Interspeech 2018
2-6 September 2018, Hyderabad

Auditory Filterbank Learning Using ConvRBM for Infant Cry Classification

Hardik B. Sailor, Hemant A. Patil

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

{sailor_hardik,hemant_patil}@daiict.ac.in

Abstract

The infant cry classification is a socially-relevant problem where the task is to classify the normal *vs*. pathological cry signals. Since the cry signals are very different from the speech signals in terms of temporal and spectral content, there is a need for better feature representation for infant cry signals. In this paper, we propose to use unsupervised auditory filterbank learning using Convolutional Restricted Boltzmann Machine (ConvRBM). Analysis of the subband filters shows that most of the subband filters are Fourier-like basis functions. The infant cry classification experiments were performed on the two databases, namely, DA-IICT Cry and Baby Chillanto. The experimental results show that the proposed features perform better than the standard Mel Frequency Cepstral Coefficients (MFCC) using various statistically meaningful performance measures. In particular, our proposed ConvRBM-based features obtained an absolute improvement of 2 % and 0.58 % in the classification accuracy on the DA-IICT Cry and the Baby Chillanto database, respectively.

Index Terms: Infant cry classification, auditory filterbank learning, ConvRBM.

1. Introduction

Humans cry to express a range and degree of emotions, such as from happiness after passing a tough exam or meeting a beloved one to grief after the death of a person or difficult situations in life [1]. On the whole, the crying is not just a simple reaction to any feeling or emotional state but rather a multifaceted behavior that can offer clues to how we process and regulate our feelings, and how we experience the world around us [1]. In humans, infants communicate their need, such as feeding, distress or pain by crying [2]. Intra-individual variation in the infant cry sounds is known to encode information on the condition, emotional status, needs and the degree of urgency. Based on the perception of the cry, parents or caretakers empirically try to understand the reason for the crying and even identify their newborn [2]. The infant cry classification is very helpful to the parents, caretakers, and pediatricians in the diagnosis of a pathology at an early stage. This may be beneficial to reduce or completely eliminate symptoms of a pathology. Many times, addressing pathology at an early stage of infant development leads to severe conditions including the death.

The research work in this direction is also important to identify the appropriate reasons for sudden infant death syndrome (SIDS) [3], [4]. The first study of SIDS case was analyzed in [5]. The SIDS is sudden unexplained death of a child less than one year of age that remains unexplained even after a complete forensic investigation [6]. The early studies suggest that infants who die from SIDS are born with brain abnormalities (in the specific brain region called *medulla oblongata* that helps in control functions, such as breathing, and blood pressure [7]) or physiological defects [6]. Hence, the study of infant cry analysis will be very much helpful to prevent SIDS cases. Another important social relevance is in the families where literacy level is lower and access to good hospitals is difficult specifically in remote villages. The infant cry classification if implemented in mobiles devices (since mobiles finds its usage in wide strata of society) can be beneficial in initial detection of the pathology in the infants through cry signals.

The are many signal processing challenges to analyze the infant cry signals, such as difficulties in extraction of source excitation and vocal-tract related features, scarcity of pathological cry samples, and unbalanced data for classification. From a signal processing perspective, our goal is to classify whether the infant is crying due to pain, hunger or some medical disease collectively called as *pathology*. To date, there is no standard publicly available database for infant cry classification. Many researchers collected their own data including our Speech Research Group at DA-IICT [8], [9]. Few studies used Baby Chillanto infant cry database [10] while most of the recent studies used their own corpus. The early work includes analysis of different features for infant cry signals [11]. The classification of full-term and preterm infant cry is presented in [12]. Recently, an automatic cry segmentation system is proposed as a pre-processing step in the infant cry classification task [13]. Other notable studies for infant cry classification include the works reported in [14–21]. The detailed discussion on the topic of infant cry classification is found in [7] and in [22], the first Ph.D. thesis from India in this area.

