SELF-SUPERVISED LEARNING FOR INFANT CRY ANALYSIS

Arsenii Gorin[⋆]*, Cem Subakan*[♮♯♭]*, Sajjad Abdoli*[⋆]*, Junhao Wang*[⋆]*, Samantha Latremouille*[⋆]*, Charles Onu*[⋆♭]

[⋆]Ubenwa Health, Montréal, Canada [♮]Université Laval, Québec City, Canada
[♯]Concordia University, Montréal, Canada [♭]Mila-Québec AI Institute, Montréal, Canada

ABSTRACT

In this paper, we explore self-supervised learning (SSL) for analyzing a first-of-its-kind database of cry recordings containing clinical indications of more than a thousand newborns. Specifically, we target cry-based detection of neurological injury as well as identification of cry triggers such as pain, hunger, and discomfort. Annotating a large database in the medical setting is expensive and time-consuming, typically requiring the collaboration of several experts over years. Leveraging large amounts of unlabeled audio data to learn useful representations can lower the cost of building robust models and, ultimately, clinical solutions. In this work, we experiment with self-supervised pre-training of a convolutional neural network on large audio datasets. We show that pre-training with SSL contrastive loss (SimCLR) performs significantly better than supervised pre-training for both neuro injury and cry triggers. In addition, we demonstrate further performance gains through SSL-based domain adaptation using unlabeled infant cries. We also show that using such SSL-based pre-training for adaptation to cry sounds decreases the need for labeled data of the overall system.

***Index Terms*—** Self-Supervised Learning, Infant Cry Classification, Audio Classification, Transfer Learning, Domain Adaptation

1. INTRODUCTION

Crying is the primary means by which babies communicate with the world. Researchers have been interested in infant cry analysis since the early 1960s [1]. Cry characteristics may help us to understand basic baby needs (hunger, pain, etc.) and, more importantly, can be analyzed for the early and non-invasive detection of various diseases [2]. For example, clinical research has reported that certain infant cry characteristics are correlated with birth asphyxia [3]. This multi-causal condition frequently leads to severe health problems, including neurological injury and even death. Various methods based on signal processing [4], statistical modeling [5, 6] and deep learning [7–10] have been explored for finding clinical and other insights using cry recordings.

One of the main challenges in baby cry analysis is data acquisition. Today, cry sounds are not part of routine medical records, so obtaining a database requires targeted efforts such as a clinical study. These are expensive to conduct and typically require the collaboration of several hospital staff over the years. Most machine learning (ML) research on pathology detection from cry sounds was done using the Baby Chillanto [11] database, which contains only six patients diagnosed with birth asphyxia.

From an ML problem point of view, cry classification is analogous to general audio classification, where deep convolutional neural networks (CNNs) have excelled as the state-of-the-art. Recently, [12] demonstrated that Pre-trained Audio Neural Networks (PANNs) - large CNNs pre-trained on generic audio - transferred to a wide range of audio pattern recognition tasks outperformed several previous state-of-the-art systems. Since then, PANNs have been widely adopted for various audio tasks, including emotion recognition from speech [13] and COVID-19 detection from cough [14].

Another popular paradigm in audio classification state-of-the-art is self-supervised learning (SSL) - a method to obtain high-quality representations by training on unlabeled data. SSL has revolutionized the fields of Natural Language Processing and Computer Vision and is currently widely adopted in audio processing [15]. A neural network (encoder) pre-trained with SSL can be seen as a non-linear mapping of an audio sequence to a hidden representation - an embedding. The embeddings can be used as input to a classifier trained on a specific task with a supervised objective (using labeled data and conventional cross-entropy loss). This approach is common for benchmarking various SSL models on multiple diverse audio tasks [16, 17]. Recently, a similarity-based contrastive learning method called SimCLR introduced in Computer Vision [18, 19] demonstrated good performance in multiple audio tasks [17, 20], including music analysis [21, 22]. SimCLR maximizes the similarity between modified (distorted) views of the same object. For audio, such distortion can be done, for example, by mixing random audio samples [17], spectrogram masking [23] in [20], or/and reverberation, pitch shifting, etc [21].

