Neural Transfer Learning for Cry-based Diagnosis of Perinatal Asphyxia

Charles C. Onu[1,2]*, Jonathan Lebensold*[1,2]*, William L. Hamilton*[1,3]*, Doina Precup*[1,4]

[1]Mila - Québec Artificial Intelligence Institute, McGill University
[2] Ubenwa Intelligence Solutions Inc
[3]Facebook AI Research
[4]Google DeepMind

{charles.onu@mail, jonathan.maloney-lebensold@mail, wlh@cs, dprecup@cs}.mcgill.ca

Abstract

Despite continuing medical advances, the rate of newborn morbidity and mortality globally remains high, with over 6 million casualties every year. The prediction of pathologies affecting newborns based on their cry is thus of significant clinical interest, as it would facilitate the development of accessible, low-cost diagnostic tools. However, the inadequacy of clinically annotated datasets of infant cries limits progress on this task. This study explores a neural transfer learning approach to developing accurate and robust models for identifying infants that have suffered from perinatal asphyxia. In particular, we explore the hypothesis that representations learned from adult speech could inform and improve performance of models developed on infant speech. Our experiments show that models based on such representation transfer are resilient to different types and degrees of noise, as well as to signal loss in time and frequency domains.

1. Introduction

Perinatal asphyxia—i.e., the inability of a newborn to breath spontaneously after birth—is responsible for one-third of newborn mortalities and disabilities worldwide [1]. The high cost and expertise required to use standard medical devices for blood gas analysis makes it extremely challenging to conduct early diagnosis in many parts of the world. In this work, we develop and analyze neural transfer models [2] for predicting perinatal asphyxia based on the infant cry. We ask the question of whether such models could be more accurate and robust than previous approaches that primarily focus on classical machine learning algorithms due to limited data.

Clinical research has shown that there exists a significant alteration in the crying patterns of newborns affected by asphyxia [3]. The unavailability of reasonably-sized clinically-annotated datasets limits progress in developing effective approaches for predicting asphyxia from cry. The Baby Chillanto Infant Cry database [4], based on 69 infants, remains the only known available database for this task. Previous work using this data has mainly focused on classical machine learning methods or very limited capacity feed-forward neural networks [4, 5].

We take advantage of freely available large datasets of adult speech to investigate a transfer learning approach to this problem using deep neural networks. In numerous domains (e.g., speech, vision, and text) transfer learning has led to substantial performance improvements by pre-training deep neural networks on some different but related task [6, 7, 8]. In our setting, we seek to transfer models trained on adult speech to improve performance on the relatively small Baby Chillanto Infant Cry dataset. Unlike newborns—whose cry is a direct response to stimuli—adults have voluntary control of their vocal organs and their speech patterns have been influenced, over time, by the environment. We nevertheless explore the hypothesis that there exists some underlying similarity in the mechanism of the vocal tract between adults and infants, and that model parameters learned from adult speech could serve as better initialization (than random) for training models on infant speech.

Of course, the choice of source task matters. The task on which the model is pre-trained should capture variations that are relevant to those in the target task. For instance, a model pre-trained on a speaker identification task would likely learn embeddings that identify individuals, whereas a word recognition model would likely discover an embedding space that characterizes the content of utterances. What kind of embedding space would transfer well to diagnosing perinatal asphyxia is not clear a priori. For this reason, we evaluate and compare 3 different (source) tasks on adult speech: speaker identification, gender classification and word recognition. We study how different source tasks affect the performance, robustness and nature of the learned representations for detecting perinatal asphyxia.

Key results. On the target task of predicting perinatal asphyxia, we find that a classical approach using support vector machines (SVM) represents a hard-to-beat baseline. Of the 3 neural transfer models, one (the word recognition task) surpassed the SVM's performance, achieving the highest unweighted average recall (UAR) of 86.5%. By observing the response of each model to different degrees and types of noise, and signal loss in time- and frequency-domain, we find that all neural models show better *robustness* than the SVM.