Most of the earlier studies used Mel Frequency Cepstral Coefficients (MFCC) as auditory-based features (to the best of authors' knowledge). Various handcrafted features are explored in [22] for classification of infant cries. Recently, representation learning (RL) is very popular to learn meaningful feature representation directly from the raw audio signals [23]. Various approaches were proposed for RL that shows significant improvements compared to handcrafted features, such as MFCC [24]. The objective of this paper is to use our proposed Convolutional Restricted Boltzmann Machine (ConvRBM) for auditory-like filterbank learning from the raw audio signals [24], [25]. We used two databases, namely, (1) Baby Chillanto [10] and (2) DA-IICT Cry database collected by our group [8]. The experimental results showed an improved performance with the proposed feature representation compared to the baseline MFCC feature set.

2. Auditory Filterbank Learning using ConvRBM

ConvRBM is an undirected probabilistic graphical model used for representation learning. It has two layers, namely, a visible layer and a hidden layer [24]. The input to a visible layer (denoted as $\mathbf{x}$) is a cry signal of length n-samples. The hid-

10.21437/Interspeech.2018-1536

den layer (denoted as $\mathbf{h}$) is divided into K-groups (i.e., number of subband filters). Each group weight has filter length of m-samples. The weights of ConvRBM are shared between visible and hidden units in each group [24]. The hidden bias is denoted as b_k for the k^{th} group. The convolutional response for the k^{th} group is given as:

$$\mathbf{I}_k = (\mathbf{x} * \tilde{\mathbf{W}}^k) + b_k, \tag{1}$$

where $\mathbf{x} = [x_1, x_2, ..., x_n]$ are samples of the infant cry signal, $\mathbf{W}^k = [w_1^k, w_2^k, ..., w_m^k]$ is a weight vector (i.e., k^{th} subband filter) and $\tilde{\mathbf{W}}$ indicates a *flipped* array [24]. Here, we used the noisy leaky rectifier linear units (NLReLU) for the inference compared to our earlier works [24]. Earlier, LReLU is proposed to avoid the limitations of ReLU, such as zero gradient for negative inputs and unbounded output for very large inputs [26]. With an NLReLU, the sampling equations for hidden and visible units are given as:

$$\begin{aligned} \mathbf{h}^k &\sim max(0, \mathbf{J}_k) + \alpha \cdot min(0, \mathbf{J}_k), \\ \mathbf{x}_{recon} &\sim \mathcal{N}\left(\sum_{k=1}^{K}(\mathbf{h}^k * \mathbf{W}^k) + c, 1\right), \end{aligned} \tag{2}$$

where $\mathbf{J}_k = \mathbf{I}_k + N(0, \sigma(\mathbf{I}_k))$ with $N(0, \sigma(\mathbf{I}_k))$ is a Gaussian noise with mean zero and sigmoid of $\mathbf{I}_k$ as a variance. α is a leaky parameter which is generally set to 0.01 as suggested in [26] and c is a visible bias which is also shared among the visible units. In ConvRBM training, an annealed dropout is applied before sampling the hidden units in both the positive and negative phase of contrastive divergence (CD) learning. Our earlier works showed that the use of annealed dropout (first proposed in [27]) resulted in an improved performance in speech recognition [28] and audio classification [29]. In an annealed dropout training, the dropout probability of the hidden units in ConvRBM is gradually decreased over the training period. The parameters of ConvRBM are updated using an Adam optimization method [30].

After ConvRBM is trained, the pooling is applied in the temporal-domain to reduce the representation of ConvRBM filter responses. The pooling operation reduces the temporal resolution from $K \times n$ samples to the $K \times F$ frames. Logarithmic nonlinearity compresses the dynamic range of features. The block diagram for feature extraction procedure is shown in Figure 1. During feature extraction stage, we have used an absolute nonlinearity $|\mathbf{I}_k|$ as an activation function of the hidden units.

