Auditory Representation Learning

Hardik B. Sailor

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

sailor_hardik@daiict.ac.in

1. Motivation

Representation of a speech and audio signal based on human auditory processing is of significant interest in developing features for speech and audio processing applications [1]. The traditional approaches to extract the auditory features are either based on computational or mathematical models of auditory system [2–5]. Recently, representation learning (RL) has gained a significant interest for feature learning in various signal processing areas including speech processing [6]. In the speech processing literature, it is also called Automatic Speech Analysis (ASA) that involves the extraction of meaningful information using computers from the speech signals [7]. There are various approaches proposed for auditory modeling using RL techniques [8–18]. Unsupervised learning is one of the important forms of representation learning since many human learning tasks are unsupervised [19], [20].

A motivation for this thesis is the first notable study conducted by M. Lewicki to show that the human auditory system (HAS) is adapted to the sound statistics [8]. The experimental results in [8] showed that the optimal auditory codes are different according to the statistics of sounds. The objective of this thesis is to propose a novel auditory representation learning model that can be used in speech and audio processing applications. In this thesis, we have developed unsupervised auditory filterbank learning using a Convolutional Restricted Boltzmann Machine (ConvRBM). Our proposed ConvRBM model has been successfully applied in automatic speech recognition (ASR) [21–25], Environmental Sound Classification (ESC) [26], spoof speech detection (SSD) [27], [28], and infant cry classification (ICC) [29], [30]. Next, we discuss the key research challenges in the development of auditory models.

2. Key Research Challenges

The Human Auditory System (HAS) is one of the engineering masterpieces of the human body that is unique and distinct from other animals. The key research challenges in developing the human auditory model are as follows:

1. The HAS is highly complex containing several layers of nonlinear transformations and physiological effects, many of which are still not clearly understood [31].
2. There is a significant importance of the temporal structures in the sounds. Most of the auditory models in speech and audio applications have used windowing for the quasi-stationary assumptions that introduce artifacts [32]. Hence, how to preserve the temporal structures in the sounds is another open research issue.
3. The standard auditory representations use a fixed auditory frequency scale and filter shapes for a variety of applications. However, the auditory system is continuously adapting to the natural sound statistics [8].
4. Understanding the cortical representation of the sounds is currently an active research area in auditory neuroscience. Due to the highly complex nature of the auditory cortex, many computational and RL models are still at an elementary-level.

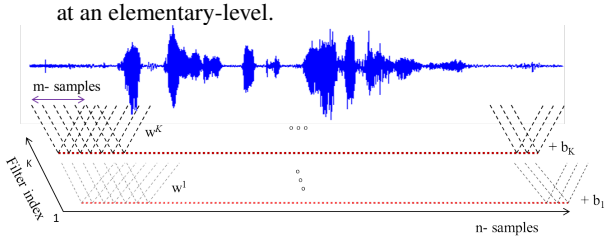

Figure 1: *Architecture of proposed model of auditory filterbank learning using ConvRBM. After [22].*

3. Contributions From the Thesis

The main contribution of the thesis is to propose a novel model of the auditory representation learning that tries to address a few of the research challenges mentioned above. The model is based on ConvRBM, an unsupervised probabilistic graphical model (PGM). Following are the key contributions in this thesis using our proposed ConvRBM model:

3.1. Proposed Model for Auditory Representation Learning

Compared to the earlier work of using ConvRBM to model the spectrograms with sigmoid units [34], we proposed to model the raw speech signals of arbitrary lengths and thus, avoiding the need of windowing. We also propose to use noisy rectified linear units (NReLU) (instead of sigmoid units in [34]) for inference in ConvRBM. We further improved our proposed ConvRBM using an annealing dropout and the Adam optimization. For noise-robust ASR task, a novel auditory-based feature representation is proposed using ConvRBM and the energy estimation using the Teager Energy Operator (TEO). An unsupervised deep auditory model (UDAM) is proposed by stacking the two ConvRBMs using a greedy layer-wise training [24]. Hence, the proposed UDAM can be seen as a simplified model of the deep auditory processing in humans.

3.2. Analysis of the Model and Representation

The subband filters, frequency scale, and the hidden unit representations of the ConvRBM are analyzed in this thesis. The comparative analysis of the subband filters and the frequency scales obtained using various sound categories are also provided (also shown in Figure 3). Specifically, frequency scales obtained via ConvRBM training on three standard ASR database are shown in Figure 2. The cross-domain experiments are performed on the ASR task to justify that ConvRBM can learn general representation across various databases of the speech signals. The mathematical justification of improved performance

Table 1: *Summary of experimental results (obtained through different types of DNN back-ends) for various standard ASR databases.*

Database	Proposed Representation	Relative Imp.
TIMIT	ConvRBM-BANK	2.56 % PER [21], [22]
WSJ0, WSJ	ConvRBM-BANK	1.35-6.82 % WER [21],, [22]
AURORA 4	ConvRBM-BANK	1.25-3.85 % WER [22]
AURORA 4	TEO-ConvRBM-BANK	1.26-11.63 % WER [25]
Gujarati Agricultural ASR System	TEO-ConvRBM-BANK	5.4 % WER [33]