In this paper, we experiment with PANNs using both supervised and self-supervised pre-training to learn representations for two downstream tasks. The first task is classifying brain injury (resulting from birth asphyxia), and the second is predicting cry triggers (pain, hunger, discomfort). The methods are tested on a unique clinical database of newborn cries collected by Ubenwa Health in collaboration with hospitals across three countries [24].

In addition, we evaluate the impact of SimCLR-based adaptation of PANNs using unlabeled cries inspired by self-supervised domain adaptation in Speech [25] and Natural Language Processing [26]. It should be noted that speech and audio SSL state-of-the-art frequently uses transformers instead of CNNs and relies on different learning objectives [15]. However, our preliminary experiments with some popular pre-trained speech and audio transformers (specifically, Wav2Vec2.0 [27], HuBERT [28], WavLM [29] and SSAST [30]) have not shown sufficient improvements but generally required many parameters to be adapted and hyperparameters tuned. We, therefore, focus on CNN and SimCLR, which demonstrated a good balance of accuracy and adaptation complexity.

2. METHODOLOGY

We illustrate our 3-stage approach in Figure 1, indicating the SSL pipeline and datasets used at different stages. In the rest of this section, we provide details on each stage.

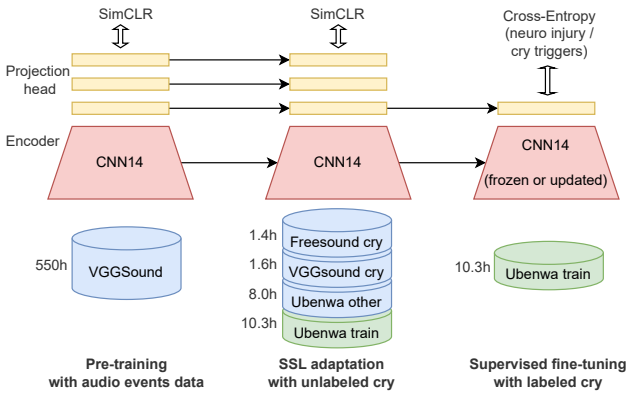

Fig. 1: Summary of the proposed SSL-based trainining pipeline. **(Left)** First the CNN14 backbone is pre-trained on the VGGSound dataset using SimCLR. **(Middle)** The CNN14 backbone is further pre-trained via SSL using cry-specific datasets. We denote this stage *SSL cry adaptation*. **(Right)** The model is finally trained with supervision on a labeled cry dataset.

2.1. Encoder Architecture

As the backbone encoder, we adopt CNN14 introduced in [12]. The encoder is pre-trained on the VGGSound database as in [20].

There are a few minor differences between the CNN14 proposed in [12] and the models that we adopted from [20]. First, [20] uses VGGSound [31] instead of AudioSet [32] for pre-training. VGGSound is an automatically curated collection of 550 hours of YouTube recordings, which is about ten times smaller than AudioSet but with a single label per audio recording and a lower noise level. Second, [20] used more log-Mel filterbanks (80 vs 64) and shorter audio chunks in training (4 seconds vs 10). The network overall has about 80 million parameters.

2.2. Pre-Training

We compare two identical CNN14 models pre-trained in a supervised and self-supervised manner on VGGSound. SimCLR pre-training relies on an additional projection head - a three-layer multi-layer perceptron with 2048 hidden units and a bottleneck layer with 512 hidden units. In both supervised and self-supervised initial pre-training, the model is updated with stochastic gradient descent using mini-batches of 32 four-second chunks randomly cut from VGGSound clips. When extracting embeddings of validation and test data, the model processes arbitrary-length audio sequences using global temporal pooling.

