International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Spectrogram Enhanced SimCLRV2 Emotional Representation Strategy for Kid's Speech using Multisource Transfer Learning in CNN

Preethi V.*[1], V. Elizabeth Jesi*[2]

Submitted:12/03/2024 **Revised**: 27/04/2024 **Accepted**: 04/05/2024

Abstract: A model that learns voice image representation, which outperform past techniques is developed. An unlabeled dataset is used to develop a semi-supervised representation contrastive learning strategy that picks and compares anchor, negative, and positive (APN) features with just 1% training data of real-time kid's speech. Most of kid's emotion detection model evaluate its model with adult's emotion dataset, which may not give accurate result. To overcome the limitation of less labelled kid's dataset we use Spectrogram Enhanced SimCLRV2 model since we can train the model with minimal available kid's dataset with which we can classify and predict kid's emotion effectively**.** Its goal is to maximize agreement across variously enhanced samples in the latent region of the input representation using contrastive loss. The quality of acquired emotional representations is greatly enhanced by the introduction of learnable nonlinear transformations between learnt emotional representations and contrastive losses. Multi-source transfer learning develops the network's capacity for accurate classification, finds the missing information in each source, and transfers it to the target data to complete it. As of our knowledge this is the first work that have utilized IMTL with SimCLR to improve target labels as well as overcomes less data. In order to train the network without labels using a self-supervised algorithm we use Ravdess song and Iemocap dataset. We use real-time kid's data as the teacher network. We employ Ravdess speech with labels as supervised algorithm. By incorporating these discoveries and analyzing those using SimCLR CSL methodologies in Zenodo children recording un-labeled dataset, which is the student network, we can dramatically outperform previous methods for semi-supervised learning in recognition rate of kid's emotion by training network with 1% real-time kid's data.

Keywords: *Tri-cut and Tri-mix augmentation, speech emotion recognition (SER), transfer learning, contrastive loss, contrastive learning, SimCLRV2, Incomplete Multi-source Transfer Learning*

1. Introduction

Unsupervised image representation learning-based research that uses contrastive learning has recently gained popularity [14, 15, and 16]. Speech emotion recognition (SER) is a technique that detects an individual's emotions from their speech signal. Contrast-based learning is used to extract representations of an input from a huge amount of un-labelled speech visual data in order to characterize children's emotions using semi-supervised learning. Given the lack of training data, this structure is utilized to complete tasks that call for a small number of labelled examples of children's speech. People may communicate with and express themselves to others thanks to a special human characteristic called speech. However, it can be quite challenging to determine a child's emotional state that can be anticipated by machines because they don't know how to communicate their feelings appropriately. Since most parents today work, the majority of kids are placed in a crèche or with an unknown person to be looked after. As a result, parents are frequently too far away to fully appreciate what is happening in their child's life. We require a strong emotion identification system for kids in order to get beyond the restrictions in training data and appropriately categorize children's emotions.

Contrastive learning is a method for structuring the work of locating similarities and differences for an ML model. This method can be used to successfully train a learning model to differentiate between similar and different emotional spectrogram. The contrastive predictive coding (CPC; [6]) and enhanced multi-scale deep information maximization (AMDIM; [13]) frameworks are the other frameworks that are based on contemporary CSL techniques. With the aim of maximizing similarity between representations derived from a single image while minimizing similarity between representations derived from other images, we use this (SimCLR) framework for contrastive method of visual learning in speech images.

Just a small amount of data is frequently provided for training reasons in situations like SER. As shown in this paper, a method known as knowledge distillation can effectively address the issue of limited training data. By utilizing an existing teacher's network that has less labelled data from various emotion classes, it resolves a classification prediction challenge. Here, when there is a limited training dataset available, we wish to recognize the emotions of children. The majority of papers use adult speech to train their networks in order to overcome this constraint. Here, it is overcome by applying the SimCLRV2

Department of Networking & Communications[1,2], pv9600@srmist.edu.in[1] , jesiv@srmist.edu.in[2]
School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai - 603203, india

model, which allows us to evaluate the student model to identify the emotion for children with great accuracy with only 1% real-time training data from children.

We can find helpful emotional representations by maximizing mutual information between features collected from various elements of speech. For voice, there exist continuous input feature representations. Observing a log-spectrogram of speech emotion from distinct masking could produce the log-spectrogram's numerous features. In order to understand what enables contrastive prediction tasks to acquire efficient representations, we examine the key elements of our system.

SimCLR is a recently proposed contrastive learning framework that is improved in three ways: (1) supervised/unsupervised pre-training, (2) supervised fine-tuning, and (3) distillation utilizing unlabeled data.

This study gives a complete analysis of the "unsupervised pre-train, supervised fine-tune" paradigm for semi-supervised learning on speech Image [21], which was inspired by recent developments in self-supervised learning of visual representations [16-20, 1]. Images are utilized without class labels (in a task-agnostic manner) during self-supervised pre-training, so the representations are not specifically customized to a given classification job. We discover that network size matters when using un-labelled data for this task-independent approach: A degradation issue arises because of vanishing gradients in the 56-layer plain network, while the 20-layer plain network had lower training and test errors [18]. In this work, the vanishing gradient problem is solved using Resnet, where 50 layer is the optimal number rather than higher layers. In addition to the size of the network, we describe a few key design decisions for contrastive representation learning that are advantageous for semi-supervised learning and supervised fine-tuning.