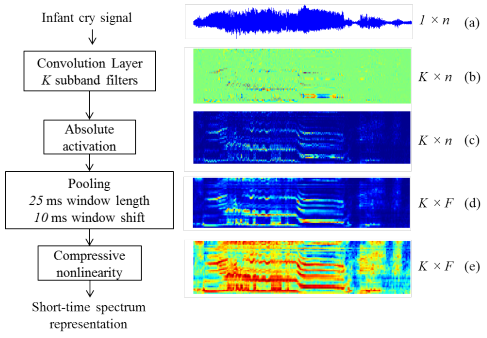


Figure 1: *Feature extraction using ConvRBM (after training): (a) Infant cry signal, (b) and (c) responses from the convolutional layer and absolute nonlinearity, respectively, (d) pooling, and (e) logarithmic compression.*

3. Analysis of Infant Cry Signals

3.1. Analysis of Subband Filters and Frequency Scale

The subband filters learned from the DA-IICT Cry and Baby Chillanto database are shown in Figure 2. We have also shown the subband filters obtained from the TIMIT speech database. It is important to note an intriguing observation that these subband filters were learned from only 37 minutes and 50 seconds duration of cry signals from the Baby Chilanto and 30 minutes of cry signals from the DA-IICT Cry database (such scarcity of larger database is all the more case in the medical scenarios [31]). Thus, it shows the applicability of our proposed model even in the very small database scenarios. The time-domain subband filters are significantly different than the one for normal TIMIT speech database. The subband filters of the infant cry databases contain more Fourier-like basis functions. The analysis of the frequency-domain subband filters revealed that many subband filters are not localized and contain *harmonic* structures. This may be due to the harmonic nature of the infant cry signals revealed via ten distinct cry modes in narrowband spectrograms [7], [15]. These results also justify the observations in [32] for animal vocalizations that are harmonic in nature. On comparing the subband filters learned from the two different databases, the subband filters from the baby Chillanto database has more lower frequency filters. However, most of the subband filters are similar in shape.

The frequency scales obtained using ConvRBM are compared with the standard auditory frequency scales in Figure 3. Unlike the frequency scale obtained through the speech database [24], here we observed two linear segments in the frequency scale, from 0 to 1 kHz and from 1 kHz to 3 kHz. After 3 kHz, it is nonlinear and follows the ERB, and Bark scales. However, the frequency scale from the DA-IICT Cry database is more away from the standard scales. The difference in the frequency scales of both the databases is may be due to variabilities in the cry signal production mechanism through language perception (Indian languages *vs.* English in the Baby Chillanto), data recording conditions, background noise, channel characteristics, microphone specifications, etc.

4. Experimental Setup

4.1. Databases

The experiments were performed with two databases described as follows:

4.1.1. DA-IICT Cry Database

The DA-IICT Cry database was collected as a part of the DST fast-track scheme for young scientist project, "Development of Infant Cry Analyzer using Source and System Features" [8]. The sampling frequency of the original recordings was 12 kHz, quantized at 16-bit PCM. For our experiments, we resample it to 11.025 kHz since at a later stage we will compare the results with another database. The statistics of DA-IICT Cry database is shown in Table 1. The healthy cry signals consist of normal and hunger cry signals. The pathological cry includes two types of pathology, namely, asphyxia (also called as Hypoxic Ischemic Encephalopathy (HIE)) and asthma.

4.1.2. Baby Chillanto

Baby Chillanto infant cry database was developed by the recordings conducted by medical doctors which is a property of INAOE-CONACyT, Mexico [10]. Each cry signal was seg-

