Robust Raw Waveform Speech Recognition Using Relevance Weighted Representations

Purvi Agrawal and Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) Lab, Dept. of Electrical Engg.,
Indian Institute of Science, Bengaluru-560012, India.

(purvia, sriramg)@iisc.ac.in

Abstract

Speech recognition in noisy and channel distorted scenarios is often challenging as the current acoustic modeling schemes are not adaptive to the changes in the signal distribution in the presence of noise. In this work, we develop a novel acoustic modeling framework for noise robust speech recognition based on relevance weighting mechanism. The relevance weighting is achieved using a sub-network approach that performs feature selection. A relevance sub-network is applied on the output of first layer of a convolutional network model operating on raw speech signals while a second relevance sub-network is applied on the second convolutional layer output. The relevance weights for the first layer correspond to an acoustic filterbank selection while the relevance weights in the second layer perform modulation filter selection. The model is trained for a speech recognition task on noisy and reverberant speech. The speech recognition experiments on multiple datasets (Aurora-4, CHiME-3, VOiCES) reveal that the incorporation of relevance weighting in the neural network architecture improves the speech recognition word error rates significantly (average relative improvements of 10% over the baseline systems).

Index Terms: Raw speech waveform, relevance weighting, cosine-modulated Gaussian filterbank, speech recognition.

1. Introduction

The broad set of methods that enable the learning of meaningful representations for a given data are referred to as representation learning methods. This can be unsupervised like principal components or supervised like linear discriminant analysis. With the growing interest in deep learning, representation learning using deep neural networks has been actively pursued. While a lot of success has been reported for text and other domains (for example, using word2vec models [1]), representation learning for speech is still challenging. This paper explores representation learning for speech using a novel modeling approach.

In the past, the main direction pursued has been to learn filterbank parameters [2–4] from raw waveforms. The objective can be either detection or classification [3, 5, 6]. Some of the efforts also attempt unsupervised learning of filterbank, eg. Sailor et. al [7] uses restricted Boltzmann machine while Agrawal et. al. [8] uses variational autoencoders. The wav2vec method by Schneider et. al. in [9] explores unsupervised pretraining for speech recognition by learning representations of raw audio. There has been some attempts to explore interpretability of acoustic filterbank recently, for eg. SincNet filterbank [10, 11]. However, compared to vector representations of text which have shown to embed meaningful semantic properties, the interpretability of speech representations from these approaches has often been limited. Further, most of the state-of-the-art systems continue to use mel filterbank [12] features.

The approach of modulation filter learning (modulation filters process the time-frequency representation and perform filtering along time (rate) and frequency (scale) dimensions) using the linear discriminant analysis (LDA) has been explored to learn the temporal modulation filters in a supervised manner [13, 14]. There have also been attempts to learn modulation filters in an unsupervised manner [15–18].

In this paper, we propose a relevance weighting mechanism that allows the interpretability of the learned representations in the forward propagation itself. The relevance weighting scheme is popular in text domain in applications such as document search, where a static relevance weight is attached to each document, based on the search term feature [19, 20]. A similar application of visual attention in the image domain uses spatial weighting to weigh different parts of the image [21]. A related work is the deep mixture of experts (MoE) model that trains multiple expert networks, each of which specializes in a different part of the input space and a gating network decides which expert to use for each input region [22]. In this work, the relevance weighting on the learned representations is achieved using a sub-network.

We propose a speech representation learning method using a two-step relevance weighting approach. The first step performs relevance weighting on the output of the first convolutional layer that learns acoustic filterbank from the raw waveform. The acoustic filters are parametric cosine-modulated Gaussian filters whose parameters are learned within the acoustic model [8]. The output is fed to the relevance sub-network to obtain the relevance weights for the filterbank outputs. The weighted filterbank representation is used as input to the second convolutional layer which is interpreted as a modulation filtering step. The kernels of the second convolutional layer are 2-D spectro-temporal modulation filters and the filtered representations are weighted using another relevance sub-network. The rest of the architecture performs the task of acoustic modeling for automatic speech recognition (ASR). All the model parameters are learned in a supervised fashion. The ASR experiments are conducted on Aurora-4 (additive noise with channel artifact) [23], CHiME-3 (additive noise with reverberation) [24] and VOiCES (additive noise with reverberation) [25] databases. The experiments show that the learned representations from the proposed framework provides considerable improvements in ASR results over the baseline methods.