2. Related Work

Detecting pathologies from infant cry. The physiological interconnectedness of crying and respiration has been long appreciated. Crying presupposes functioning of the respiratory muscles [9]. In addition, cry generation and respiration are both coordinated by the same regions of the brain [10, 11]. The study of how pathologies affect infant crying dates back to the 1970s and 1980s with the work of Michelsson et al [3, 12, 13]. Using spectrographic analysis, it was found that the cries of asphyxiated newborns showed shorter duration, lower amplitude, increased higher fundamental frequency, and significant increase in "rising" melody type.

The Chillanto Infant Cry database. In 2004, Reyes et al. [4] collected the Chillanto Infant Cry database with the objective of applying statistical learning techniques in classifying deafness, asphyxia, pain and other conditions. The authors experimented with audio representations as linear predictive coefficients (LPC) and mel-frequency cepstral coefficients (MFCC), training a time delay neural network as the classifier. They achieved a precision and recall of 72.7% and 68%. Building on this work, Onu et al. [5] improved the precision and recall to

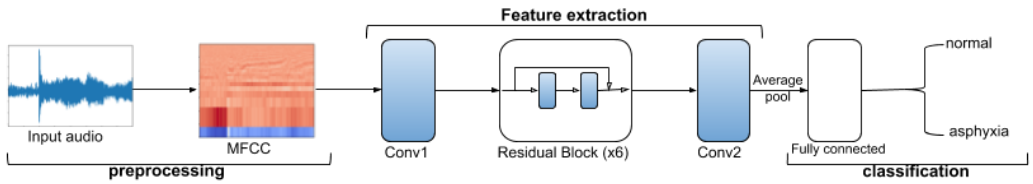

Figure 1: *Structure of learning pipeline. Weights of feature extraction stage were pre-loaded during transfer learning.*

73.4% and 85.3%, respectively, using support vector machines (SVM). It is worth noting that both works represent an overestimate of performance as authors split train/test set by examples, not by subjects.

Weight initialization and neural transfer learning. Modern neural networks often contain millions of parameters, leading to highly non-linear decision surfaces with many local optima. The careful initialization of the weights of these parameters has been a subject of continuous research, with the goal of increasing the probability of reaching a favorable optimum [14, 15]. Initialization-based transfer learning is based on the idea that instead of hand-designing a choice of random initialization, the weights from a neural network trained on similar data or task could offer better initialization. This pre-training could be done in an unsupervised [16] or supervised [17, 18] manner.

3. Methods

In this section, we describe our approach to designing and evaluating transfer learning models for the detection of perinatal asphyxia in infant cry. We present the source tasks selected along with representative datasets. We further describe preprocessing steps, choice of model architectures as well as analysis of trained models.

3.1. Tasks

3.1.1. Source tasks

We choose 3 source tasks — speaker identification, gender classification, word recognition — with corresponding audio datasets: VCTK [19], Speakers in the Wild (SITW) [20], and Speech Commands [21] . Table 1 briefly describes the datasets used for each task.

3.1.2. Target task: Perinatal asphyxia detection

Our target task is the detection of perinatal asphyxia from newborn cry. We develop and evaluate our models using the Chillanto Infant Cry Database. The database contains 1,049 recordings of normal infants and 340 cry recordings of infants clinically confirmed to have perinatal asphyxia. Audio recordings were 1-second long audio and sampled at frequencies between 8kHz to 16kHz with 16-bit PCM encoding.