- Res-Net50 Models for CNN Encoder is used instead of 152-layer Res-Net, selective kernels
- Tri-cut and Tri-mix region Augmentation is used instead of basic augmentation in order to avoid loss of more information and performs effective predictive tasks
- Use two dense layers instead of one projection head.
- An incomplete multi-source transfer learning technique is used to extract useful patterns from a multi-source dataset for various emotions using self-supervised learning.
- The classification accuracy was greatly increased when the semi-supervised transfer learning method was tested on audio emotion classification datasets such the Ravdess song, Iemocap, and Zenodo children's speech recordings.

Large models are not necessary for learning visual representations as long as the job at concern is specific. Therefore, the model's efficacy in forecasting can be enhanced and applied to a smaller network with the task-specific use of tagged data.

We go on to illustrate the significance of the nonlinear transformation (also known as the projection head) employed in SimCLR after the convolutional layers for semi-supervised learning. When fine-tuning from a second layer of the projection head, a deeper projection head not only increases semi-supervised performance but also the representation quality as determined by linear evaluation.

We introduce the region augmentation technique [38], which enables you to increase the number of views, preserve the portion of a Speech image that contains the most information, and provide additional training data.

We combine these results to reach a brand-new benchmark for semi-supervised learning on the abridged Ravdess speech image. Enhance spectrogram SimCLRv2 achieves 85.58% top-1 accuracy under the supervised algorithm, a 25.5% relative improvement on the existing most recent task [36], while using a self-supervised algorithm, it achieves 85.55%, a 29.74% relative improvement over the prior state-of-the-art [36]. It achieves 89.7% top-1 accuracy when refined on just 1% of labelled cases and reduced to the same architecture using unlabelled data, which is a relative improvement of 25.95% over the prior state-of-the-art [36].

We improve self-supervised performance on Speech Image by 0.65% when compared to a conventional ResNet and outperform supervised Image pre-training by 0.62% by merging both technological contributions (self-supervised SimCLR and IMTL) into a single model.

2 Related Works

A straightforward method for contrastive analysis of visual information is developed utilising a base encoder network f (•) and a projection head g (•) that have been trained to optimise agreement using a contrastive loss. The model is evaluated using linear evaluation models (ResNet-50) trained with various batch sizes and epochs (Chen Ting et al., 2020). Images are encoded into a representation space using two contrastive learning optimisation methods, and pair wise affinities are calculated. SimCLR considerably enhances the end-to-end variance of instance discrimination in three ways: by producing more negative samples with larger batches (4k or 8k); by switching the output fc projection head for an MLP head; and by using greater data augmentation. (Chen, Xin lei, et al., 2020). It is demonstrated that the current variation lower bounds are insufficient. To address this problem, a set of lower bounds

that go beyond previous boundaries is formed. When the MI is high, it starts to drop, showing either a lot of bias or a lot of volatility. (Ben Poole et al., 2019).

BERT aims to pre-train in-depth bidirectional model from un-labelled text by configuring both left and right context at all levels. Modern models can be produced by fine-tuning the pre-trained BERT model with just one additional output layer (Devlin, Jacob, et al., 2018). By using self-supervised representation learning, information is captured to maximise the mutual information between features from these points of view that discuss high-level issues that affect many different points of view. Assessment utilising standard datasets: CIFAR10, CIFAR100, STL10 are done (Bachman, Philip, et al., 2019). Compact latent embedding spaces make it simpler to model conditional predictions. They use complex autoregressive models in this latent space to produce predictions. It makes use of noise-contrastive estimation. Together, the encoder and autoregressive model are trained to optimise the InfoNCE loss, which is based on NCE. (Aaron van den Oord et al., 2019). Pre-trained networks have had great success classifying images. Many open-source, highly effective image classification networks have been created. A pre-trained image classification network can be used to tackle the voice classification problem by rewriting it as an image classification problem, according to recent study. (Stolar et al., 2017).

Using transfer learning and pre-trained networks, a computationally efficient method that might be used to a small training data set. The SER issue had to be reframed as a spectrogram categorization problem since almost all of the pre-trained networks now in use were built for image classification and are used for speech recognition. Labelled voice clips were stored into brief time intervals in order to achieve this. The pre-trained CNN was given each block's individual spectral amplitude spectrogram array as input after it had been generated and transformed into an RGB image format. The experienced CNN was prepared to derive emotional categories from an unlabelled audio utilising the similar voice-to-image transformation procedure after only a brief training (fine-tuning) period. (Margaret Lech et al., 2020).