2.3. Supervised Fine-Tuning and Evaluation

To study various aspects of pre-trained models, we conduct three evaluations. First, we use *linear probing*, where the frozen encoder extracts features for a linear classifier. Second, in addition to learning the linear classifier, we also update statistics of batch normalization parameters of the encoder while still keeping other parameters frozen. The primary motivation is to compensate for the difference between pre-training and target data characteristics [33, 34]. This also allows us to understand what portion of improvement obtained by SSL fine-tuning with cry data may be attributed to a simple update of normalization parameters occurring naturally during SimCLR. Finally, aiming to improve classification results further, we perform *end-to-end fine-tuning*, where the encoder parameters are optimized jointly with the classifier on target tasks.

The supervised training is done for 50 epochs in all three settings, and the model with the best validation score is selected. We use weighted random sampling to balance the distribution of classes during training. We use Adam optimizer [35] with a learning rate reduced two times if validation loss does not improve for three epochs. For end-to-end supervised fine-tuning, a much smaller and separately tuned learning rate is used for the encoder. Also, the encoder learning rate is linearly increased from zero to target one over the first ten epochs. In all experiments, the learning rates of the classifier and encoder are optimized using grid search. For the model with the best validation score, we repeat the experiment 10 times and report the mean area under the receiver operating characteristic curve (AUC) along with standard error. For multi-class classification (triggers), the macro averaged AUC is computed using a one-versus-rest approach.

Similar to [19], in our preliminary studies, we found that keeping one layer of projection head of the SSL pre-trained model leads to slightly better results. Therefore, we always transfer from layer 1 of the projection head.

2.4. Self-Supervised Cry Adaptation

To improve the quality of SSL representations, we further explore a second stage of self-supervised domain adaptation. Our goal is to adapt the encoder from the domain of general audio sounds to the domain of cry sounds, using unlabeled cry data (middle column of Figure 1). This SSL adaptation is done by reusing the encoder and the projector from CNN14 trained on VGGSound and running 100 more epochs of SimCLR using infant cry data with the same learning rate and schedule as the initial pre-training. The only difference is that we use batch size 200 for faster training and because larger batches performed better for SimCLR in the literature [21]. Notably, we did not find a significant difference trying to improve the initial SSL pre-trained model of [20] by using more training epochs and large batches without cry sounds.

Similar to the initial SSL pre-trained models, the cry-adapted ones are evaluated with linear probing and end-to-end settings described in the previous section.

3. EXPERIMENTAL SETUP

3.1. Dataset Description

This study is based on a subset of a larger Ubenwa newborn cry clinical database collected from five hospitals in Nigeria, Brazil, and Canada since 2020 [24]. For most infants, one recording is done after birth and one before discharge. A neurological exam was conducted on all infants, and the level of neuro injury was recorded using a four-scale measure called Sarnat score [36]. For classification, we categorize the recordings into two groups: normal (no neuro injury) and injured (mild, moderate, or severe injury). We further split the data into train, validation, and test, making sure the recordings of a given patient go to one subset. Table 1 summarizes key statistics of the data.

In addition, the recordings are annotated with a trigger - the primary reason for crying determined by the medical or research staff. In this study, we use a subset of three main triggers resulting in 267 recordings of discomfort, 200 hunger, and 682 pain.

	Healthy			Neuro Injury		
	Train	Val	Test	Train	Val	Test
# recordings	1360	247	238	92	40	45
# patients	885	165	163	75	33	38
# hours	10.3	1.9	2.0	0.8	0.3	0.3

Table 1: The description of our neurological injury dataset.

Compared to the commonly used Chillanto database [11], our dataset has much more patients with neurological injury (146 vs 6) and more annotated cry signals in general (14.2 vs 0.6 hours). In our database, cry recording is a segment of arbitrary length (a second to a few minutes). Conversely, in Chillanto, the recordings correspond to 1-second cry expirations annotated as belonging to the healthy or sick infant. However, there is insufficient evidence to determine whether every cry expiration of a sick infant has distinct characteristics from healthy infants or if only some expirations have them. Furthermore, cry expirations of a single infant are generally quite similar, so if recordings are split randomly for training, testing, and validation without considering infant identities (for example, in [10]), the resulting performance may be over-estimated.