Table 1: *Source tasks and corresponding datasets used in pre-training neural network. Size: number of audio files.*

Dataset	Description	Size
VCTK	Speaker Identification. 109 English speakers reading sentences from newspapers.	44K
SITW	Gender classification. Speech samples from media of 299 speakers.	2K
Speech commands	Word recognition. Utterances from 1,881 speakers of a set of 30 words.	65K

3.2. Pre-processing

All audio samples are pre-processed similarly, to allow for even comparison between source tasks and compatibility with target task. Raw audio recordings are downsampled to 8kHz and converted to mel-frequency cepstral coefficients (MFCC). To do this, spectrograms were computed for overlapping frame sizes of 30 ms with a 10 ms shift, and across 40 mel bands. For each frame, only frequency components between 20 and 4000 Hz are considered. The discrete cosine transform is then applied to the spectrogram output to compute the MFCCs. The resulting coefficients from each frame are stacked in time to form a spatial (40×101), 2D representation of the input audio.

3.3. Model Architecture and Transfer Learning

We adopt a residual network (ResNet) [22] architecture with average pooling, for training. Consider a convolutional layer that learns a mapping function $F(x)$ of the input, parameterized by some weights. A residual block adds a shortcut or skip connection such that the output of the layer is the sum of $F(x)$ and the input x, i.e., $y = F(x) + x$. This structure helps control overfitting by allowing the network to learn the identity mapping $y = x$ as necessary and facilitates the training of even deeper networks.

ResNets represent an effective architecture for speech, achieving several state-of-the-art results in recent years [23]. To assure even comparison across source tasks, and to facilitate transfer learning, we adopt a single network architecture: the *res8* as in Tang et al. [23]. The model takes as input a 2D MFCC of an audio signal, transforms it through a collection of 6 residual blocks (flanked on either side by a convolutional layer), employs average pooling to extract a fixed dimension embedding, and computes a k-way softmax to predict the classes of interest. Fig 1 shows the overall structure of our system. Each convolutional layer consists of 45, 3×3 kernels.

We train the *res8* on each source task to achieve performance comparable with the state of the art. The learned model weights (except those of the softmax layer) are used as initialization for training the network on the Chillanto dataset. During this post-training, the entire network is tuned.

3.4. Baselines

We implement and compare the performance of our transfer models with 2 baselines. One is a model based on a radial basis function Support Vector Machine (SVM), similar to [5]. The other is a *res8* model whose initial weights are drawn randomly from a uniform Glorot distribution [14] i.e., according to $U(-k, k)$ where $k = \frac{\sqrt{6}}{n_i + n_o}$, and n_i and n_o are number of

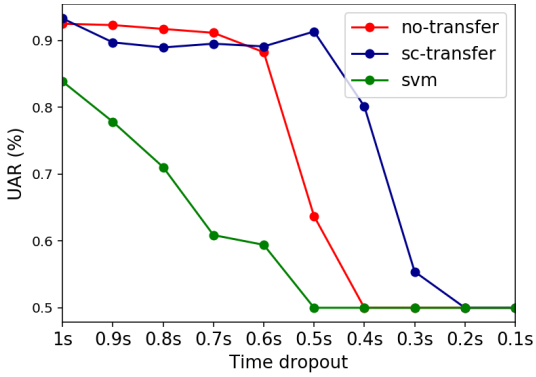

Figure 2: *Audio length analysis highlighting the impact of using shorter amounts of input audio on UAR performance.*

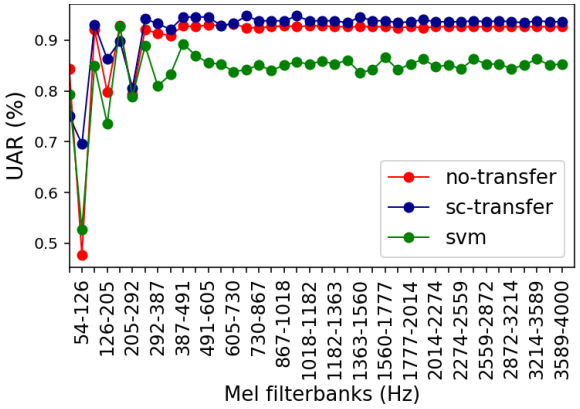

Figure 3: *Frequency response analysis of the relative importance of different Mel filterbanks on UAR performance. Each point represents the performance after removing the corresponding Mel filterbank.*

units in the input and output layers, respectively. This initialization scheme scales the weights in such a way that they are not too small to diminish or too large to explode through the network's layers during training.