A transfer learning technique for processing variable-length input from speech recognition trained on vast volume of speaker-labelled data that makes use of an experienced residual (ResNet) model with a statistics-pooling layer. It uses a spectrogram augmentation technique to apply random time-frequency masks to log-mel spectrograms to produce extra training data samples. (Padi, Sarala, and others, 2021). A novel framework for deep model-based voice processing called nnAudio conducts temporal to spectrum domain translation utilising 1D convolutional neural networks. Its quick processing speed makes it possible to extract spectrograms instantly, doing away with the necessity to save them on a disc. This method also enables back propagation on the waveforms-to-image converted layer, making the conversion mechanism capable and enabling for additional optimisation for the particular application (Cheuk, Kin Wai, et al., 2020).

It is demonstrated using nnAudio, a GPU-based audio processing tool, how to compute Mel-spectrograms, Constant Q Transforms (CQT), linear-frequency spectrograms, and log-frequency spectrograms. One-dimensional (1D) convolution is used extensively in nnAudio for audio processing and transformations, allowing the transformation kernels to be integrated into larger neural networks and further trained/optimized for particular applications. Audio, Cheuk, K. W., Agres, Kat, et al., 2019. By enhancing and balancing the voice samples with data enhancement and dataset balancing, a novel method for SER (DCNN-BLSTMwA) is applied. We generate three-channel log Mel-spectrograms (static, delta, and delta-delta) from the input of the DCNN. Segment-level characteristics are then produced using the DCNN model, which has been previously trained on the ImageNet dataset. The utterance-level features are built up from these sentence-level features. The Bidirectional Long Short-Term Memory with Attention (BLSTMwA) model is then used to acquire high-level emotional characteristics that concentrate on emotionally significant variables for temporal summarization (Zhang, Hua, et al., 2021).

The learning effort in the focus region is facilitated by the use of incomplete numerous sources for successful information transfer. Cross-domain transfer from individual input to output and cross-source transfer are three methods for incomplete multisource transfer learning that have been proposed (Ding, Zhengming, et al., 2016

3 Method

Given that the most effective experienced neural model was trained on adult's speech labelled data to predict emotion for kid's spectrogram categorization. To employ these models to the issue of SER with less kid's emotion training data; the audio signal needs to be converted into a spectrogram [7]. Figure 1 shows spectrogram Visualization of kid's speech. Given that, most of the programmable voice interaction solutions use speech augmentation such as frequency masking and time masking on the real-time SER, here we additionally investigate Tri cut-mix augmentation.

Once a convolutional network is pre-trained and fine-tuned, we find that its task-specific or task-agnostic predictions can be further improved and distilled into a smaller network. To this end, we make use of unlabelled data for a second time to encourage the student model to mimic the teacher model's label predictions. Thus, the distillation phase of our method using unlabelled data is reminiscent of the use of

pseudo labels in self-training [19], but without extra complexity.

3.1 Methodology

We implement emotion categorization using a complete framework system built on the Resnet50 model. Many contemporary SER systems now use deep network models that acquire from spectral image or even initial waveforms to capture and express distinct emotions in speech, as opposed to traditional SER systems that relied on a high volume of minimal temporal- and frequency-domain variables [20]. Because of this, we develop and assess a Resnet-based approach using log-Mel spectrograms as input feature. Eight emotions, including joy, sadness, anger, surprise, neutral, disgust, fear, and calm, are included in our emotion model. We make use of the Ravdess speech dataset, which we split into train and test sets of emotion-based speech labels. Fig.1. shows the spectrogram visualization of kid's speech.

To train the model for various patterns, we use the unlabelled datasets from the Iemocap and the Ravdess song datasets. By dividing this contrastive learning strategy into its three main components—data augmentation, encoding, and loss minimization—we can analyse it [1]. By modifying SimCLRV2, which contains a teacher-student model for which 1% of real kids' speech data is used, we will expand on this strategy [21]. Using adult speech data, contrastive learning is carried out in both supervised and unsupervised ways.

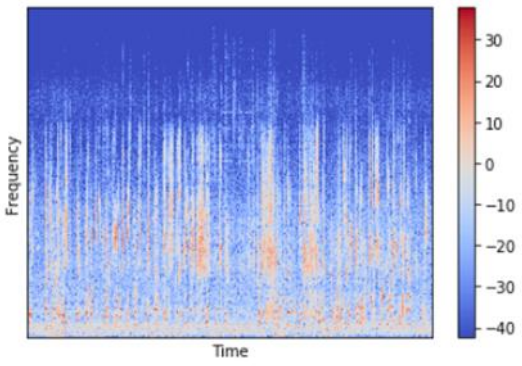

Fig. 1. spectrogram Visualization of speech of kid's data

Algorithm 1: Spectrogram Enhanced SimCLRv2

Input: SO1, SO2, SO3, SO4, DS1, DS2, DS3, DS4

1. for SO1:

    1a. train the model with labelled DS1 using supervised alg

    1b. calculate loss

    1c. Apply IMTL alg, calculate MD and CD to fix target labels

Repeat 1 until the loss minimizes

2. for SO2 & SO3:

    2a. train the model with un-labelled DS2 (DS3) using self-supervised alg

    2b. calculate loss

    2c. Apply IMTL alg, calculate MD and CD to fix target labels

Repeat 2 until the loss minimizes

4. Consider 1% real-time labelled kid's data as Teacher network

5. for SO4:

    5a. consider this as student network and impute labels for un-labelled DS4 using 4 and know ledged transferred from other sources using semi-supervised learning.

