Unsupervised Representation Learning for Robust Speech Recognition

Purvi Agrawal

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

purvia@iisc.ac.in

Abstract

The performance of the state-of-the-art automatic speech recognition (ASR) systems remain fragile in high levels of noise and reverberation. The noise robustness can be partly addressed by multi-condition training (utilizing noisy training data from multiple environments) [1]. 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. In this work, various unsupervised speech representation learning approaches are explored, proposed and compared for robust speech recognition system. The work can be broadly categorized in two parts for representation learning: robust representation learning (processing the spectrogram representation) and representation learning from raw waveform.

1. Representation learning for robust ASR

The first part of the work deals with obtaining robust representations for ASR system. It is motivated by the auditory processing studies having shown the importance of modulation filtered spectrogram representations in human speech recognition [2]. It involves filtering the input speech spectrogram along the temporal and spectral axis with filters that retain only relevant speech information. Inspired by these evidences, we propose speech representation learning paradigm using data-driven 1-D and 2-D spectro-temporal modulation filtering technique [3, 4]. In particular, modulation filters are learned in unsupervised manner using various generative models.

The framework of unsupervised learning can be divided into distribution learning, representation learning or clustering methods [5]. We use convolutional restricted Boltzmann machine (CRBM) as distribution learning method for unsupervised modeling [3, 6]. The RBM model assumes a Boltzmann distribution for the joint density function of the observation and latent variable. An autoencoder (AE) is a neural network which aims at representation learning at the hidden layers by mapping the input to the output using mean square error (MSE) cost [7]. In this work, we explore the use of convolutional autoencoder (CAE) incorporating convolutional layers in an AE for modulation filter learning [8]. A second approach of representation learning using generative adversarial network (GAN) attempts to modify the CAE approach with an additional adversarial cost function. It also aim to use a latent data distribution to generate the observed data. The model uses a discriminative loss function (fake versus real) to further correct the generative model [9]. The convolutional variational autoencoder (CVAE) minimizes the MSE of the reconstructed data along with a latent loss [10, 11].

The learning of modulation filters using generative modeling framework is shown in block diagram in Fig. 1. The input to the network is either 1-D temporal and spectral trajectories [3] or 2-D patches of the speech spectrograms to learn spectro-temporal characteristics jointly [4, 11]. The modulation filtering layer incorporates convolutional layer with the aim of learning non-overlapping irredundant set of modulation filters. We explore and propose several approaches to learn non-overlapping filters - residual approach (similar to matching pursuit algorithm) [3], modified cost function [11], and skip connection based residual approach [12]. The output of the modulation filtering layer is then fed to the either of the generative model and the network parameters (including modulation filters) are iteratively updated using the gradients of the respective loss function.

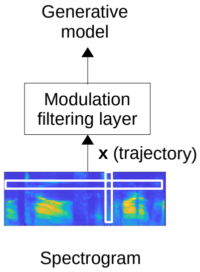

Figure 1: *Block diagram to learn modulation filters from spectrogram in Modulation filtering layer of generative model.*

The learnt modulation filters are then used to process the log Mel spectrogram features and the filtered spectrograms are used as features for ASR experiments on noisy and reverberant speech. In our work, we select the temporal modulation filter with bandpass characteristic while both the spectral filters are used for ASR [3]. While this was partly motivated by the previous studies on human perception of modulation [13], we observed and validate the claim that the important modulations for ASR lie in the bandpass region of temporal domain and the entire modulation range of the spectral domain, much similar to the human perceptual experiments [2].

The proposed approach is compared with knowledge based filtering approach such as RASTA filtering [13] and other noise-robust front ends. Several speech recognition experiments are performed on a set of tasks consisting of databases with additive noise with channel artifacts (Aurora-4), reverberation (REVERB Challenge) and additive noise with reverberation (CHiME-3). In these experiments, the proposed framework shows significant improvements over the baseline (spectrogram) features as well as various other noise robust front-ends. The different approaches are compared as well and the results indicate that the generative modeling framework of CVAE provides the best ASR performance in comparison with other models [11].

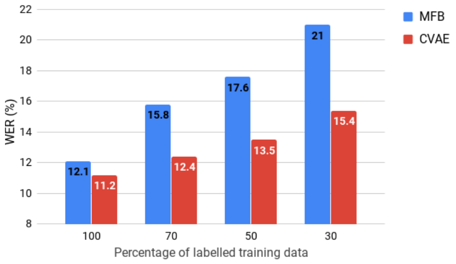

Figure 2: *ASR performance in terms of word error rate (WER in %) in Aurora-4 database (average of all test conditions) using lesser amount of labeled training data (70%, 50%, 30%).*

Semi-supervised ASR training: The application of the proposed filtering for semi-supervised ASR training is investigated where reduced amounts of labeled training data is available for ASR. This is partly motivated by the fact that, while data collection in real noisy environments may be relatively easy, the labeling of noisy data is cumbersome and more expensive than in clean recording conditions. Given the unsupervised learning paradigm of the proposed approach, the modulation filters could be learned from the entire unlabeled training data and applied for ASR training with the labeled data. We validate the ASR experiments with reduced labeled data (70, 50 and 30% random selection of the original training data) and compare the performance with baseline ASR system. We observed that the baseline ASR system of log Mel spectrogram (MFB) features has a drastic degradation in performance when the amount of training data is reduced, as shown in Fig. 2. The proposed CVAE features using the learnt modulation filters are more resilient to the reduced amounts of labelled training data for ASR.