For SSL experiments, we also use an additional 8 hours of Ubenwa unlabeled cries along with 1.6 hours available in VGGSound and about 1.4 hours collected from Freesound website[1] using search query "baby cry".

3.2. Baselines

While our primary focus is on SSL pre-training and fine-tuning, we use two supervised approaches as baselines that do not rely on pre-training. The first system is a statistical model using ComParE 2016 [37] acoustic features extracted with OpenSmile toolkit [38]. The feature set contains 6373 recording-level derivatives (mean, standard deviation, etc.) of various acoustic descriptors (MFCC, pitch, jitter, etc.). It is commonly used in computational paralinguistics and infant cry classification [6]. The model and hyperparameters are selected using grid search and 10-fold cross-validation, maximizing the average AUC score. For this study, random forest, logistic regression, and support vector machine were considered in model selection. The second baseline is CNN14 described in Section 2 using random initialization, no pre-training and end-to-end supervised fine-tuning described in Section 2.3 The performance of baselines on neurological injury and triggers is summarized in Table 2 for five experiments with different random seeds. CNN14 without pre-training does not outperform the statistical baseline, which is not surprising given that our training datasets are quite small for such a model.

Model	**AUC % (mean and standard error)**	
	Neuro Injury	Cry Triggers
Statistical	75.1 ± 0.4	71.1 ± 0.2
CNN14	74.6 ± 1.7	59.8 ± 0.9

Table 2: Baseline performance obtained without any pre-training

4. SSL EXPERIMENTS

The main results of neurological injury and cry trigger experiments are summarized in Tables 3 and 4. The first row in both tables refers to supervised training with random initialization and is provided to give an idea about the model performance without pre-training.

The last three columns in Table 3 and Table 4 summarize the performances obtained after fine-tuning the network with supervised training. From left to right, the results in the tables correspond to the following:

1. Evaluation with linear probing, where a linear layer is fine-tuned on top of the frozen encoder weights (Linear).
2. Evaluation with linear probing, with batch-norm layers updated during fine-tuning (Linear+BN).
3. End-to-end fine-tuning where the linear layer and whole encoder are updated (End-to-end).

	Pre-training	**SSL cry adapt. Dataset**	**% AUC after fine-tuning** Linear	Linear+BN	End-to-end
1	–	–	60.6 ± 1.5	60.4 ± 1.3	74.6 ± 1.7
2	supervised	–	75.5 ± 0.6	75.9 ± 0.7	80.0 ± 0.7
3	SSL	–	71.3 ± 0.9	78.5 ± 1.2	83.9 ± 0.6
4	SSL	train set	78.8 ± 0.5	78.0 ± 0.7	80.8 ± 0.8
5	SSL	+ 11h cry	79.8 ± 0.4	81.3 ± 0.5	81.3 ± 0.7
6	SSL	+ replayVGG	80.8 ± 0.5	83.3 ± 0.6	**85.0** ± 0.9

Table 3: Performance of neuro injury classification under various types of pre-training and fine-tuning strategies. **Column 1** indicates the type of pre-training. Note that for the first row, no pre-training is applied. For the second row, supervised pre-training on the VGGSound dataset is applied. The rest of the rows use SSL-based pre-training on the VGGSound. **Column 2** indicates the datasets used in SSL cry adaptation. Note that in rows 4-5-6, the datasets used in SSL cry adaptation are cumulative. The 4th row uses neuro injury train, the 5th adds 11h cry to the neuro injury train, and the 6th adds a replay buffer from the VGG Sound dataset to the previous datasets from rows 4-5. **Columns 3-5** show the % AUC (with mean and standard error) obtained with different supervised fine-tuning strategies (after the SSL fine-tuning as shown in Figure 1).