    5b. calculate loss

    5c. Apply IMTL alg, calculate MD and CD to fix target labels

Repeat 5 until the loss minimizes as well as target data obtained

End.

6

3.1.1 Spectrogram Augmentation

A cost-effective data augmentation technique called spectrogram image augmentation has recently shown promise for voice emotion identification job [30]. The spectrogram image augmentation approach uses random time-frequency masks on spectrograms to add more training data samples, preventing over-fitting and enhancing the generalization of voice recognition models. We use a loop to supplement the training data, utilizing Tri-cut for the first comparison which compares with time, frequency masking and Tri-mix that is performed jointly. In this way spectrogram image is sent to the network. The second fc result is used to compare results for similarity. For augmented versions of the same image, the estimated similarity must be small, and for a different image, it must be significant. The calculated loss must be minimal for the same image and large for different image

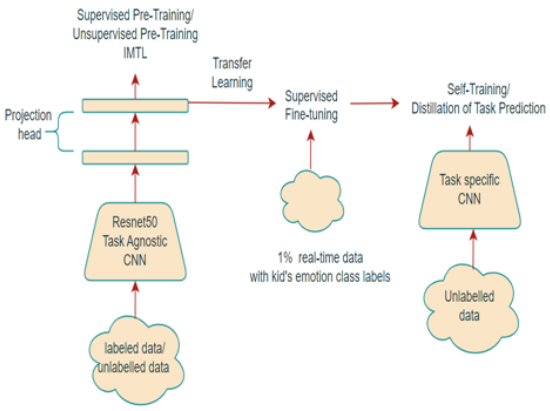

Fig. 2. The proposed semi-supervised learning framework for supervised/Unsupervised pre-trained network with Multisource Transfer Learning

3.1.2 Pre-trained Supervised and Unsupervised learning

We used supervised and self-supervised learning to extract usable representations from samples without the need for additional labels as a pre-training technique. The SimCLR framework [27] was used in our experiment as a supervised and self-supervised learning technique. Fig. 2. Shows the proposed semi-supervised framework.

Stochastic data augmentation module that, at random, splits each provided data example into two associated perspectives of the similar example, identified by the labels $x_{i,}$ and x_j, which we consider to be a positive pair. In this investigation, three straightforward augmentations are applied one after the other. Tri-Cut and Tri-Mix [38] are two policies that we look into for systematizing the use of spectrogram augmentation to SER. In order to obtain good performance, we combine image-level augmentation techniques like Time masking, frequency masking, and mixing of high information triangular parts (Tri-Mix) and a triangular region with little information (Tri-Cut). Since these two are effectively two distinct renditions of the same image, we want the model to learn that they are "similar." In my earlier paper, I looked into and employed this augmentation, and it produced good results when compared to other straightforward augmentation methods. Algorithm 1 summarizes the proposed work where SO refers to Source and DS refers to Dataset. To calculate MD and CD references is taken from [23].

Algorithm 1 summarizes the proposed work

The representations utilized in subsequent tasks are those produced by the encoder. We utilized transformer layers as the building blocks for the encoder section of Speech SimCLR, since transformer-based models have been demonstrated to be particularly effective for a variety of speech problems [34]. After the encoder, we used average pooling to transfer the encoder output to the area where contrastive loss would be performed. We employed a non-linear MLP with two hidden layers as the projection head in addition to SimCLR. With an MLP and two hidden layers, the projection head's output is specifically z_i = g (h_i) = $W^{(2)}(W^{(1)}\ h_i)$ where σ is a ReLU non-linearity. Our images are compressed into a latent space representation, which enables the model to learn the key attributes. In reality, we can envision that the model is learning clusters of related data points in the latent space as we continue to train it to maximize the vector similarity between similar images. Since this is what we are teaching the model to learn, for instance, happy representations will be closer together but farther apart than sad, angry, disgust, and fear representations. We use the cosine of the angle between the two vectors as a measure of similarity. We will get a high similarity when the angle is close to 0, and a low similarity otherwise, we also need a loss function that can be minimized. NT-Xent (Normalized Temperature-Scaled Cross-Entropy Loss) [35] is one choice.

$$\ell_{i,j} = -log \frac{\exp\left(\frac{sim(z_i,z_j)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp\left(\frac{sim(z_i,z_k)}{\tau}\right)} \quad \textbf{(1)}$$

3.1.2 Teacher-Student Model

Here, we directly use the un-labelled data for the final job so as to further develop the network for that task. We develop a student model utilising the fine-tuned network as a teacher, drawing inspiration from [23, 11, 22, 24, and 12]. Where no actual labels are used, we specifically reduce the following distillation loss:

$$\int^{distill} = -\sum_{x_{i,}\in D} \left[\sum_y P^T(y/x_{i\,};\mathfrak{c}) \log P^S(y/x_{i\,};\mathfrak{c})\right] \quad \textbf{- (2)}$$

P(y/x_i) = exp (f^{task}(x_i)[y]/ ɕ) / $\sum_{y'}$ **exp** (f^{task} (x_i)[y']/ ɕ) and ɾ is a scalar temperature parameter. Only the student network, which generates$P^S(y/x_i)$, is trained; the teacher network, which generates$P^T(y/x_i)$, is fixed during the distillation.

