$ to the next convolutional layer. The next layer (also having one kernel) then learns the modulation characteristics from the residual and generates output $\mathbf{h}_2 = (\mathbf{x} - \mathbf{h}_1) * \mathbf{r}_2$ for rate filter learning and $\mathbf{h}_2 = (\mathbf{x} - \mathbf{h}_1) * \mathbf{s}_2$ for scale filter learning. We add the two filtered (hidden) representations and the non-linear activations $tanh(\mathbf{h}_1 + \mathbf{h}_2)$ is fed to FC layers of the encoder. The latent vector $\mathbf{z}$ is calculated from the encoder output as discussed in Section 2. The decoder then reconstructs the 1-D trajectory from the $\mathbf{z}$ [15].

3.1. Filter Characteristics

The filters $\mathbf{r}_1, \mathbf{r}_2$ $(\mathbf{s}_1, \mathbf{s}_2)$ are iteratively updated using the gradients of the total loss function. The CVAE is trained using multi condi-

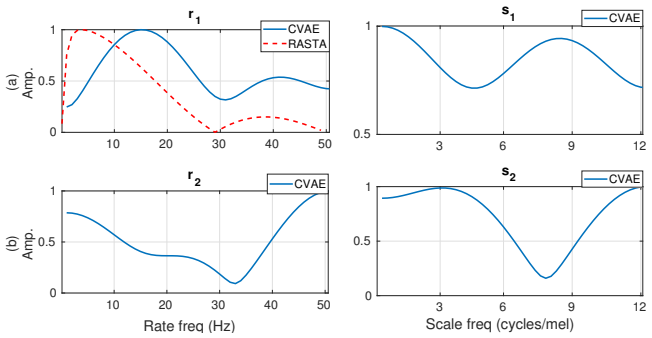

Fig. 2. Frequency modulation characteristics of the two rate ($\mathbf{r}_1, \mathbf{r}_2$) and scale filters ($\mathbf{s}_1, \mathbf{s}_2$) learned from the CVAE model with skip connections using the Aurora-4 dataset. The RASTA filter is also shown in the $\mathbf{r}_1$ plot for reference.

tion data of different databases separately. We begin with random initialization of the filters and the weights and allow the CVAE to learn modulation filter characteristics from data. Figure 2 shows the normalized magnitude frequency response of the filters learned using Aurora-4 database (details of the Aurora-4 dataset are given in Sec. 4). The x axes for the rate and scale filters are rate frequencies (measured in Hz) and scale frequencies (measured in cycles/mel) respectively.

In our analysis, we find that the two rate filters learned from the input mel spectrogram have invariably band-pass and band-stop characteristics. The scale filters jointly span the entire spectral modulation space. We hypothesize that the filters will learn the common underlying representation of all types of input noisy speech, which would be dominated by clean speech. The first row of Fig. 2 also shows the comparison with the RASTA filter [5]. As seen here, the learnt data driven rate filter resembles the perceptual knowledge driven RASTA filter. Also, it is interesting to note that the range of modulations captured by $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{s}_1, \mathbf{s}_2$ are quite similar to the modulation filters found in human perceptual studies [3].

3.2. Comparison with prior work

The previous work in this direction to learn irredundant 1-D and 2-D modulation filters in an unsupervised manner is reported in [11, 16, 17]. In the previous work, the filter learning is performed by learning one filter at a time using convolutional restricted Boltzmann machine (CRBM). The residual is computed externally and is fed to the network for the learning of second filter. Thus, the filters are not jointly optimized. The proposed method in this paper uses a single filter learning framework with CVAE model and skip connections.

3.3. Feature extraction for ASR

The features for ASR are derived by filtering the log mel spectrogram using filters learned from the proposed approach. In this work, we select the rate filter with bandpass characteristic as it has been observed earlier to be crucial for ASR performance [5, 11], while both the scale filters are used for ASR. The log mel spectrograms are filtered using filters ($\mathbf{r}_1, \mathbf{s}_1$) and ($\mathbf{r}_1, \mathbf{s}_2$) separately and are concatenated to derive features for ASR. This is motivated from the neurophysiological evidences suggesting the processing of speech signals along parallel pathways that encode complementary information in the signal [18, 19]. For ASR training, the features are mean-variance normalized at the utterance level before the acoustic model training.

Table 2. Word error rate (%) in Aurora-4 database for multi condition training condition with various feature extraction schemes and the proposed CVAE modulation filtering approach.