	Pre-training	**SSL cry adapt. Dataset**	**% AUC after fine-tuning** Linear	Linear+BN	End-to-end
1	–	–	57.1 ± 2.6	60.7 ± 0.5	59.8 ± 0.9
2	supervised	–	67.9 ± 1.9	68.0 ± 1.7	68.1 ± 1.6
3	SSL	–	65.9 ± 0.8	69.5 ± 0.7	69.0 ± 0.9
4	SSL	neuro injury train	71.7 ± 0.5	**75.4** ± 0.8	72.4 ± 1.4
5	SSL	+ 11h cry	74.5 ± 0.4	74.7 ± 0.4	72.0 ± 1.8
6	SSL	+ replayVGG	74.2 ± 0.4	**75.6** ± 0.6	74.4 ± 0.7

Table 4: Performance of cry trigger classification. We follow the same structure used in Table 3, therefore the same caption applies.

The second and third rows in both tables compare supervised and self-supervised initial pre-training with VGGSound. In these experiments, the cry database is used only for supervised fine-tuning (In other words, no-additional SSL cry adaptation stage is applied). We observe that, while simple linear probing performs better for supervised pre-training, the self-supervised pre-training achieves better results when updating BN statistics. Also, we see that with the end-to-end fine-tuning strategy SSL pre-training performs significantly better on neuro injury task.

Next, in rows 4-5-6 of Table 3 and Table 4 we show the results when additional SSL cry adaptation stage is employed. First, we

[1] https://freesound.org

fine-tune with SimCLR using only the neuro injury training dataset, as shown in row 4 of Table 3). For neuro injury, we observe that this significantly improves AUC for linear evaluation (71.3 to 78.8), but the improvement vanishes when BN is updated (78.5 and 78.0). We hypothesize that SimCLR in this experiment performs better mostly due to a significant domain mismatch between VGGSound that a simple BN update can compensate for. For cry triggers, we observe that SSL cry adaptation in general improves the performance obtained after supervised fine-tuning.

Next, as shown in the 5th row of Table 3 and Table 4 we further double the amount of unlabeled data for SimCLR SSL cry adaptation stage. This is achieved by adding an 8-hour portion of previously unused Ubenwa cry along with some unlabeled cry sounds from VGGSound and freesound. In total, these SSL cry adaptation datasets amount to approximately 11 hours of recording (hence it is called 11h cry in Table 3, and Table 4). This significantly improves the performance of linear evaluation with and without BN update for neuro injury.

This, however, is not the case for end-to-end fine-tuning, where the initial VGGSound pre-training results in better transfer for neuro injury (row 3 of Table 3). We hypothesize that the model loses its generalization properties that are useful for fine-tuning due to catastrophic forgetting [39] when we further adapt the model with SSL.

To mitigate this forgetting effect and preserve generalization properties that seem to be important for transfer learning, we perform SSL cry adaptation using replay technique from continual learning literature [40]. We show this on the last row of both tables. This is done by replaying 50% of the VGGSound dataset when applying SSL cry adaptation stage. We, therefore, observe that SSL cry adaptation with replay performs significantly better in all evaluations for neuro injury. We also obtain the best results on trigger classification using linear+BN evaluation with this approach.

5. TRAINING WITH SUBSETS OF LABELED DATA

Obtaining labeled, high-quality medical data is extremely time-intensive and expensive. Therefore, in this section, we aim to understand if a small amount of labeled data can still yield decent performance. Specifically, we aim to analyze how SSL cry adaptation helps in such a scenario and which supervised fine-tuning method gives the best performance. To this end, we experiment with linear+BN and end-to-end fine-tuning using randomized subsets of labeled neuro injury dataset.

In Figure 2 we show results for neuro injury classification using two models: the model pre-trained with SSL on VGGSound without SSL cry adaptation stage (row 3 of Table 3) and the model that obtains the best performance with SSL cry adaptation (last row of Table 3).