While we concentrate on distillation using solely Zenoda children's speech unlabelled instances in this study, one can also combine the distillation loss with ground-truth labelled examples using a weighted combination when the number of labelled examples is significant.

$\int$ = - (1 - α) $\sum_{x_i,y_i\in D} \log P^S(x_i/\ y_i)$ – α $\sum_{x_{i,}\in D}$ $\left[\sum_y P^T(y/\varkappa_{i\,};\mathfrak{c})\ \log P^S\ (y/\varkappa_{i\,};\mathfrak{c})\right]$ --------- (3)

Students with the same model architecture can be used in this technique (self-distillation), which enhances task-specific performance, or with a smaller model architecture, which results in a compact model.

3.2 Multi-source Transfer Learning

In the proposed study, the network has been pre-trained using the ravdess speech datasets as labelled datasets and the ravdess song and Iemocap dataset as an unlabelled dataset so that it can accurately categorize eight emotions using the supervised and self-supervised algorithm. Resnet50 was trained with a dataset for this pre-trained model using a supervised and self-supervised approach. The proposed semi-supervised learning framework, which contrasts supervised and unsupervised pre-trained networks, is shown in Figure 2. Basic image augmentation techniques like temporal and frequency masking are utilized in combination with tri-mix and tri-cut augmentation. We train the network with augmentation and simCLR technique to gain more information from spectrograms and generalize the network using the aforementioned data because some

words in the kid's speech dataset would not be obvious to identify emotion. Additionally, we use the Multi-source IMTL method [23]. To aid in the transfer of knowledge from each source, we combine a typical low rank transfer learning framework with an iterative structure term with a latent component. Latent variable would also help each source recover any missing categories. The underlying framework of the target data was preserved by two cross-source regularizes that were created to combine the strongly related samples of various input in both supervised and unsupervised modes. Marginal distribution (MD) and conditional distribution (CD) disparities were both reduced by two-direction transfer. Low rank reconstruction is used to solve the MD difference in order to reduce the gap between two domains in the converted space.

In the new environment, every target instance would be relatively to the similar class as the input instance, thus we minimize the MD difference. We employ iterative structure learning for the CD, which promotes intended information to only be associated with source data from the same class. The CD difference would therefore be lessened.

The student network is then trained using 1% real-time kid's speech tagged data that has been fine-tuned. Finally, utilizing supervised and unsupervised pre-trained models, we can improve accuracy through the patterns learned for children's emotion classification and prediction tasks. Here, we identify and forecast the emotions using recordings of Zenoda children's speech.

4. Experiment Analysis

In this part, we compare SimCLR and multi-source transfer learning to other approaches in order to assess their effectiveness on common datasets. The standard dataset is described in Section 4.1,data labelling process is explained in 4.2, the experimental setup is explained in detail in Section 4.3, the datasets used are evaluated and presents a comparison of the SimCLR method against baseline in Section 4.4, the Multi-source transfer learning process is discussed in Section 4.5 and the student network and teacher network are fine-tuned in Section 4.6.

4.1 Data

In this study, we use a model that was originally created for emotion recognition as a characteristic predictor. More precise, Resnet50 is trained by using a sizable amount of emotion-labelled Ravdess speech data. The FC structure of the already trained model are then swapped out with brand-new FC layers with random initialization. Finally, we retrain the new FC layers for a SER using multi-source transfer learning using the Ravdess music dataset. The Ryerson Audio-Visual Database of Song (1012 songs) and Emotional Speech (1440 files) 24 professional actors—12 men and 12 women—are expressing two phrases that are lexically similar in a neutral North American accent in the database. The dataset includes emotions like as happy, sad, angry, surprise, neutral, disgust, fear, and calm

The downstream job for voice emotion recognition uses the IEMOCAP database [27], which provides labels for the target data that are not full. We chose the recordings from which the majority of annotators agreed on the labels for the four different emotion types—angry, happy, sad, and neutral. So as to control the number of instances in each emotion category, happy and exciting feelings were mixed as happy. The dataset consists of 5,531 utterances divided into 5 sessions (1,103 angry, 1,636 happy, 1,708 neutral, and 1,084 sad). On this dataset, 5-fold cross validation was done.

We collected Real-time voice data of children reading stories and interacting with each other at ages 2(M) and 5(F) through microphone, which represents 1% of the labelled data used for the supervised fine-tuned teacher network.

For fine-tuning, ZENODO children speech recording unlabelled dataset is used. The collection has 672 files and includes audio recordings (lossless WAV) of 11 young children (age M=4.9 years; 5 girls, 6 males).