Cond	MFB	PFB	ETS	RAS	LDA	MHE	CVAE
A. Clean with same Mic							
Clean	4.2	4.1	4.5	4.6	4.7	4.0	**3.4**
B: Noisy with same Mic							
Airport	7.5	7.9	8.0	8.1	10.1	8.2	6.8
Babble	7.7	7.9	7.9	8.7	9.9	8.6	7.0
Car	4.7	4.9	5.6	5.0	5.8	4.9	4.4
Rest.	9.8	10.2	11.0	11.0	12.6	11.1	8.9
Street	8.6	8.8	10.0	9.0	10.6	8.8	7.9
Train	8.7	8.3	9.3	9.1	10.6	8.4	8.3
Avg.	7.8	8.0	8.6	8.5	9.9	8.3	**7.2**
C: Clean with diff. Mic							
Clean	8.4	7.8	8.0	9.7	10.0	8.1	**7.2**
D: Noisy with diff. Mic							
Airport	19.7	20.9	18.5	20.1	22.3	20.8	17.7
Babble	20.3	20.9	19.3	20.0	22.5	21.3	18.3
Car	11.8	13.1	14.1	12.5	14.5	12.8	10.3
Rest.	21.7	23.7	21.8	23.1	25.2	23.1	19.7
Street	19.1	20.0	19.4	18.9	21.2	20.5	17.2
Train	18.3	19.6	19.6	19.9	21.6	18.9	17.6
Avg.	18.5	19.7	18.8	19.1	21.2	19.6	**16.8**
Avg. of all conditions							
Avg.	12.1	12.7	12.6	12.8	14.4	12.8	**11.0**

4. EXPERIMENTS AND RESULTS

4.1. Kaldi ASR framework

The speech recognition Kaldi toolkit [20] is used for building the ASR. For the ASR experiments on Aurora-4 and CHiME-3 Challenge, we use a deep neural network (DNN) with 6 hidden layers having 21 frames of input temporal context. The baseline model does not incorporate any modulation filtering. The DNNs with sigmoid nonlinearity are layer wise pre-trained with a deep belief network. Then, the models are discriminatively trained using the training data with cross entropy loss. A hidden Markov model - Gaussian mixture model (HMM-GMM) system is used to generate the alignments for training the DNN based model and a tri-gram language model is used in the ASR decoding. The performance of the ASR system is analyzed using word-error-rate (WER).

We compare the ASR performance of the proposed modulation filtering approach (CVAE) with traditional mel filter bank energy (MFB) features, power normalized filter bank energy (PFB) features [21], advanced ETSI front-end (ETS) [22], RASTA features (RAS) [5], LDA based features (LDA) [9], and MHEC features (MHE) [23]. In particular, the RASTA features (RAS) and LDA features are included as they both perform modulation filtering in the temporal domain using a knowledge driven filter and a supervised data driven filter, respectively.

4.2. Aurora-4 ASR

The Wall Street Journal (WSJ) Aurora-4 corpus consists of continuous read speech recordings of 5000 words corpus, recorded under clean and noisy conditions (street, train, car, babble, restaurant, and airport noises at $10-20$ dB SNR). The training data has 7138 multi condition recordings from 84 speakers. The validation data has 1206 recordings for multi condition setup. The test data has 330 recordings (8 speakers) for each of the 14 test conditions (clean and noisy).

Table 3. Word error rate (%) in CHiME-3 Challenge database for multi-condition training (real+simulated) with test data from simulated and real noisy environments.

Test Cond	MFB	PFB	RAS	MHE	CVAE
Sim_dev	14.3	13.7	14.6	14.4	**12.2**
Real_dev	11.6	12.0	11.8	12.0	**9.6**
Avg.	13.0	12.9	13.2	13.2	**10.9**
Sim_eval	25.5	25.1	23.1	26.4	**19.1**
Real_eval	22.6	23.0	21.6	22.9	**17.9**
Avg.	24.1	24.1	22.4	24.7	**18.5**

Table 4. WER (%) for each noise condition in CHiME-3 dataset with the baseline features and the proposed feature extraction.

Dev Data				
Cond.	Sim		Real	
	MFB	CVAE	MFB	CVAE
BUS	12.6	**10.6**	14.2	**11.5**
CAF	17.0	**15.7**	11.4	**9.8**
PED	12.0	**9.9**	8.5	**7.0**
STR	15.7	**12.5**	12.3	**10.3**
Eval Data				
Cond.	Sim		Real	
	MFB	CVAE	MFB	CVAE
BUS	18.3	**13.3**	29.2	**22.8**
CAF	26.3	**20.5**	23.7	**18.0**
PED	29.1	**20.6**	21.1	**16.8**
STR	28.3	**22.1**	16.4	**13.9**

The test data is classified into four groups, A - clean data, B - noisy data, C - clean data with channel distortion, and D - noisy data with channel distortion.

The results of various ASR experiments on Aurora-4 dataset is shown in Table. 2. The ASR results have also been separately reported for different noisy conditions (conditions A,B,C,D). As seen in this Table, most of the noise robust front-ends do not improve over the baseline mel-filter bank (MFB) performance, as the acoustic models are trained using multi-condition noisy training data. The proposed features provide significant improvements in ASR performance over the baseline system (average relative improvements of 9%). Furthermore, the improvements in ASR performance are consistently seen across all the noisy conditions of Aurora-4 dataset.