PER=Phone Error Rate, WER = Word Error Rate

Table 2: *Summary of experimental results in audio classification applications*

Application	Database	Proposed Representation	Absolute Improvement
ESC	ESC-50	ConvRBM-BANK	10.65-18.70 % Acc [26]
Synthesic SSD	ASVSpoof 2015	ConvRBM-CC	4.76 % EER [27]
Replay SSD	ASVSpoof 2017	AM-FM part of ConvRBM-CC	5.28-7.48 % EER [28]
Infant Cry Classification	Baby Chillanto, DA-IICT	ConvRBM-CC	0.58-2 % Acc [29]

EER = Equal Error Rate, Acc=Classification Accuracy

in the noise-robust ASR task is given using the Lipschitz continuity conditions derived for the ConvRBM [24].

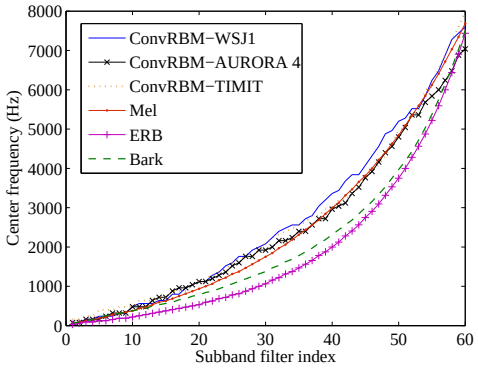

Figure 2: *Comparison of filterbank learned using ConvRBM with auditory filterbanks. After [22].*

3.3. Applications

The first motivation to develop the ConvRBM is to use it as a front-end in the ASR task. As a part of the MeitY, Govt. of India sponsored consortium project at DA-IICT, ConvRBM is also applied in the development of a speech-based access system for the agricultural commodities in the Gujarati language. Later, the ConvRBM is applied in a variety of speech and audio processing applications, namely, Environmental Sound Classification (ESC), Spoof Speech Detection (SSD), and Infant Cry Classification (ICC). In all these applications, our proposed model gave consistently better performance compared to the respective baselines. The overall contributions of this thesis in various applications are summarized in Figure 3.

4. Results and Discussions

The summary of experimental results on various standard ASR databases using is shown in Table 1. In TIMIT, WSJ, and AURORA 4 task, ConvRBM filterbank (denoted as ConvRBM-BANK) is used as a front-end and DNN is used as a back-end. Noise robustness for AURORA 4 and agricultural ASR task was further improved using TEO applied on ConvRBM-BANK. The summary of experimental results on the various audio classification task is shown in Table 2. We achieved state-of-the-art results on the ESC-50 classification task [26] as described in the literature [35] (also referred in [36]). Except for the ESC task, we used ConvRBM Cepstral Coefficients (ConvRBM-CC) obtained by applying Discrete Cosine Transform (DCT) on ConvRBM-BANK on synthetic and replay SSD task and infant cry classification.

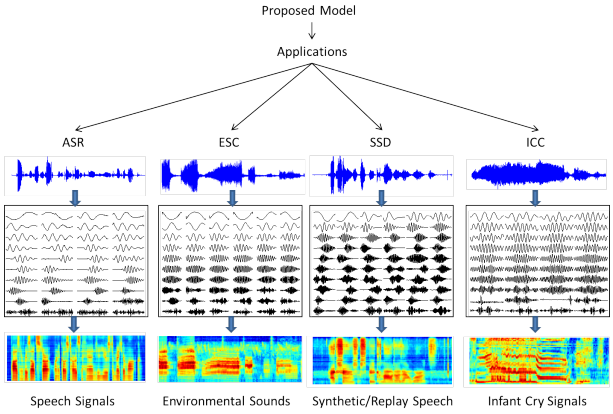

Figure 3: *Proposed model applied in different applications along with the subband filters for particular sound categories.*

5. Summary and Conclusions

In this thesis work, a novel auditory representation learning framework using ConvRBM is presented. The significance of our proposed model is to learn subband filters directly from the raw speech signals of arbitrary lengths. The model is applied in several ASR and audio classification tasks. The learned subband filters are adapted with respect to specific sound categories. Our proposed approach of model development and analysis could be a good starting point for those who would like to advance the research in auditory representation learning.

6. Acknowledgments

I am very thankful to my Ph.D. supervisor Prof. Hemant A. Patil for his excellent guidance and motivation towards research publications. I would also like to thank the MeitY, Govt. of India, for two sponsored projects (1) TTS Phase-II, (2) ASR Phase-II and the authorities of DA-IICT. I also acknowledge NVIDIA for the hardware grant of TITAN X GPU to the Speech Research Lab, DA-IICT.