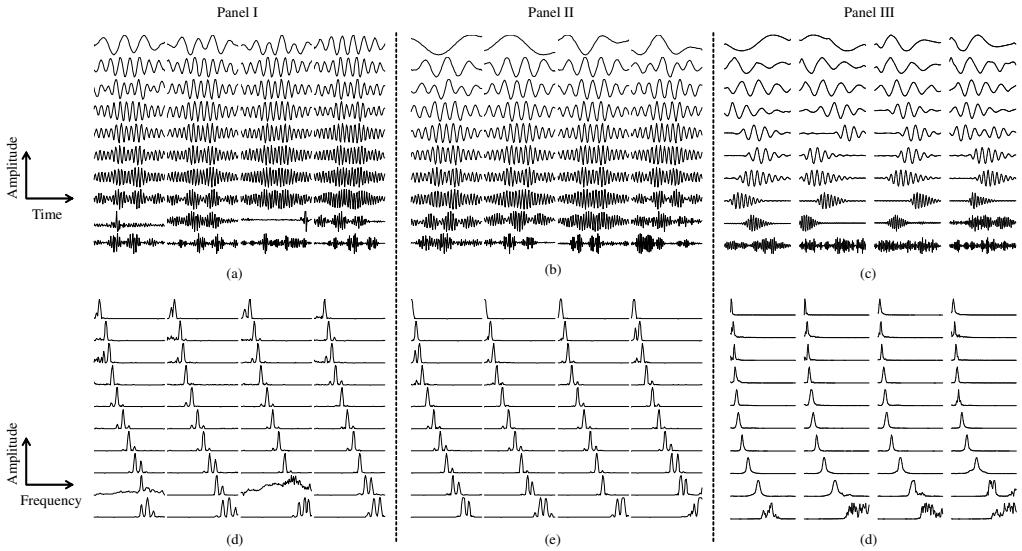


Figure 2: *The subband filters trained on DA-IICT Cry (Panel I), Baby Chillanto (Panel II), and TIMIT (Panel III) databases, respectively: (a)-(c) in the time-domain, (d)-(f) corresponding frequency responses.*

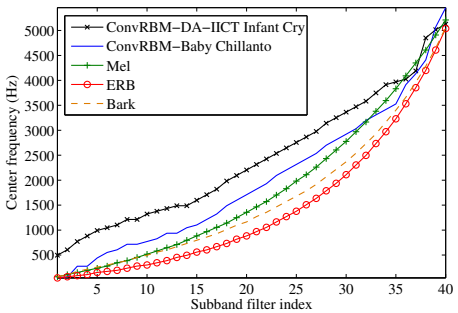


Figure 3: *Analysis of the auditory frequency scales.*

mented into one-second duration (which represent one sample) and are grouped into five categories as shown in Table 1. Since the sampling rate of cry signals is different in all the categories, we kept the sampling rate of 11.025 kHz for all the categories. Two groups were formed for binary classification of healthy *vs.* pathology. Healthy cry signals include three categories, namely, normal, hungry, and pain resulting in 1049 cry samples. Pathology cry signals include two categories, namely, asphyxia, and deaf resulting in 1219 cry samples.

Table 1: *Number of samples contained in DA-IICT Cry database (D1) and Baby Chillanto (D2)*

Class	Category	Samples in D1	Samples in D2
Healthy	Normal	793*	507
	Hungry	-	350
	Pain	-	192
Pathology	Asphyxia	215	340
	Asthma	182	-
	Deaf	-	879

*Samples include both the normal and hunger cry

4.2. Training of ConvRBM and Feature Extraction

The ConvRBM is trained using an annealed dropout with dropout probability p=0.3 that decayed to zero (i.e., $p = 0$) during the training. The learning rate was chosen to be 0.001 and decayed according to the learning rate schedule as suggested in [30]. The moment parameters of Adam optimization chosen to be β_1=0.9 and β_2=0.999. The model is trained using 40 subband filters (i.e., K) with window length m= 88 samples (i.e., 8 ms). After the ConvRBM was trained, the features were extracted from the cry signals. The Discrete Cosine Transform (DCT) was applied to reduce the dimension retaining only first 13 dimensional (D) coefficients. The delta and double-delta features were also appended resulting in 39-D cepstral features (denoted as ConvRBM-CC). The baseline MFCC features are extracted from the cry signals with 25 ms window length and 10 ms window shift.

4.3. Binary Classification and Evaluation

Since both the cry databases are very small in size, the Gaussian Mixture Models (GMM) is used for the binary classification. Healthy cry features belong to one class and pathology cry features belong to another class. The GMMs with different mixture components were trained using the MFCC and ConvRBM-CC. The decision of the test cry signal being healthy or pathology is based on the log-likelihood ratio (LLR). The results are predicted using LLR scores with 10-fold cross-validation (CV). Since the number of samples in the two classes are different, we applied 10-fold CV separately for each class and then combine respective test folds. For each fold, we noted % classification accuracy. The final result is presented as averaged % classification accuracy over 10 CV folds. The performance of the classification task is evaluated using F-measure (also called as F1-ratio), Youden's J-statistic or informedness [33], and Matthews Correlation Coefficient (MCC) [34] obtained from the confusion matrix [35]. The range of F-measure is [0,1] while for J-statistic and MCC, the range is [-1,1] (higher is better for all the measures). MCC is considered as a balanced measure which can be used even if the classes are of very different sizes.