Domain specific versus cross-domain filter learning: In a subsequent analysis, we perform a cross-domain ASR experiment, i.e., we learn the filters from one of the datasets (either Aurora-4, REVERB Challenge or CHiME-3) and use those filters to train/test ASR on the other two datasets. The results suggests that the filter learning process is relatively robust to the domain of the training data used in the CVAE model.

2. Raw waveform representation learning

The next part of the work extends the unsupervised representation learning approach directly from raw speech waveform [14]. In particular, the representation learning is carried out as a two-layer process in CVAE, as shown in Fig. 3. First, an acoustic filterbank is learnt from the raw waveforms using the first convolutional layer in CVAE. We use cosine-modulated Gaussian functions as acoustic filters with center frequency and bandwidth as the learnable parameters and random initialization as the starting point. The convolution is carried out in time domain, and the output of the layer is pooled and log transformed to obtain time-frequency representation. The next layer learns the spectral and temporal modulation filters from the obtained representations as discussed in the first part of the work.

The experiments are performed on Aurora-4 and CHiME-3 dataset. The acoustic filters in the acoustic filterbank layer and the filters in the modulation filtering layer are iteratively updated using the gradients of the total loss function. The CVAE is trained using data of different databases separately. From the filter characteristics, we observed that the learned filterbank also has nonlinear relationship between center frequencies and the filter index with more number of filters in lower frequencies compared to higher frequencies, similar to traditional acoustic filterbanks, shown in Fig. 4. For both databases, the time-frequency representation obtained after first layer of CVAE perform same as Mel filterbank features in ASR, preserving all information such as formant contours, voiced and unvoiced sounds, even when filters are learnt with a fully unsupervised objective. The modulation filtered representations (2nd layer output) as ASR features provides considerable noise robustness over the acoustic filterbank layer output.

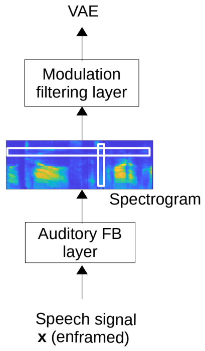

Figure 3: *Block diagram of CVAE architecture to learn auditory filters in Auditory FB layer, and modulation filters in Modulation filtering layer [12].*

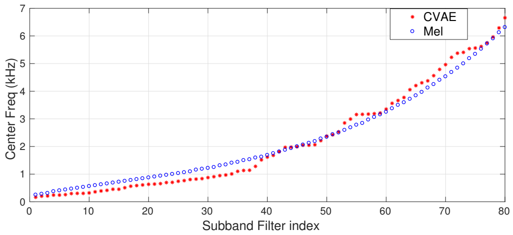

Figure 4: *Comparison of center frequency of filterbank learnt using CVAE with center frequencies of Mel filterbank.*

3. Future Work

The ongoing work is fine-tuning the learnt filters for the task of speech recognition. Since the acoustic filters and modulation filters are learnt in an unsupervised manner, fine-tuning them (updating) for the task in hand may further enhance the ASR system performance, and can make the representation task-specific. Hence, the transfer learning approach with using learnt filters as initialization point in ASR may benefit.

In addition, since these filter learning approaches are unsupervised, the learnt representations may be capturing other speech characteristics as well, like speaker information, accent, or language information. Hence, analysis of the obtained representation for other tasks may prove to be crucial.

4. Acknowledgement

I would like to acknowledge my Ph.D. supervisor Dr. Sriram Ganapathy, Dept. of Electrical Engg. Indian Institute of Science, Bengaluru, for the valuable discussions and motivation for the work. Also I would like to acknowledge the funding source Ministry of Human Resource and Development (MHRD), Government of India, for Postgraduate Research Scholarship, 2015-2020, and GARP fund for travel assistance.

5. References

[1] M. L. Seltzer, D. Yu, and Y. 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.

[2] T. M. Elliott and F. E. Theunissen, “The modulation transfer function for speech intelligibility,” *PLoS computational biology*, vol. 5, no. 3, p. e1000302, 2009.

[3] 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.

[4] P. Agrawal and Ganapathy, “Speech representation learning using unsupervised data-driven modulation filtering for robust ASR,” pp. 2446–2450, *Proc. INTERSPEECH*, 2017.

[5] P. Agrawal and S. Ganapathy, “Comparison of unsupervised modulation filter learning methods for ASR.” in *Proc. of INTERSPEECH*, 2018, pp. 2908–2912.

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

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

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

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[11] P. Agrawal and S. Ganapathy, “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.

[12] P. Agrawal and Ganapathy, “Deep variational filter learning models for speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5731–5735.

[13] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[14] P. Agrawal and S. Ganapathy, “Unsupervised raw waveform representation learning for ASR,” in *(to appear) in Proc. INTERSPEECH*, 2019.