- Free speech (retelling the children's tale "Frog, Where Are You?" by Mercer Mayer) is one of the recordings.
- Repetition of five short, pre-defined statements (such as "The horse is in the stable").
- Reciting the numbers 1 through 10

Both native English speakers and non-native speakers can be heard on the recordings.

Three sources are used to create each sample:

• The two front microphones of the Aldebaran NAO robot; a studio-quality microphone (Rode NT1-A); a portable microphone (Zoom H1).

4.2 Data Labeling of Real-time kid's data

In order to study emotion, the emotional behavior of children has been recorded in databases. These are natural databases made up of recordings made of kid's interaction with

sibling. These databases recorded through microphones, telling stories, and children fighting. The children are 2 and 5 years old, Male and Female respectively. The data labelling is determined by the listener's perception. Eight emotions—happy, sad, angry, surprised, neutral, disgusted, fearful, and calm—are used to underpin the judgement. 5 cross validation was done to confirm the exact labelling

4.3. Experiments Settings

The baseline network was set to ResNet-50 [30], and the same parameters from a prior study [28, 31] were employed. It was maximized by the Adam algorithm with step size and

regularization of 1e-4 and 1e-3, respectively. The networks were also trained for 80 epochs on an NVIDIA TITAN-Xp GPU with a batch size of 34. ResNet-50 pre-trained with via self-supervised [28] approaches were used. We further verified the outcomes by 5-fold cross validation using the pre-existing training/validation folds for three benchmarks, namely RAVDESS, IEMOCAP, and ZENODO children speech.

4.4. Evaluation Results on RAVDESS, IEMOCAP, and ZENODO children speech

The performance of semi-supervised learning is improved by well-learned features from speech images [31] Table 1 makes clear that the pre-trained networks outperformed the earlier efforts [28, 31] in terms of classification performance as seen in Fig. 3. In prior research, however, employed supervised transfer learning, which gave the model a head starts by pre-training them on a sizable dataset of images with matching labels. Our self-supervised transfer learning technique, in contrast to supervised learning, delivers a comparable level of accuracy without the need for a separate label for pre-training. The accuracy of the supervised and unsupervised methods for UrbanSound8K was somewhat inferior to that of the supervised and self-supervised methods using the Ravdess speech and song dataset where a drop in average accuracy of 2.24%p and 2.25%p, respectively was discovered.

Additionally, the pre-trained network's loss convergence occurred faster in almost 20 epochs. In order to transfer learning from Ravdess speech and song images to emotion recognition in children, both pre-trained networks displayed same characteristics in their training loss, which indicates similarity between the two pre-training mechanisms (Fig. 4). Both pre-trained networks in these tests used different volumes of instances (Ravdess speech and song). The creation of large pre-training image datasets is constrained by the labor-intensive process of manually annotating each image's multiple labels [29, 1].

Self-supervised learning, on the other hand, doesn't need labels and can make use of big image datasets [1]. The recognition rate of un-supervised pre-trained model has achieved 85.5% over previous works [1, 28, and 36] showing less effectiveness on spectrogram categorization application than the one trained with supervised learning methods. Our findings also showed that employing semi-supervised pre-training, transfer learning between different sources and domains might perform better than previous self-supervised pre-training schemes. Therefore, transferring knowledge from speech images to emotion recognition in children can be done simpler and effectively with pre-training using enhanced simclrv2 semi-supervised learning.

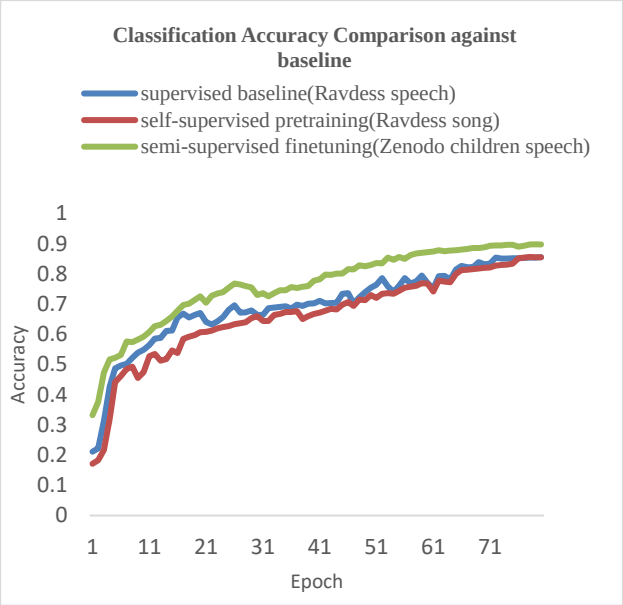

Fig. 3. shows semi-supervised classification accuracy comparison with supervised and un-supervised learning

4.5 multi-source and multi-domain training improves model knowledge

Studies using three different data sets have demonstrated that IMTL is superior to other approaches in addressing the issue of incomplete multiple sources [28]. Unlike pre-trained supervised and unsupervised transfer learning, IMTL un-supervised scheme delivers a greater standard of accuracy with target labels. For IEMOCAP database, target labels were missing. The target labels were achieved as well as the accuracy of the unsupervised method has improved slightly than that of the supervised and self-supervised method pre-training using Ravdess speech and song dataset by 0.62%p and 0.65%p respectively.