The SSL cry-adapted model (solid lines) consistently outperforms the not-adapted ones (dashed lines) with a larger margin as the size of the supervised subset is reduced. Another point to note is that we observe that for the not-adapted models, as the amount of labeled fine-tuning data decreases, the end-to-end fine-tuning strategy decreases more rapidly in performance. This could perhaps explain why end-to-end adaptation was performing worse than Linear+BN for cry triggers (Table 4), where less labeled data is available for fine-tuning compared to neuro injury.

We see that, very interestingly, with only 3% (a few dozen samples) of labeled data, we can still achieve about 70% AUC by using a cry-adapted model. Also, with only 20% of data, the adapted model significantly outperforms our supervised baselines. This showcases that SSL cry-adaptation has huge potential to obtain satisfactory model performance by only incorporating a small amount of labeled data in the supervised fine-tuning stage.

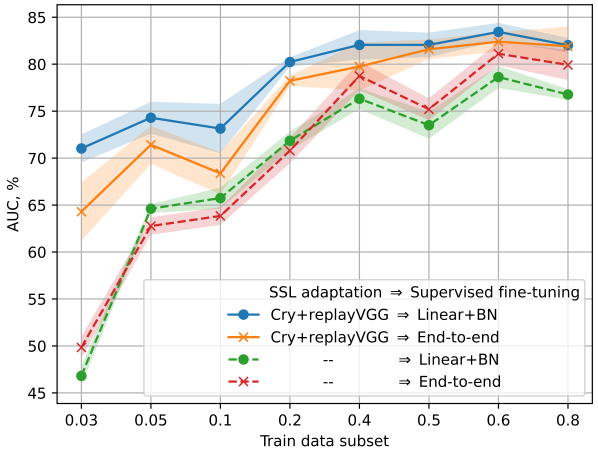

Fig. 2: Performance using subsets of labeled neuro injury data in supervised fine-tuning (Linear+BN and End-to-end). Solid lines - model with SimCLR cry adaptation (last row of Table 3), dashed - same model without adaptation (row 3 of Table 3). The filled areas show the standard error of AUC from five runs with different random seeds.

6. CONCLUSIONS

In this paper, we explored large-scale SSL pre-training for infant cry analysis, namely for detecting neurological injury and cry triggers. We observe that SSL pre-training performs significantly better than the conventional supervised pre-training, and both perform significantly better than training from random initialization. Furthermore, with limited annotated data, we observe that SSL adaption on cry-specific unlabeled data significantly decreases the need for labeled data in the supervised fine-tuning stage. We show that when we adapt the encoder through SSL using unlabeled cry data, the downstream performance for neurological injury is significantly improved. We therefore believe that, with many unlabeled cry recordings, this opens a promising research direction where it would be possible to train a classifier to detect new diseases using only a small amount of annotated cry sounds from the target population.

7. ACKNOWLEDGEMENT

We would like to thank the principal investigators from each site where cry data was collected: Dr. Uchenna Ekwochi (Enugu State University Teaching Hospital, Nigeria), Dr. Boma West and Dr. Datonye Briggs (River State University Teaching Hospital, Nigeria), Dr. Peter Ubuane (Lagos State University Teaching Hospital, Nigeria) Dr. Gabriel Variane (Santa Casa de Misericórdia de São Paulo, Brazil), and Dr. Guilherme Sant'Anna (Montreal Children's Hospital, Canada). We also thank Zhepei Wang for providing pre-trained models and his valuable advice. Finally, we are indebted to all infants, their families, and the medical and research staff at all sites for their participation and help with the acquisition of the cry data.

8. REFERENCES

[1] O Wasz-Höckert et al., "The identification of some specific meanings in infant vocalization," *Experientia*, vol. 20, 1964.

[2] Chunyan Ji et al., "A review of infant cry analysis and classification," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

[3] Katarina Michelsson et al., "Pain cry in full-term asphyxiated newborn infants correlated with late findings," *Acta Pædiatrica*, vol. 66, no. 5, 1977.

[4] Lichuan Liu et al., "Infant cry language analysis and recognition: an experimental approach," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, 2019.