7. References

[1] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 34–43, Nov 2012.

[2] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America (JASA)*, vol. 118, no. 2, pp. 887–906, 2005.

[3] R. M. Stern and N. Morgan, "Features based on auditory physiology and perception," in *Techniques for Noise Robustness in Automatic Speech Recognition*. T. Virtanen, B. Raj, and R. Singh, (Eds.) John Wiley and Sons, Ltd, New York, NY, USA, 2012, pp. 193–227.

[4] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *The Journal of the Acoustical Society of America (JASA)*, vol. 124, no. 1, pp. 422–438, 2008.

[5] J. Anden and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.

[6] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[7] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, July 2017.

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

[9] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[10] A. Bertrand, K. Demuynck, V. Stouten, and H. V. hamme, "Unsupervised learning of auditory filter banks using non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP) 2008, Las Vegas, Nevada, USA*, 2008, pp. 4713–4716.

[11] Y.-H. Chiu, B. Raj, and R. Stern, "Learning-based auditory encoding for robust speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Dallas, Texas, USA*, March 2010, pp. 4278–4281.

[12] 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), Prague, Czech Republic*, 22–27 May 2011, pp. 5884–5887.

[13] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, Dresden, Germany, 6–10 Sept. 2015, pp. 1–5.

[14] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *INTERSPEECH*, Dresden, Germany, 6-10 Sep. 2015, pp. 26–30.

[15] D. Palaz, M. Magimai-Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *40^{th} International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD*, 19–24 April 2015, pp. 4295–4299.

[16] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *INTERSPEECH*, Singapore, 14–18 Sep. 2014, pp. 890–894.

[17] Y. Tokozume and T. Harada, "Learning environmental sound with end-to-end convolutional neural network," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, New Orleans, USA, 2017, pp. 2721–2725.

[18] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain, 2016, pp. 892–900.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[20] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 831–843, Nov. 2004.

[21] H. B. Sailor and H. A. Patil, "Filterbank learning using convolutional restricted Boltzmann machine for speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 20-25 March 2016, pp. 5895–5899.

[22] H. B. Sailor and H. A. Patil, "Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 12, pp. 2341–2353, Dec. 2016.

[23] 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), Budapest, Hungary*, 29 Aug. - 2 Sept. 2016, pp. 873–877.

[24] H. B. Sailor and H. A. Patil, "Unsupervised deep auditory model using stack of convolutional RBMs for speech recognition," in *INTERSPEECH*, San Francisco, California, USA, 8–12 September 2016, pp. 3379–3383.

[25] H. B. Sailor and H. A. Patil, "Auditory feature representation using convolutional restricted Boltzmann machine and Teager energy operator for speech recognition," *Journal of Acoustical Society of America Express Letters (JASA-EL)*, vol. 141, no. 6, pp. EL500–EL506, June. 2017.

[26] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3107–3111.

[27] H. B. Sailor, M. R. Kamble, and H. A. Patil, "Unsupervised representation learning using convolutional restricted Boltzmann machine for spoof speech detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2601–2605.

[28] H. B. Sailor, M. R. Kamble, and H. A. Patil, "Auditory filterbank learning for temporal modulation features in replay spoof speech detection," *to appear in INTERSPEECH, Hyderabad*, Sept. 2018.

[29] H. B. Sailor and H. A. Patil, "Auditory filterbank learning using ConvRBM for infant cry classification," *to appear in INTERSPEECH, Hyderabad*, Sept. 2018.

[30] H. B. Sailor and H. A. Patil, "Unsupervised auditory filterbank learning for infant cry classification," in *submitted in Voice Technologies for Reconstruction and Enhancement*. Hemant A. Patil and Neustein, Amy,(Eds.), De Gruyter Series in Speech Technology and Text Analytics in Medicine and Healthcare, 2018, pp. 1–18.

[31] J. W. Schnupp, I. Nelken, and A. J. King, *Auditory Neuroscience: Making Sense of Sound*. The MIT Press, First Edition, 2012.

[32] E. Smith and M. S. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Comput.*, vol. 17, no. 1, pp. 19–45, Jan. 2005.

[33] H. B. Sailor and H. A. Patil, "Representation learning for speech recognition system in agricultural commodity for Gujarati (poster presentation)," in *Global Conference on Cyberspace (GCCS), Organized by MeitY, Govt. of India under National e-Governance Division (NeGD), New Delhi, India*, 2017.

[34] H. Lee, P. T. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *23^{rd} Annual Conference on Neural Information Processing Systems (NIPS), Canada, 7-10 December*, 2009, pp. 1096–1104.

[35] K. J. Piczak, "ESC-50: Dataset for environmental sound classification," URL: https://github.com/karoldvl/ESC-50, {Last Accessed: 10 July 2018}.

[36] B. Korbar, D. Tran, and L. Torresani, "Co-training of audio and video representations from self-supervised temporal synchronization," *arXiv preprint arXiv:1807.00230*, 2018.