3.5. Analysis

3.5.1. Performance

We evaluate the performance of our models on the target task by tracking the following metrics: sensitivity (recall on asphyxia class), specificity (recall on normal class), and the unweighted average recall (UAR). We use the UAR on the validation set for choosing best hyperparameter settings. The UAR is a preferred choice over accuracy since the classes in the Chillanto dataset are imbalanced.

3.5.2. Robustness

Noise. We analyze our models for robustness to 4 different noise situations: Gaussian noise $\mathcal{N}(0, 0.1)$, sounds of children playing, dogs barking and sirens. In each case, we insert the noise in increasing magnitude to the test data and monitor the impact on classification performance of the model.

Audio length. We also evaluate the response of each model to varying lengths of audio, since in the real-world a diagnostic system must be able to work with as much data as is available. To achieve this, we test the models on increasing lengths of the test data, starting from 0.1s to the full 1s segment, in 0.1 increments.

Frequency response. The response of the models to variations in frequency domain is important as this could reveal underlying characteristics of the data. We know as well that perinatal asphyxia alters the frequency patterns in cry. To discover what range of frequencies are most sensitive in detecting perinatal asphyxia, we conduct an ablation exercise where features extracted from a different filterbanks in the MFCC are zeroed out. We measure the response of our models by monitoring the drop in performance for the frequency ranges in each mel-filterbank.

3.5.3. MFCC Embeddings

In order to further investigate the nature of the embedding learned by each model, we apply principal component analysis (PCA) to the learned final-layer embeddings for all models [24]. By applying PCA, we hope to gain insight on the extent to which the embedding space captures unique information.

Table 2: *Performance – mean (standard error) - of different models in predicting perinatal asphyxia.*

Model	UAR %	Sensitivity %	Specificity %
SVM	84.4 (0.4)	81.6 (0.7)	87.2 (0.2)
no-transfer	80.0 (2.5)	71.8 (5.8)	88.1 (0.8)
sc-transfer	**86.5 (1.1)**	**84.1 (2.2)**	**88.9 (0.4)**
sitw-transfer	81.1 (1.7)	72.7 (3.5)	89.5 (0.2)
vctk-transfer	80.7 (1.0)	72.2 (2.1)	89.1 (0.3)

4. Experiments

4.1. Training details

There were a total of 1,389 infant cry samples (1,049 normal and 340 asphyxiated) in the Chillanto dataset. The samples were split into training, validation and test sets, with a 60:20:20 ratio, and under the constraint that samples from the same patients were placed in the same set.

Each source task was trained, fine-tuning hyperparameters as necessary to obtain performance comparable with the literature. For transfer learning on the target task, models were trained for 50 epochs using stochastic gradient descent with an initial learning rate of 0.001 (decreasing to 0.0001 after 15 epochs), a fixed momentum of 0.9, batch size of 50, and hinge loss function. We used a weighted balanced sampling procedure for mini-batches to account for class imbalance. We also applied data augmentation via random time-shifting of the audio recordings. Both led to up to 7% better UAR scores when training source and target models.

4.2. Performance on source tasks

Our model architecture achieved accuracies of 94.8% on word recognition task (Speech Commands), 91.9% on speaker identification (VCTK) and 90.2% on gender classification (SITW). These results are comparable to previous work. See [23, 25] for reference [1].