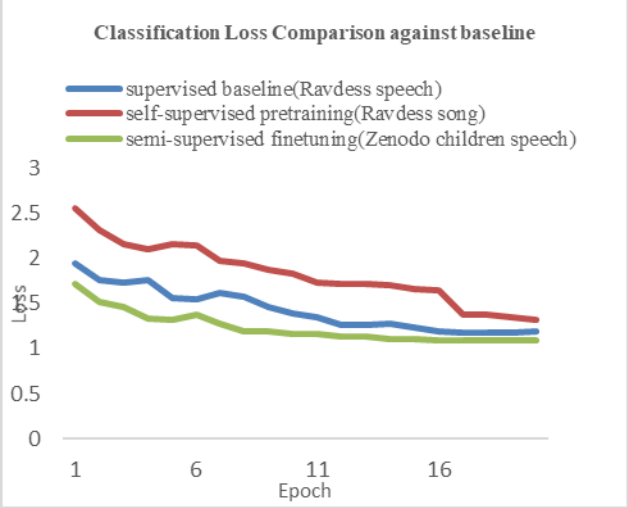

Fig. 4. shows semi-supervised classification loss comparison with supervised and un-supervised learning

4.6 Fine-tuning with teacher network

Distillation with unlabeled data boosts the performance of fine-tuned models in two ways: (1) when the student model has a smaller architecture than the teacher model, it increases model efficiency by transferring task-specific knowledge to a student model, and (2) even when the student model has the same architecture as the teacher model (with the exception of the projection head after ResNet encoder) [32].

Table 1: classification comparison of semi-supervised learning for Enhanced simclrv2 approach with previous works

Model	Dataset	Accuracy
Pre-trained (supervised)[28]	UrbanSound8K	83.34
Pre-trained (Unsupervised)[28]	UrbanSound8K	83.3
fine-tuned semi-supervised [28]	UrbanSound8K	
Pre-trained (supervised)[36]	STL-10	60.08
Pre-trained (Unsupervised)[36]	STL-10	55.81
fine-tuned semi-supervised [36]	STL-10	63.75
Pre-trained supervised(ours)	Ravdess speech	85.58
Pre-trained Unsupervised (ours)	Ravdess song	85.55
Pre-trained Unsupervised (ours)	IEMOCAP database	86.2
fine-tuned semi-supervised (ours)	Zenodo children speech recording	89.7

Table 2: predicted labels of semi-supervised learning for Enhanced simclrv2 approach

S. No	Actual Label	Predicted Label
1	happy	happy
2	Fear	Sad
3	Angry	Angry
4	Surprised	Surprised
5	Sad	Sad

With the un-labelled ZENODO children speech recording dataset, we achieve the greatest performance for smaller ResNet model. We contrast our semi-supervised learning outcomes in Table 1. with those of our work with supervised and unsupervised learning as well as with the latest SOTA semi-supervised learning techniques [28] on UrbanSound8K. We can show that for small ResNet that after fine-tuning the recognition results have dramatically improved than pre-training supervised and unsupervised learning, our approach significantly outperforms previous results. We may increase the recognition rate and the number of missing labels by using 1% real-time children's speech data as a teacher network and the IMTL method. The predicted labels as seen in Table 2. shows effective prediction where most of the labels are predicted correctly

5 Conclusions

We investigated the semi-supervised transfer learning from spectrogram. Effective prediction tasks are defined by the composition of data augmentations. We pre-trained the CNN network using self-supervised learning techniques that teach the relationship among visual representation without labels, and then we fine-tuned it using the target voice mel-spectrograms to extract the well-tuned features. The CNN trained with spectrogram using an un-supervised method has more or less equal performance as supervised schemes at a similar level without the use of labels. Furthermore, both fine-tuned pre-trained systems beat supervised and self-supervised networks by a significant margin.

Consequently, our research has been applied to children's emotion recognition tasks in the audio image domain to obtain notable performance increases. The information shared from individual instance utilizing the IMTL framework was facilitated by the inclusion of an iterative structural term with a latent variable, which helped to reduce the problems of missing categories, MD, and CD has been mitigated effectively. This framework has been combined with the Simclr framework to efficiently recognize children's emotions; a difficult task due to a lack of data.

Acknowledgements

We thank our colleagues who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. We thank our Supervisor and Reviewers for assistance and for comments that greatly improved the manuscript.

Author contributions

Preethi V: Conceptualization, Methodology, Software, Field study, Visualization, **V Elizabeth Jesi** Investigation, Writing-Reviewing and Editing, Validation.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," CoRR, vol. abs/2002.05709, 2020.

[2] Chen, Xinlei, et al. "Improved baselines with momentum contrastive learning," arXiv preprint arXiv:2003.04297 (2020).