The rest of the paper is organized as follows. Section 2 describes the proposed two-step representation learning approach using relevance weighting. Section 3 describes the ASR experiments with the various front-ends followed by a summary.

This work was partly funded by grants from the Department of Science and Technology project DST0 (ECR01341), Govt. of India, and Indian Institute of Science.

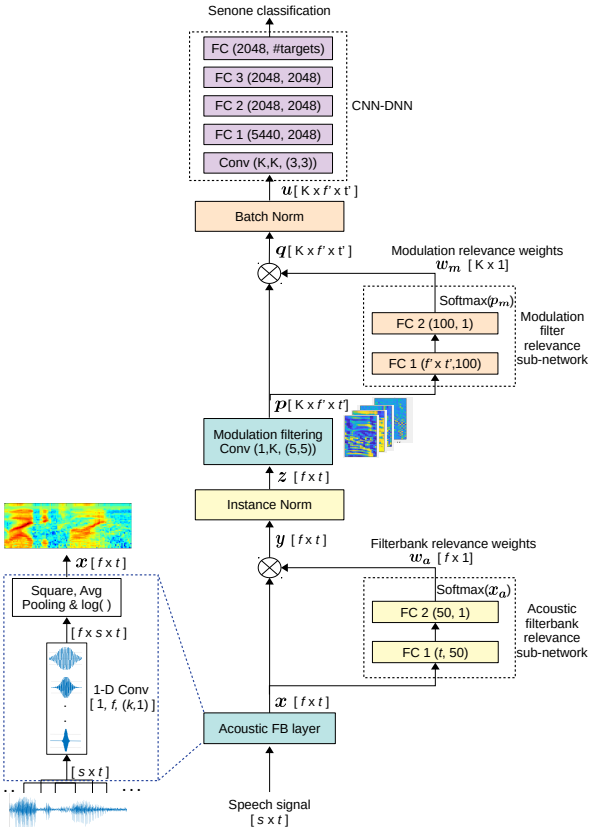

Figure 1: *Block diagram of the proposed representation learning from raw waveform using relevance weighting approach. Here, FC represents fully connected layer.*

2. Relevance Based Representation Learning

The block schematic of the proposed relevance weighting based two-step representation learning model is shown in Figure 1.

2.1. Step-1: Acoustic Filterbank representation

The input to the neural network are raw samples windowed into s samples per frame with a contextual window of t frames. Each block of s samples is referred to as a frame. This matrix of size $s \times 1$ raw audio samples are processed with a 1-D convolution using f kernels (f denotes the number of sub-bands in filterbank decomposition) each of size k. The kernels are modeled as cosine-modulated Gaussian function [8],

$$g_i(n) = \cos 2\pi\mu_i n \times \exp\left(-n^2\mu_i^2/2\right) \quad (1)$$

where $g_i(n)$ is the i-th kernel ($i = 1, .., f$) at time n, μ_i is the center frequency of the ith filter (in frequency domain). The parametric approach to filterbank (FB) learning generates filters with a smooth frequency response. The mean parameters are updated in a supervised manner for each dataset. The convolution with the cosine-modulated Gaussian filters generates f feature maps. These outputs are squared, average pooled within each frame and log transformed. This generates $\boldsymbol{x}$ as f dimensional features for each of the t contextual frames, as shown in Figure 1. The $\boldsymbol{x}$ can be interpreted as the "learned" time-frequency representation (spectrogram). We refer to the first layer as the acoustic filterbank (FB) layer.