[1] SITW to our knowledge has not been used for gender classification, even though this data is available

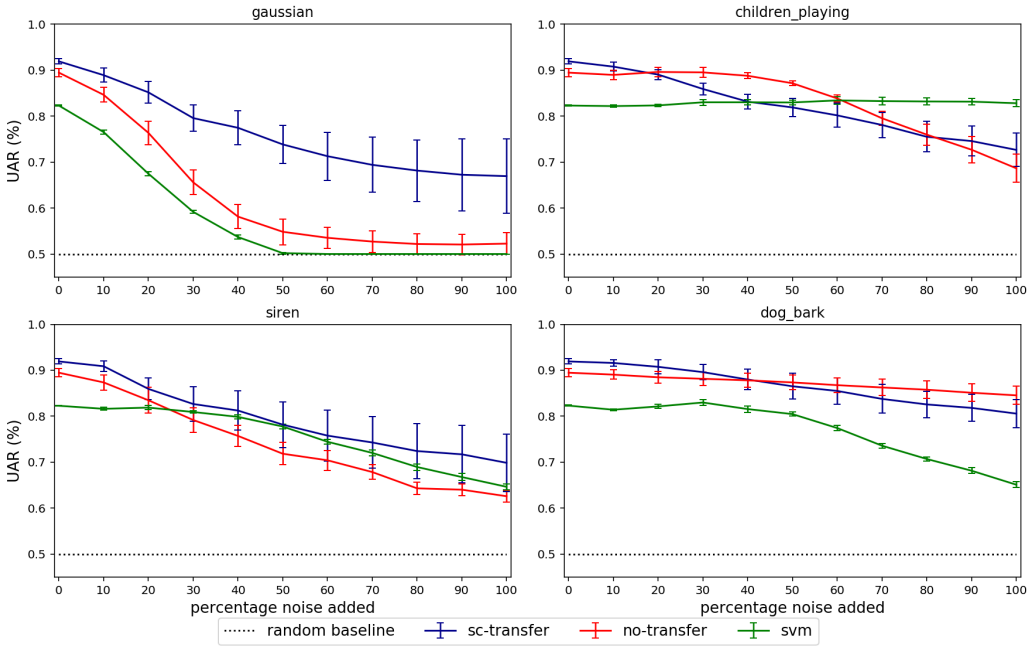

Figure 4: *Performance of models under different noise conditions.*

4.3. Performance on target task

Table 2 summarizes the performance of all models on the target task. The best performing model was pre-trained on the word recognition task (sc-transfer) and attained a UAR of 86.5%. This model also achieves the highest sensitivity and specificity 84.1% and 88.9% respectively. All other transfer models performed better than *no-transfer* , suggesting that transfer learning resulted in better or at least as good an initialization. The SVM was the second best performing model and had the lowest variance among all models in its predictions.

4.4. Robustness Analysis

In most cases, our results suggest that neural models have overall increased robustness. We focused on the top transfer model *sc-transfer*, *no-transfer* and the SVM. Figure 4, shows the response of the models to different types of noise, revealing that in all but one case the neural models degrade slower than the SVM. Results from Figure 2 suggest that the neural models are also capable of high UAR scores for short audio lengths, with *sc-transfer* maintaining peak performance when evaluated on only half (0.5s) of the test signals.

From our analysis of the models' responses to filterbank frequencies (Figure 3), we observe that (i) the performance of all models (unsurprisingly) only drops in the range of the fundamental frequency of infant cries, i.e. up to 500Hz [26] and (ii) *sc-transfer* again is the most resilient model across the frequency spectrum.

4.5. Visualization of embeddings

Figure 5 shows cumulative variance explained by the principal components (PC) of the neural model embeddings. Whereas in *no-transfer*, the top 2 PCs explain nearly all variance in the data (91%), in *sc-transfer* they represent only 52%—suggesting that the neural transfer leads to an embedding that is intrinsically higher dimensional and richer than the *no-transfer* counterpart.