4.3. CHiME-3 ASR

The CHiME-3 corpus for ASR contains multi-microphone tablet device recordings from everyday environments, released as a part of 3rd CHiME challenge [24]. Four varied environments are present, cafe (CAF), street junction (STR), public transport (BUS) and pedestrian area (PED). For each environment, two types of noisy speech data are present, real and simulated. The real data consists of 6-channel recordings of sentences from the WSJ0 corpus spoken in the environments listed above. The simulated data was constructed by artificially mixing clean utterances with environment noises. The training data has 1600 (real) noisy recordings and 7138 simulated noisy utterances. We use the beamformed audio for filter learning using CVAE, and for ASR training and testing. The development (_dev) and evaluation (_eval) data consists of the 410 and 330 utterances respectively. For each set, the sentences are read by four different talkers in the four CHiME-3 environments. This results in 1640 (410×4) and 1320 (330×4) real development and evaluation utterances in total. Identically-sized, simulated dev and eval sets are made by mixing recordings captured in the recording booth with the environmental noise recordings.

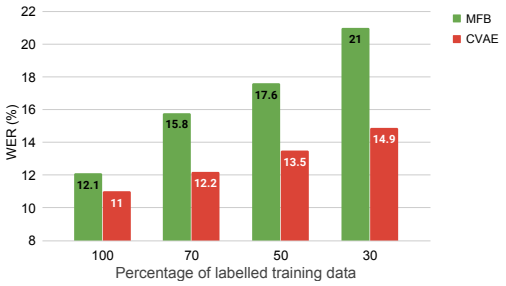

Fig. 3. ASR performance in terms of WER (%) in Aurora-4 database (average of 14 test conditions) for multi condition training using lesser amount of labeled training data (70%, 50%, 30%).

The results for the CHiME-3 dataset are reported in Table 3. The proposed approach to feature extraction provides significant improvements over the baseline system as well as the other noise robust front-ends considered here. On the average, the proposed approach provides relative improvements of 16% in the development set and 23% in the evaluation set. The detailed results on different noises in CHiME-3 are reported in Table 4. For all the noise conditions in CHiME-3 in simulated and real environments, the proposed approach shows significant improvements over the baseline MFB features. In the evaluation dataset, the relative improvements over the baseline features for most of the noise conditions are above 22%.

4.4. Semi-supervised training

The semi-supervised ASR requires modeling speech without labels and then utilizing minimal supervision for ASR training. In many real-world scenarios, collection of noisy data may be relatively easy while the labeling process may be quite cumbersome and expensive. For semi-supervised ASR training, the Aurora-4 training set up is used with 70, 50 and 30% of the labeled training data. The modulation filters are learnt using full unsupervised training data. The performance comparison of ASR with semi-supervised training is shown in Figure 3 only for MFB features (as MFB features performed relatively better than other features in Table 2) and the proposed CVAE feature scheme (average WER of all 14 test data conditions). These results indicate that the proposed features are much more resilient to reduced amounts of labeled training data as compared to the baseline system. These features perform significantly better than MFB features (relative improvement of 29% over the baseline for the case with 30% labelled training data).

5. SUMMARY

The major contributions from this work are as follows:

- Proposal of an unsupervised data-driven approach to learn spectral and temporal modulation filters with a random initialization.
- Obtaining multiple irredundant data-driven filters using skip connection approach in deep variational network.
- 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.

6. REFERENCES

[1] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[2] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7398–7402.

[3] Taffeta M Elliott and Frédéric E Theunissen, "The modulation transfer function for speech intelligibility," *PLoS computational biology*, vol. 5, no. 3, pp. e1000302, 2009.

[4] Nima Mesgarani, Malcolm Slaney, and Shihab 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.

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

[6] Michael Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[7] Tony Ezzat, Jake Bouvrie, and Tomaso Poggio, "Spectro-temporal analysis of speech using 2-d gabor filters," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[8] György Kovács, László Tóth, and Dirk Van Compernolle, "Selection and enhancement of gabor filters for automatic speech recognition," *International Journal of Speech Technology*, vol. 18, no. 1, pp. 1–16, 2015.

[9] Sarel van Vuuren and Hynek Hermansky, "Data-driven design of RASTA-like filters.," in *Eurospeech*, 1997, vol. 1, pp. 1607–1610.

[10] Jeih-Weih Hung and Lin-Shan 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.

[11] Purvi Agrawal and Sriram Ganapathy, "Unsupervised modulation filter learning for noise-robust speech recognition," *Journal of the Acoustical Society of America*, vol. 142, no. 3, pp. 1686–1692, 2017.

[12] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[14] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[15] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International Conference on Artificial Neural Networks*. Springer, 2011, pp. 52–59.

[16] Purvi Agrawal and Sriram Ganapathy, "Speech representation learning using unsupervised data-driven modulation filtering for robust ASR," in *INTERSPEECH*, 2017, pp. 2446–2450.

[17] Purvi Agrawal and Sriram Ganapathy, "Comparison of unsupervised modulation filter learning methods for ASR," in *Proc. INTERSPEECH*, 2018, pp. 2908–2912.

[18] Gregory Hickok and David Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393, 2007.

[19] Taishih Chi, Powen Ru, and Shihab A Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[20] Daniel Povey et al., "The KALDI speech recognition toolkit," in *IEEE ASRU*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

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

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

[23] Seyed Omid Sadjadi, Taufiq Hasan, and John HL Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[24] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.