2.2. Acoustic FB relevance weighting

The relevance weighting paradigm for acoustic FB layer is implemented using a relevance sub-network fed with the $f \times t$ time-frequency representation $\boldsymbol{x}$. A two layer deep neural network (DNN) with a softmax output generates acoustic FB relevance weights $\boldsymbol{w}_a$ as f dimensional vector with weights corresponding to each sub-band filter. Let the output from the relevance sub-network be denoted as $\boldsymbol{x}_a$, then the relevance weights $\boldsymbol{w}_a$ are generated using the softmax function as,

$$w_a^i = \frac{e^{x_a^i}}{\sum_j e^{x_a^j}}; \quad \text{where } i = 1, 2, ..., f. \quad (2)$$

These weights $\boldsymbol{w}_a$ are multiplied element-wise with each frame of $\boldsymbol{x}$ to obtain weighted filterbank representation $\boldsymbol{y}$. The relevance weights in the proposed framework are different from typical relevance weights used in text search problem [20] as well as the attention mechanism [26]. In proposed framework, relevance weighting is applied on the representation as soft feature selection weights without performing a linear combination. We also smooth the first layer outputs ($\boldsymbol{y}$) using instance norm [27, 28]. Let $y_{j,i}$ denote the relevance weighted filterbank output for frame j ($j = 1, .., t$) of sub-band i ($i = 1, .., f$). The soft weighted output $z_{j,i}$ is given as,

$$z_{j,i} = \frac{y_{j,i} - m_i}{\sqrt{\sigma_i^2 + c}} \quad (3)$$

where m_i is the sample mean of $y_{j,i}$ computed over j and σ_i is the sample std. dev. of $y_{j,i}$ computed over j. The constant c is $1e-4$. The output of relevance weighting ($\boldsymbol{z}$) is propagated to the subsequent layers for the acoustic modeling.

In our experiments, we use $t = 101$ whose center frame is the senone target for the acoustic model. We also use $f = 80$ sub-bands and acoustic filter length $k = 129$. This value of k corresponds to 8 ms in time for a 16 kHz sampled signal which has been found to be sufficient to capture temporal variations of speech signal [29]. The value of s is 400 corresponding to 25 ms window length and the frames are shifted every 10ms. Thus, the input to the acoustic filter bank layer with $t = 101$ contains about 1 sec. of audio segment. In our experiments, we also find that after the normalization layer, the number of frames t can be pruned to the center 21 frames for the acoustic model training without loss in performance. This has significant computational benefits and the pruning is performed to keep only the 21 frames around the center frame (200 ms of context).

The soft relevance weighted time-frequency representation $\boldsymbol{z}$ obtained from the proposed approach is shown in Figure 2(c) for an utterance with airport noise from Aurora-4 dataset (the waveform is plotted in Figure 2(a)). The corresponding mel spectrogram (without relevance weighting) is plotted in Figure 2(b). It can be observed that, in the learned filterbank representation (Figure 2(c)), the formant frequencies appear to be shifted upwards because of the increased number of filters in the lower frequency region. Also, the relevance weighting modifies the representations propagated to the higher layers.

2.3. Step-2: Relevance Weighting of Modulation Filtered Representation

The representation $\boldsymbol{z}$ from acoustic filterbank layer is fed to the second convolutional layer which is interpreted as modulation filtering layer (shown in Figure 1). The kernels of this convolutional layer are interpreted as 2-D spectro-temporal modulation

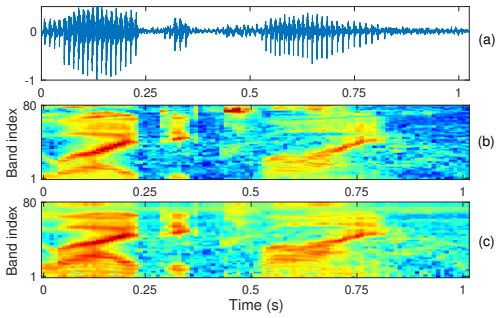

Figure 2: *(a) Speech signal from Aurora-4 dataset with airport noise, (b) mel spectrogram representation (c) acoustic FB representation with soft relevance weighting ($\boldsymbol{z}$ in Figure 1).*

filters, learning the rate-scale characteristics from the data. This step is partly inspired by the neuro-physiological evidences of multi-stream feature framework for ASR [30, 31]. The modulation filtering layer generates K parallel streams, corresponding to K modulation filters $\boldsymbol{w}_K$. The modulation filtered representations $\boldsymbol{p}$ are max-pooled with window of 3×1, leading to feature maps of size $f' \times t'$. These are weighted using a second relevance weighting sub-network (referred to as the modulation filter relevance sub-network in Figure 1). Let $\boldsymbol{p}_m$ denote the K-dimensional output of modulation filter relevance sub-network. The softmax function is applied on the output to generate modulation relevance weights $\boldsymbol{w}_m$ over K modulation filters,

$$w_m^i = \frac{e^{p_m^i}}{\sum_j e^{p_m^j}}; \text{ where } i = 1, 2, ..., K. \quad (4)$$