5. Experimental Results

In this Section, the classification results and evaluation using various performance measures are presented.

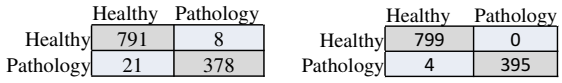
5.1. Results on the DA-IICT Cry Database

The classification accuracy for the DA-IICT Cry database using MFCC and ConvRBM-CC feature sets are shown in Table 2. ConvRBM-CC obtained higher % classification accuracy compared to MFCC for all the GMM components. For MFCC, the optimal results obtained using 200 GMM components. For the ConvRBM-CC, the optimal results obtained using 400 GMM components. We achieved an absolute improvement of 2 % in the classification accuracy compared to the MFCC feature set. The confusion matrices for the classification experiment are shown in Figure 4. The false positive (FP) and false negative (FN) rate of the MFCC are quite high compared to the ConvRBM-CC feature set. From Figure 4 (b), it can be seen that the ConvRBM-CC has no FP and only 4 FN compared to 21 FN using MFCC (Figure 4 (a)). Hence, with ConvRBM-CC, there is no chance that the normal cry signal is considered as pathological cry signal.

The performance measures of the classification experiments on the DA-IICT Cry database are shown in Table 3. The ConvRBM-CC gave a significantly better performance for all the measures than MFCC. Since F-measure do not consider the true negatives (TN) into account, the values of F-measure are closer for both the feature sets. The MCC and J-measure values are higher for ConvRBM-CC compared to the MFCC. From Table 3, it can be observed that the difference in MCC and J-statistic for MFCC and ConvRBM-CC is higher compared to % accuracy. This is due to the fact that % accuracy does not consider FP and FN in the confusion matrix. Hence, MCC and J-statistic are more meaningful performance measure than % classification accuracy alone.

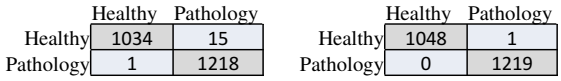
5.2. Results on the Baby Chillanto Database

The experimental results using the Baby Chillanto database is shown in Table 2 for MFCC and ConvRBM-CC with different GMM mixture components. Compared to the Cry database, both the features were able to perform well in the classification of normal and pathology cry signals. However, ConvRBM-CC consistently performs better than the MFCC for all GMM components. The best classification accuracy of 99.87 % was achieved using ConvRBM-CC (0.58 % absolute improvement compared to the MFCC) obtained with 300 GMM mixture components. The confusion matrices for both the feature sets are shown in Figure 5. The false positive rate of the MFCC is quite high than ConvRBM-CC (15 *vs.* 1), while there are no false negative when ConvRBM-CC is used in the classification task. Hence, with the ConvRBM-CC feature set, all the cry samples are correctly classified with only one false negative. The significance of this improvement using ConvRBM-CC feature set can also be seen from the performance measures in Table 3. The F-measure is similar for both the ConvRBM-CC and MFCC. The MCC and J-statistic are higher for the ConvRBM-CC with a value 0.999 (close to 1). The difference in values of MCC and J-statistic indicates that ConvRBM-CC performs better than the MFCC evenif % accuracy is similar.