[5] Gustavo Z Felipe et al., "Identification of infants' cry motivation using spectrograms," in *International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2019.

[6] Joanna J Parga et al., "Defining and distinguishing infant behavioral states using acoustic cry analysis: is colic painful?," *Pediatric research*, vol. 87, no. 3, 2020.

[7] Charles C Onu et al., "Neural transfer learning for cry-based diagnosis of perinatal asphyxia," *INTERSPEECH*, 2019.

[8] Najla Al Futaisi et al., "VCMNet: Weakly supervised learning for automatic infant vocalisation maturity analysis," in *International Conference on Multimodal Interaction*, 2019.

[9] Turgut Ozseven, "Infant cry classification by using different deep neural network models and hand-crafted features," *Biomedical Signal Processing and Control*, vol. 83, 2023.

[10] Hemant A Patil et al., "Constant Q Cepstral coefficients for classification of normal vs. Pathological infant cry," in *ICASSP*. IEEE, 2022.

[11] 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 *SPECOM*, 2004.

[12] Qiuqiang Kong et al., "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 28, 2020.

[13] Andreas Triantafyllopoulos and Björn W Schuller, "The role of task and acoustic similarity in audio transfer learning: insights from the speech emotion recognition case," in *ICASSP*. IEEE, 2021.

[14] Edresson Casanova et al., "Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in ComParE 2021.," in *INTERSPEECH*, 2021.

[15] Abdelrahman Mohamed et al., "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[16] Shu-wen Yang et al., "SUPERB: Speech processing universal performance benchmark," *INTERSPEECH*, 2021.

[17] Luyu Wang et al., "Towards learning universal audio representations," in *ICASSP*. IEEE, 2022.

[18] Ting Chen et al., "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[19] Ting Chen et al., "Big self-supervised models are strong semi-supervised learners," *NeurIPS*, vol. 33, 2020.

[20] Zhepei Wang et al., "Learning representations for new sound classes with continual self-supervised learning," *IEEE Signal Processing Letters*, vol. 29, 2022.

[21] Janne Spijkervet and John Ashley Burgoyne, "Contrastive learning of musical representations," *ISMIR*, 2021.

[22] Matthew C McCallum et al., "Supervised and unsupervised learning of audio representations for music understanding," *ISMIR*, 2022.

[23] Daniel S Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," *INTERSPEECH*, 2019.

[24] CC Onu et al., "Ubenwa: Cry-based diagnosis of birth asphyxia," *NIPS Workshop on Machine Learning for the Developing World*, 2017.

[25] Zhengyang Chen et al., "Self-supervised learning based domain adaptation for robust speaker verification," in *ICASSP*, 2021.

[26] Suchin Gururangan et al., "Don't stop pretraining: Adapt language models to domains and tasks," *ACL*, 2020.

[27] Alexei Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NIPS*, vol. 33, 2020.

[28] Wei-Ning Hsu et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021.

[29] Sanyuan Chen et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, 2022.

[30] Yuan Gong et al., "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36.

[31] Honglie Chen et al., "VGGSound: A Large-scale Audio-Visual Dataset," in *ICASSP*, 2020.

[32] Jort F. Gemmeke et al., "AudioSet: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[33] Jonathan Frankle et al., "Training batchnorm and only batchnorm: On the expressive power of random features in CNNs," *ICLR*, 2020.

[34] Moslem Yazdanpanah et al., "Revisiting learnable affines for batch norm in few-shot transfer learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[35] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.

[36] Harvey B Sarnat and Margaret S Sarnat, "Neonatal encephalopathy following fetal distress: a clinical and electroencephalographic study," *Archives of neurology*, vol. 33, no. 10, 1976.

[37] Björn Schuller et al., "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *INTERSPEECH*, 2016.

[38] Florian Eyben et al., "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. of the 18th ACM international conference on Multimedia*, 2010.

[39] Robert M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, vol. 3, no. 4, 1999.

[40] Anthony Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, vol. 7, no. 2, 1995.