NEURAL TRANSFER LEARNING FOR CRY-BASED DIAGNOSIS OF PERINATAL ASPHYXIA

Charles C. Onu, Jonathan Lebensold, William L. Hamilton & Doina Precup
School of Computer Science
McGill University
Montral, QC H3A 0G4, Canada
{charles.onu, jonathan.maloney-lebensold}@mail.mcgill.ca
{wlh, 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 based on wearables and smartphones. 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 the task of predicting 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.

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 (World Health Organisation, 2017). 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 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.

We train source tasks on freely available large datasets of adult speech in order to improve performance on the relatively small Baby Chillanto Infant Cry dataset (Reyes-Galaviz & Reyes-Garcia, 2004). 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.

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

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

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 (LaGasse et al., 2005). In addition, cry generation and respiration are both coordinated by the same regions of the brain (Lester et al., 1990; Zeskind & Lester, 2001). The study of how pathologies affect infant crying dates back to the 1970s and 1980s with the work of Michelsson et al. (1977a;b; 2002). 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-Galaviz & Reyes-Garcia (2004) collected the Chillanto Infant Cry database with the objective of applying statistical learning techniques in classifying deafness, asphyxia, pain and other conditions. Cry recordings from 69 subjects were broken into 1 second segments to form a database of 1,389 examples. 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 (2014) improved the precision and recall to 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 (Glorot & Bengio, 2010; He et al., 2015). 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 (Erhan et al., 2010) or supervised (Yosinski et al., 2014; Bengio et al., 2007) 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.

3.1 Source tasks

We choose 3 source tasks — speaker identification, gender classification, word recognition — with corresponding audio datasets: VCTK (Veaux, 2017), Speakers in the Wild (SITW) (McLaren et al., 2016), and Speech Commands (Warden, 2018) . Table 1 briefly describes the datasets used for each task.

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 30ms with a 10ms 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) (He et al., 2016) architecture with average pooling, for training. To assure even comparison across source tasks, and to facilitate transfer learning, we adopt a single network architecture: the *res8* as in (Tang & Lin, 2018). 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. 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 (Onu, 2014). The other is a *res8* model whose initial weights are drawn randomly from a Glorot distribution (Glorot & Bengio, 2010).

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

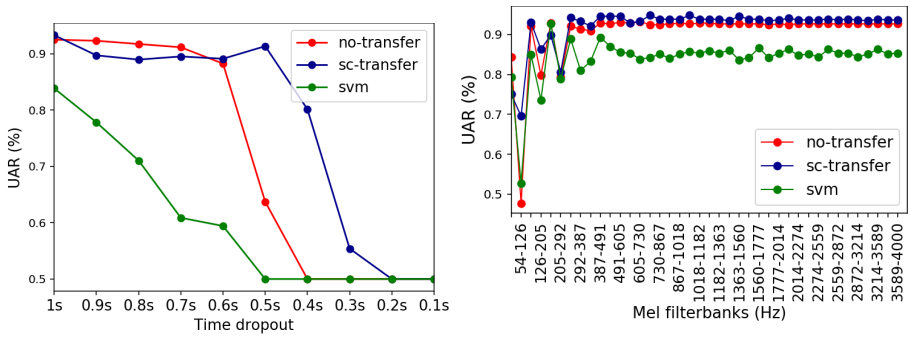
Noise. We analyze our models for robustness to 4 different noise situations: Gaussian noise $\mathcal{N}(0, 0.1)$, sounds of children playing, dogs barking and sirens.

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.

Frequency response. To discover what range of frequencies are most sensitive to detecting perinatal asphyxia, we conduct an ablation exercise where features extracted from a different filterbanks in the MFCC are zeroed out.

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 (Jolliffe, 2011). By applying PCA, we hope to gain insight on the extent to which the embedding space captures unique information.



(a) (b)

Figure 1: Response of models to missing signals in time (a) and frequency (b) domain.

4 EXPERIMENTS

4.1 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 (Tang & Lin, 2018; Arik et al., 2018) for reference [1].

4.2 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.

4.3 ROBUSTNESS ANALYSIS

In most cases, our results suggest that neural models have overall increased robustness. In Figure 1(a), we see 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 duration of test signals.

Our analysis of the models' responses to filterbank frequencies (Figure 1(b)), revealed 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 (Daga & Panditrao, 2011) and (ii) *sc-transfer* again is the most resilient model across the frequency spectrum.

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

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)

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.

References

Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*, pp. 10040–10050, 2018.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pp. 153–160, 2007.