(a) (b)

Figure 4: *Confusion matrices for experiments on the DA-IICT Cry database using: (a) MFCC, and (b) ConvRBM-CC.*



(a) (b)

Figure 5: *Confusion matrices for experiments on the Baby Chillanto database using: (a) MFCC, and (b) ConvRBM-CC.*

Table 2: *The % classification accuracy using the DA-IICT Cry database (D1) and Baby Chillanto database (D2) for various GMM components (columns)*

Dataset	Feature Set	200	256	300	400	512
D1	MFCC	97.57	97.32	97.24	97.24	96.9
D1	ConvRBM-CC	99.58	99.58	99.58	**99.66**	99.57
D2	MFCC	99.12	99.25	99.16	99.29	98.94
D2	ConvRBM-CC	99.82	99.82	**99.87**	99.91	99.96

Table 3: *Performance measures for the classification experiments using the DA-IICT Cry database (D1) and Baby Chillanto database (D2)*

Dataset	Feature Set	MCC	F-measure	J-statistic
D1	MFCC	0.945	0.963	0.937
D1	ConvRBM-CC	**0.993**	**0.995**	**0.99**
D2	MFCC	0.986	0.994	0.985
D2	ConvRBM-CC	**0.999**	**0.999**	**0.999**

6. Summary and Conclusions

In this study, we proposed to use ConvRBM-based auditory filterbank learning for the infant cry classification task. The subband filters learned from the two infant cry databases shows that most of the learned subband filters are the Fourier-like basis functions. The filterbank scale is also different than the standard auditory frequency scales since the ConvRBM is adapted to represent the cry signals. The classification experiments for the healthy *vs.* pathological cry signals are presented. The experimental results using standard performance measures show that the proposed ConvRBM-based features perform significantly well for the infant cry classification task. Our future work includes developing an infant cry classifier in a mobile application framework that will be helpful to the doctors and society as every infant may not have luxuary of access to the pediatricians.

7. Acknowledgments

Authors would like to thank the Ministry of Electronics and Information Technology (MeitY), Govt. of India, Department of Science and Technology (DST), Govt. of India, and the authorities of DA-IICT. We would also like to thank INAOE-CONACyT, Mexico for providing the Baby Chillanto database.

8. References

[1] O. Aragn, "Why do we cry?" *Scientific American Mind*, vol. 28, no. 2, p. 74, April 2017.

[2] E. Gustafsson, F. Levrro, D. Reby, and N. Mathevon, "Fathers are just as good as mothers at recognizing the cries of their baby," *Nature Communications*, vol. 4, no. 1698, pp. 1–6, 2013.

[3] M. J. Corwin, B. M. Lester, C. Sepkoski, M. Peucker, H. Kayne, and H. L. Golub, "Newborn acoustic cry characteristics of infants subsequently dying of sudden infant death syndrome," *Pediatrics*, vol. 96, no. 1, pp. 73–77, 1995.

[4] M. P. Robb, D. H. Crowell, and P. Dunn-Rankin, "Sudden infant death syndrome: Cry characteristics," *International Journal of Pediatric Otorhinolaryngology, Elsevier*, vol. 77, no. 8, pp. 1263–1267, 2013.

[5] R. H. Colton and A. Steinschneider, "The cry characteristics of an infant who died of the sudden infant death syndrome," *Journal of Speech and Hearing Disorders*, vol. 46, no. 4, pp. 359–363, 1981.

[6] U. D. of Health and H. Services, "Sudden infant death syndrome (SIDS)," URL: https://www.nichd.nih.gov/health/topics/sids/Pages/default.aspx, {Last Accessed on 10 March, 2018}.

[7] H. A. Patil, "Cry baby: Using spectrographic analysis to assess neonatal health status from an infant's cry," in *Advances in Speech Recognition Mobile Environments, Call Centers and Clinics*, A. Neustein, Ed. Springer, 2010, pp. 323–348.

[8] N. Buddha and H. A. Patil, "Corpora for analysis of infant cry," in *Int. Conf. on Speech Databases and Assessments, Oriental COCOSDA, Hanoi, Vietnam*, Dec. 2007, pp. 43 – 48.

[9] A. Chittora and H. A. Patil, "Data collection of infant cries for research and analysis," *Journal of Voice, Elsevier*, vol. 31, no. 2, pp. 252.e15 – 252.e26, 2017.