The weights are multiplied with the representation $\boldsymbol{p}$ to obtain weighted representation $\boldsymbol{q}$. This weighting is interpreted as the selection of different modulation filtered representations (with different rate-scale characteristics). The resultant weighted representation $\boldsymbol{q}$ is fed to the batch normalization layer [32]. The training data statistics of batch norm, including affine parameters, are used in the test phase. The value of the normalization factor c in denominator for batch norm is chosen to be 10^{-4} empirically. We use the value of $K = 40$ in the work. Following the acoustic filterbank layer and the modulation filtering layer (including the relevance sub-networks), the acoustic model consists of series of CNN and DNN layers. The configuration details are given in Figure 1.

The proposed two stage processing is loosely modeled based on our understanding of the human auditory system, where the cochlea performs acoustic frequency analysis while early cortical processing performs modulation filtering [30]. The relevance weighting mechanism attempts to model the feature selection/weighting inherently present in the auditory system (based on the relative importance of the representation for the downstream task).

3. Experiments and Results

The speech recognition system is trained using PyTorch [33] while the Kaldi toolkit [34] is used for decoding and language modeling. The ASR is built on three datasets, Aurora-4, CHiME-3 and VOiCES respectively. The models are discriminatively trained using the training data with cross entropy loss and Adam optimizer [35]. A hidden Markov model - Gaussian mixture model (HMM-GMM) system is used to generate the senone alignments for training the CNN-DNN based model. The ASR results are reported with a tri-gram language model and the best language model weight is obtained from the development set.

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

Cond	MFB	PFB	Sinc	A	MFB-R	A-R	S-R,M-R	A-R,M-R
A. Clean with same Mic								
Clean	4.2	4.0	4.0	4.1	4.0	**3.6**	3.8	**3.6**
B: Noisy with same Mic								
Airport	6.8	7.1	6.9	6.4	7.0	6.0	6.3	**5.9**
Babble	6.6	7.4	6.7	6.3	6.8	**6.1**	6.2	**6.1**
Car	4.0	4.5	4.0	4.0	4.2	4.0	**3.9**	**3.9**
Rest.	9.4	9.6	9.4	8.5	9.4	7.7	8.4	**6.8**
Street	8.1	8.1	8.4	7.8	8.0	7.1	7.5	**6.9**
Train	8.4	8.6	8.3	7.9	8.6	7.3	7.4	**7.2**
Avg.	7.2	7.5	7.3	6.8	7.3	6.4	6.6	**6.1**
C: Clean with diff. Mic								
Clean	7.2	7.3	7.3	7.3	7.1	8.1	6.8	**6.0**
D: Noisy with diff. Mic								
Airport	16.3	18.0	16.2	17.3	16.6	15.4	**13.9**	14.1
Babble	16.7	18.9	17.6	17.4	16.7	16.0	16.0	**15.4**
Car	8.6	11.2	9.0	9.0	9.0	9.4	7.9	**7.7**
Rest.	18.8	21.0	19.0	18.2	18.5	**16.9**	19.2	18.6
Street	17.3	19.5	17.3	17.8	17.5	16.9	**16.6**	16.8
Train	17.6	18.8	18.1	17.8	18.1	**16.2**	16.6	**16.2**
Avg.	15.9	17.9	16.2	16.2	16.1	15.1	15.0	**14.8**
Avg. of all conditions								
Avg.	10.7	11.7	10.8	10.7	10.8	10.0	10.0	**9.6**

For each dataset, we compare the ASR performance of the proposed approach of learning acoustic representation from raw waveform with acoustic FB (A) with relevance weighting (A-R) and modulation FB (M) with relevance weighting (M-R) denoted as (A-R,M-R), with the model having only the acoustic FB relevance weighting (A-R), traditional mel filterbank energy (MFB) features, and power normalized filterbank energy (PFB) features [36]. For CHiME-3 dataset, we also compare with RASTA features that perform modulation filtering (RAS) [37], and mean Hilbert envelope (MHE) features [38]. All the baseline features are processed with cepstral mean and variance normalization (CMVN) on a 1 sec. running window. The neural network architecture shown in Figure 1 (except for the acoustic filterbank learning layer, the acoustic FB relevance sub-network and modulation filter relevance sub-network) is used for all the baseline features.