[3] Poole, Ben, et al. "On variational bounds of mutual information," International Conference on Machine Learning. PMLR, 2019.

[4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805* (2018).

[5] Bachman, Philip, R. Devon Hjelm, and William Buchwalter. "Learning representations by maximizing mutual information across views," *arXiv preprint arXiv:1906.00910* (2019).

[6] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748* (2018).

[7] Stolar, Melissa N., et al. "Real time speech emotion recognition using RGB image classification and transfer learning," *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*. IEEE, 2017.

[8] Lech, Margaret, et al. "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers in Computer Science* 2 (2020): 14.

[9] Padi, Sarala, et al. "Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation," *arXiv preprint arXiv:2108.02510* (2021).

[10] Cheuk, Kin Wai, et al. "nnaudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks," IEEE Access 8 (2020): 161981-162003.

[11] Cheuk, K. W., Kat Agres, and D. Herremans. "nnaudio: A pytorch audio processing tool using 1D convolution neural networks," ISMIR–Late breaking demo (2019).

[12] Zhang, Hua, et al. "Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition," Frontiers in Physiology 12 (2021).

[13] Falcon, William, and Kyunghyun Cho. "A framework for contrastive self-supervised learning and designing a new approach," arXiv preprint arXiv:2009.00104 (2020).

[14] Tian, Yonglong, Dilip Krishnan, and Phillip Isola. "Contrastive multiview coding," Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer International Publishing, 2020.

[15] Wu, Zhirong, et al. "Unsupervised feature learning via non-parametric instance discrimination," Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[16] Ye, Mang, et al. "Unsupervised embedding learning via invariant and spreading instance feature," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[17] Guan, Qing, Yunjun Wang, Bo Ping, Duanshu Li, Jiajun Du, Yu Qin, Hongtao Lu, Xiaochun Wan, and Jun Xiang. "Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study," Journal of Cancer 10, no. 20 (2019): 4876.

[18] Harzallah, H., Jurie, F., & Schmid, C. (2009, September). Combining efficient object localization and image classification. In 2009 IEEE 12th international conference on computer vision (pp. 237-244). IEEE.

[19] Anirudh Shenoy "Pseudo-Labeling to deal with small datasets," Published in Towards Data Science, 2019.

[20] [20] Tzirakis, Panagiotis, Jiehao Zhang, and Bjorn W. Schuller. "End-to-end speech emotion recognition using deep neural networks," In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5089-5093. IEEE, 2018.

[21] Chen, Ting, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. "Big self-supervised models are strong semi-supervised learners," Advances in neural information processing systems 33 (2020): 22243-22255.

[22] Schneider, Steffen, et al. "wav2vec: Unsupervised pre-training for speech recognition," arXiv preprint arXiv:1904.05862 (2019).

[23] Ding, Zhengming, Ming Shao, and Yun Fu. "Incomplete multisource transfer learning," IEEE transactions on neural networks and learning systems 29.2 (2016): 310-323.

[24] Sajjad, Muhammad, and Soonil Kwon. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," IEEE access 8 (2020): 79861-79875.

[25] Alzubaidi, Laith "Novel transfer learning approach for medical imaging with limited labeled data," Cancers,13,7,1590, MDPI. (2020).

[26] Dataset - https://datasetsearch.research.google.com/

[27] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," Lang. Resour. Evaluation, vol. 42, 2008.

[28] Shin, Sungho, Jongwon Kim, Yeonguk Yu, Seongju Lee, and Kyoobin Lee. "Self-supervised transfer learning from natural images for sound classification," Applied Sciences 11, no. 7 (2021): 3043.
[29] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge," International journal of computer vision 115 (2015): 211-252.
[30] He, Kaiming, 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.
[31] Palanisamy, Kamalesh, Dipika Singhania, and Angela Yao. "Rethinking CNN models for audio classification," arXiv preprint arXiv:2007.11154 (2020).
[32] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, “Unsupervised visual domain adaptation using subspace alignment,” in Proc. IEEE Int.Conf. Comput. Vis., Dec. 2013, pp. 2960–2967.
[33] Gat, Itai, Hagai Aronowitz, Weizhong Zhu, Edmilson Morais, and Ron Hoory. "Speaker normalization for self-supervised speech emotion recognition," In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7342-7346. IEEE, 2022.
[34] Karita, Shigeki, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki et al. "A comparative study on transformer vs rnn in speech applications." In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 449-456. IEEE, 2019.
[35] Jiang, Dongwei, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li. "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning." arXiv preprint arXiv:2010.13991 (2020).
[36] András Béres “Semi-supervised image classification using contrastive pretraining with SimCLR,” Keras.io. https://keras.io/examples/vision/semisupervised_simclr/
[37] Dataset- https://www.researchgate.net/post/Anyone-know-of-a-free-download-of-an-emotional-speech database/5e62f7d1f8ea52d5cd35f0fc/citation/download.
[38] Preethi, V., & Jesi, V. E. (2024). Triangular Region Cut-Mix Augmentation Algorithm based Speech Emotion Recognition system with Transfer Learning Approach. IEEE Access.