[10] A. Rosales-Prez, C. A. Reyes-Garca, J. A. Gonzalez, O. F. Reyes-Galaviz, H. J. Escalante, and S. Orlandi, "Classifying infant cry patterns by the genetic selection of a fuzzy model," *Biomedical Signal Processing and Control, Elsevier*, vol. 17, no. 1, pp. 38 – 46, 2015.

[11] R. Prescott, "Infant cry sound: developmental features," *The Journal of the Acoustical Society of America (JASA)*, vol. 57, no. 5, pp. 1186–1191, 1975.

[12] S. Orlandi, C. A. R. Garcia, A. Bandini, G. Donzelli, and C. Manfredi, "Application of pattern recognition techniques to the classification of full-term and preterm infant cry," *Journal of Voice, Elsevier*, vol. 30, no. 6, pp. 656–663, 2016.

[13] L. Abou-Abbas, C. Tadj, and H. A. Fersaie, "A fully automated approach for baby cry signal segmentation and boundary detection of expiratory and inspiratory episodes," *The Journal of the Acoustical Society of America*, vol. 142, no. 3, pp. 1318–1331, 2017.

[14] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "Classification of infant cry vocalizations using artificial neural networks (ANNs)," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Detroit, Michigan USA*, May 1995, pp. 3475–3478.

[15] Q. Xie, R. Ward, and C. Laszlo, "Hidden Markov model method for estimating normal infants distress levels from their cry sounds," *IEEE Transctions on Speech and Audio Processing*, vol. 4, no. 4, pp. 253–265, July 1996.

[16] T. Etz, H. Reetz, C. Wegener, and F. Bahlmann, "Infant cry reliability: Acoustic homogeneity of spontaneous cries and pain-induced cries," *Speech Communication*, vol. 58, no. 1, pp. 91 – 100, 2014.

[17] H. F. Alaie, L. Abou-Abbas, and C. Tadj, "Cry-based infant pathology classification using GMMs," *Speech Communication*, vol. 77, no. 1, pp. 28 – 52, 2016.

[18] L. Abou-Abbas, C. Tadj, C. Gargour, and L. Montazeri, "Expiratory and inspiratory cries detection using different signals' decomposition techniques," *Journal of Voice, Elsevier*, vol. 31, no. 2, pp. 259.e13 – 259.e28, 2017.

[19] M. A. R. Daz, C. A. R. Garca, L. C. A. Robles, J. E. X. Altamirano, and A. V. Mendoza, "Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis," *Biomedical Signal Processing and Control, Elsevier*, vol. 7, no. 1, pp. 43 – 49, 2012.

[20] S. Ntalampiras, "Audio pattern recognition of baby crying sound events," *Journal of Audio Engineering Society (JAES)*, vol. 63, no. 5, pp. 358–369, 2015.

[21] R. Torres, D. Battaglino, and L. Lepauloux, "Baby cry sound detection: a comparison of hand crafted features and deep learning approach," in *International Conference on Engineering Applications of Neural Networks, Athens, Greece*. Springer, 2017, pp. 168–179.

[22] A. Chittrora, "Crying for a reason: A signal processing based approach for infant cry analysis and classification," *Ph.D. Thesis, Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India*, Jan. 2017.

[23] 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.

[24] 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.

[25] 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.

[26] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. International Conference on Machine Learning (ICML), Atlanta, USA*, vol. 30, 2013, pp. 1–6.

[27] S. J. Rennie, V. Goel, and S. Thomas, "Annealed dropout training of deep networks," in *IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, California and Nevada*, 2014, pp. 159–164.

[28] 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.

[29] 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.

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR), San Diego*, 2015, pp. 1–11.

[31] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone, and S. Narayanan, "Pathological speech processing: State-of-the-art, current challenges, and future directions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China*, March 2016, pp. 6470–6474.

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

[33] W. J. Youden, "Index for rating diagnostic tests," *Cancer, Wiley Subscription Services, Inc., A Wiley Company*, vol. 3, no. 1, pp. 32–35, 1950.

[34] B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure, Elsevier*, vol. 405, no. 2, pp. 442 – 451, 1975.

[35] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters, Elsevier*, vol. 27, no. 8, pp. 861 – 874, 2006.