3.1. Aurora-4 ASR

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 7138 multi condition recordings (84 speakers) with total 15 hours of training data. The validation data has 1206 recordings for multi condition setup. The test data has 330 recordings (8 speakers) for each of the 14 clean and noise conditions. The test data are classified into group A - clean data, B - noisy data, C - clean data with channel distortion, and D - noisy data with channel distortion.

The ASR performance on the Aurora-4 dataset is shown in Table 1 for each of the 14 test conditions. We also compare the ASR performance with the acoustic filterbank representation (A) without relevance weighting. In addition, we also experiment with the application of the relevance weighting over pre-trained mel filterbank features (MFB-R).

As seen in the results, most of the noise robust front-ends do not improve over the baseline mel filterbank (MFB) performance. The raw waveform acoustic FB performs similar to MFB baseline features on average while performing better than the baseline for Cond. A and B. The MFB-R features, which denote the application of the acoustic FB relevance weighting

Table 2: *Word error rate (%) in CHiME-3 Challenge database for multi-condition training (real+simulated).*

Test Cond	MFB	PFB	RAS	MHE	A-R	A-R,M-R
Sim_dev	12.9	13.3	14.7	13.0	12.5	**12.0**
Real_dev	9.9	10.7	11.4	10.2	9.9	**9.6**
Avg.	11.4	12.0	13.0	11.6	11.2	**10.8**
Sim_eval	19.8	19.4	22.7	19.7	19.2	**18.5**
Real_eval	18.3	19.2	20.5	18.5	17.3	**16.6**
Avg.	19.1	19.3	21.6	19.1	18.2	**17.5**

Table 3: *WER (%) for cross-domain ASR experiments.*

Filters Learned on	**ASR Trained and Tested on**	
	Aurora-4	CHiME-3
Aurora-4	9.6	14.3
CHiME-3	9.7	14.2

over mel filterbank features, also doesn't improve over baseline MFB features. The features with acoustic filterbank learning + relevance weighting (A-R) improves over the raw (A) features with average relative improvements of 6%. The proposed (A-R,M-R) representation learning (two-stage relevance weighting) provides considerable improvements in ASR performance over the baseline system with average relative improvements of 11% over the baseline MFB features. Furthermore, the improvements in ASR performance are consistently seen across all the noisy test conditions.

We also compare with the SincNet method [10] where our cosine modulated Gaussian filterbank is replaced with the sinc filterbank [1] as kernels in first convolutional layer (acoustic FB layer in Fig. 1). The ASR system with sinc FB (Sinc) is trained jointly without any relevance weighting, and with 2-stage relevance weighting (S-R,M-R) keeping rest of the architecture same as shown in Fig. 1. From results in Table 1, it can be observed that the parametric sinc FB (without weighting) performs similar to MFB and our acoustic FB features (A). The relevance weighting over sinc FB (S-R,M-R) improves over the baseline MFB with average relative improvements of 6%.

3.2. 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, constituting a total of 18 hours of training data. We use the beamformed audio in our ASR training and testing. The development (dev) and evaluation (eval) data consists of 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.

The results for the CHiME-3 dataset are reported in Table 2. The initial approach of raw waveform filter learning with acoustic FB relevance weighting improves over the baseline system as well as the other noise robust front-ends considered here. The proposed approach of 2-stage relevance weighting over learned acoustic and modulation representations provides significant improvements over baseline features. On the average, the proposed approach provides relative improvements of 10% over MFB features in the eval set.

[1] https://github.com/mravanelli/SincNet/

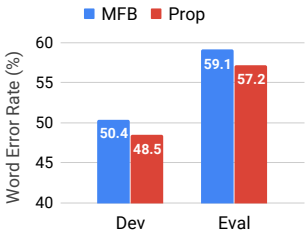

Figure 3: *ASR performance in WER (%) for VOiCES database.*

3.3. Representation transfer across tasks

In a subsequent analysis, we perform a cross-domain ASR experiment, i.e., we use the acoustic filterbank learned from one of the datasets (either Aurora-4 or CHiME-3 challenge) to train/test ASR on the other dataset. The results of these cross-domain filter learning experiments are reported in Table 3. The rows in the table show the database used to learn the acoustic FB and the columns show the dataset used to train and test the ASR (all other layers in Figure 1 are learned in the ASR task). The performance reported in this table are the average WER on each of the datasets. The results shown in Table 3 illustrate that the filter learning process is relatively robust to the domain of the training data, which suggest that the proposed representation learning approach can be generalized for other "matched" tasks.