Raina P Daga and Anagha M Panditrao. Acoustical analysis of pain cries in neonates: Fundamental frequency. *Int. J. Comput. Appl. Spec. Issue Electron. Inf. Commun. Eng ICEICE*, 3:18–21, 2011.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ian Jolliffe. *Principal component analysis*. Springer, 2011.

Linda L LaGasse, A Rebecca Neal, and Barry M Lester. Assessment of infant cry: acoustic cry analysis and parental perception. *Mental retardation and developmental disabilities research reviews*, 11(1):83–93, 2005.

Barry M Lester, CF Zachariah Boukydis, Cynthia T Garcia-Coll, and William T Hole. Colic for developmentalists. *Infant Mental Health Journal*, 11(4):321–333, 1990.

Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. The speakers in the wild (sitw) speaker recognition database. In *Interspeech*, pp. 818–822, 2016.

Katarina Michelsson, Pertti SirviÖ, and OLE WASZ-HÖCKERT. Pain cry in full-term asphyxiated newborn infants correlated with late findings. *Acta Pædiatrica*, 66(5):611–616, 1977a.

Katarina Michelsson, Pertti SirviöM. A, and Ole Wasz-Höckert. Sound spectrographic cry analysis of infants with bacterial meningitis. *Developmental Medicine & Child Neurology*, 19(3):309–315, 1977b.

Katarina Michelsson, Kenneth Eklund, Paavo Leppänen, and Heikki Lyytinen. Cry characteristics of 172 healthy 1-to 7-day-old infants. *Folia phoniatrica et logopaedica*, 54(4):190–200, 2002.

Charles 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)*, pp. 1–4. IEEE, 2014.

Orion F Reyes-Galaviz and Carlos Alberto 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.

Raphael Tang and Jimmy Lin. Deep Residual Learning for Small-footprint Keyword Spotting. Technical report, 2018.

Junichi; MacDonald Kirsten. Veaux, Christophe; Yamagishi. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit, 2017.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition, 2018.

World Health Organisation. Children: reducing mortality. *Media Centre*, 2017.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328, 2014.

PHILIP SANFORD Zeskind and BM Lester. Analysis of infant crying. *Biobehavioral assessment of the infant*, pp. 149–166, 2001.

6 APPENDIX

6.1 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.

6.2 NOISE ANALYSIS

Figure 2, 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.

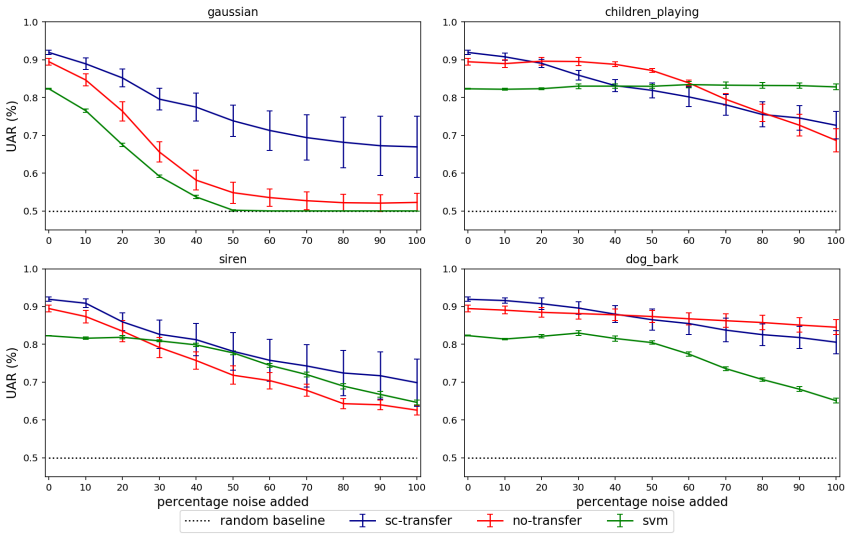


Figure 2: Performance of models under different noise conditions.

6.3 VISUALIZATION OF EMBEDDINGS

Figure 3 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.

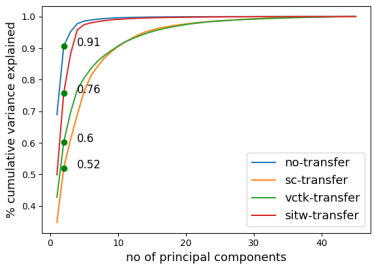


Figure 3: Cumulative variance explained by all principal components (left) and the top 2 principal components on the Chillanto test data (right) based on embeddings of *no-transfer* model.