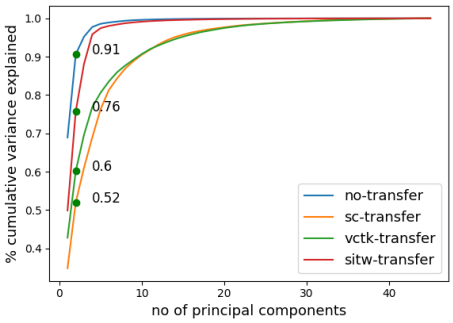

Figure 5: *Cumulative variance explained by all principal components (left) and the top 2 principal components on the Chillanto test data (right) based on embeddings of* no-transfer *model.*

5. Conclusion and Discussion

We compared the performance of a residual neural network (ResNet) pre-trained on several speech tasks in classifying perinatal asphyxia. Among the transfer models, the one based on a word recognition task performed best, suggesting that the variations learned for this task are most analogous and useful to our target task. The support vector machine trained directly on MFCC features proved to be a strong benchmark, and if variance in predictions was of concern, a preferred model. The SVM, however, was clearly less robust to pertubations in time- and frequency-domains than the neural models. This work reinforces the modelling power of deep neural networks. More importantly, it demonstrates the value of a transfer learning approach to the task of predicting perinatal asphyxia from the infant cries—a task of critical relevance for improving the accessibility of pediatric diagnostic tools.

6. References

[1] World Health Organisation, "Children: reducing mortality," *Media Centre*, 2017.

[2] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–36.

[3] K. Michelsson, P. Sirviö, and O. Wasz-Höckert, "Pain cry in full-term asphyxiated newborn infants correlated with late findings," *Acta Pædiatrica*, vol. 66, no. 5, pp. 611–616, 1977.

[4] O. F. Reyes-Galaviz and C. A. Reyes-Garcia, "A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks," in *9th Conference Speech and Computer*, 2004.

[5] C. C. Onu, "Harnessing infant cry for swift, cost-effective diagnosis of perinatal asphyxia in low-resource settings," in *2014 IEEE Canada International Humanitarian Technology Conference-(IHTC)*. IEEE, 2014, pp. 1–4.

[6] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018.

[7] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[9] L. L. LaGasse, A. R. Neal, and B. M. Lester, "Assessment of infant cry: acoustic cry analysis and parental perception," *Mental retardation and developmental disabilities research reviews*, vol. 11, no. 1, pp. 83–93, 2005.

[10] B. M. Lester, C. Z. Boukydis, C. T. Garcia-Coll, and W. T. Hole, "Colic for developmentalists," *Infant Mental Health Journal*, vol. 11, no. 4, pp. 321–333, 1990.

[11] P. S. Zeskind and B. Lester, "Analysis of infant crying," *Biobehavioral assessment of the infant*, pp. 149–166, 2001.

[12] K. Michelsson, P. Sirviö. A, and O. Wasz-Höckert, "Sound spectrographic cry analysis of infants with bacterial meningitis," *Developmental Medicine & Child Neurology*, vol. 19, no. 3, pp. 309–315, 1977.

[13] K. Michelsson, K. Eklund, P. Leppänen, and H. Lyytinen, "Cry characteristics of 172 healthy 1-to 7-day-old infants," *Folia phoniatrica et logopaedica*, vol. 54, no. 4, pp. 190–200, 2002.

[14] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[16] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 625–660, 2010.

[17] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[18] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, 2007, pp. 153–160.

[19] J. M. K. Veaux, Christophe; Yamagishi, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.

[20] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database." in *Interspeech*, 2016, pp. 818–822.

[21] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] R. Tang and J. Lin, "Deep Residual Learning for Small-footprint Keyword Spotting," Tech. Rep., 2018.

[24] I. Jolliffe, *Principal component analysis*. Springer, 2011.

[25] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 040–10 050.

[26] R. P. Daga and A. M. Panditrao, "Acoustical analysis of pain cries in neonates: Fundamental frequency," *Int. J. Comput. Appl. Spec. Issue Electron. Inf. Commun. Eng ICEICE*, vol. 3, pp. 18–21, 2011.