3.4. VOiCES ASR

The Voices Obscured in Complex Environmental Settings (VOiCES) corpus is a creative commons speech dataset being used as part of VOiCES Challenge [25]. The training data set of 80 hours has 22, 741 utterances sampled at 16kHz from 202 speakers, with each utterance having $12 - 15$s segments of read speech. We performed a 1-fold reverberation and noise augmentation of the data using Kaldi [34]. The ASR development set consists of 20 hours of distant recordings from the 200 VOiCES dev speakers. It contains recordings from 6 microphones. The evaluation set consists of 20 hours of distant recordings from the 100 VOiCES eval speakers and contains recordings from 10 microphones. The ASR performance of VOiCES dataset with baseline MFB features and our proposed approach of 2-step relevance weighting is reported in Figure 3. These results suggest that proposed model is also scalable to relatively larger ASR tasks with large vocabulary where consistent improvements can be obtained with the proposed approach.

4. Summary

The key contributions of the work are:

- Proposing a novel relevance weighted representation learning neural architecture for speech modeling.
- Modeling of 2-stage relevance weighting based architecture over learnt acoustic filterbank features from raw waveform.
- Improved acoustic modeling illustrated using performance gains in word error rates for multiple ASR tasks.

5. References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proc. of International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1301.3781*, 2013.

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

[3] 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.

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

[5] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4624–4628.

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

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

[8] P. Agrawal and S. Ganapathy, "Unsupervised raw waveform representation learning for ASR," *Proc. of Interspeech 2019*, pp. 3451–3455, 2019.

[9] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Proc. of Interspeech*, pp. 3465–3469, 2019.

[10] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with SincNet," *in Proc. of Neural Information Processing Systems (NIPS)*, 2018.

[11] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *Proc. of Interspeech*, pp. 161–165, 2019.

[12] 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.

[13] S. Van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters." in *Eurospeech*, vol. 1, 1997, pp. 1607–1610.

[14] 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.

[15] H. B. Sailor and H. A. Patil, "Unsupervised learning of temporal receptive fields using convolutional rbm for asr task," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 873–877.

[16] P. Agrawal and S. 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.

[17] ——, "Modulation filter learning using deep variational networks for robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 244–253, 2019.

[18] ——, "Deep variational filter learning models for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5731–5735.

[19] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information science*, vol. 27, no. 3, pp. 129–146, 1976.

[20] S. E. Robertson and S. Walker, "On relevance weights with little relevance information," in *Proc. of the 20th annual international ACM SIGIR Conference on Research and development in information retrieval*, 1997, pp. 16–24.

[21] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[22] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," *arXiv preprint arXiv:1312.4314*, 2013.

[23] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[24] J. Barker, R. Marxer, E. Vincent, and S. 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.

[25] M. K. Nandwana, J. Van Hout, C. Richey, M. McLaren, M. A. Barrios, and A. Lawson, "The VOiCES from a distance challenge 2019," *Proc. of Interspeech*, pp. 2438–2442, 2019.

[26] Y. Zhang, P. Zhang, and Y. Yan, "Attention-based LSTM with multi-task learning for distant speech recognition," *Proc. of Interspeech*, pp. 3857–3861, 2017.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, p. 533, 1986.

[28] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

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

[30] 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.

[31] 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.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proc. of International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.

[33] A. Paszke, S. Gross, S. Chintala, and G. Chanan, "PyTorch," *Computer software. Vers. 0.3*, vol. 1, 2017.

[34] D. Povey *et al.*, "The KALDI speech recognition toolkit," in *IEEE ASRU*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. of International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1412.6980*, 2015.

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

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

[38] S. O. Sadjadi, T. Hasan, and J. H. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition," in *Proc. of Interspeech*, 2012.