Hindawi
Scientific Programming
Volume 2021, Article ID 1093698, 7 pages
https://doi.org/10.1155/2021/1093698

Research Article

Application of Transfer Learning Algorithm and Real Time Speech Detection in Music Education Platform

Hexue Shen

College of Music, Chongqing Arts and Sciences University, Chongqing 440000, China

Correspondence should be addressed to Hexue Shen; rslavova@wisc.edu

Received 19 July 2021; Revised 6 September 2021; Accepted 22 September 2021; Published 11 October 2021

Academic Editor: Mian Ahmad Jan

Artificial intelligence (AI), particularly machine learning (ML) and neural networks (NN), has various applications and has sparked a lot of interest in the recent years due to its superior performance in a variety of tasks. Automatic speech recognition (ASR) is a technique that is becoming more important with the passage of time and is being used in our daily lives. Speech recognition is an important application of ML and NN, which is the auditory system of machines that realize the communication between humans and machines. In general, speech recognition methods are divided into three types, i.e., based on the channel model and speech knowledge method, template matching scheme, and the use of NN method. The main problem associated with the existing speech recognition methods is the low recognition accuracy and more computation time. In order to overcome the problem of low recognition accuracy of existing speech recognition techniques, a speech recognition technology based on the combination of deep convolution neural network (DCNN) algorithm and transfer learning techniques, i.e., VGG-16, is proposed in this study. Due to the limited application range of DCNN, when the input and output parameters are changed, it is necessary to reconstruct the model that leads to a long training time of the architecture. Therefore, the migration learning method is conducive to reducing the size of the dataset. Various experiments have been performed using different dataset constructs. The simulation results show that transfer learning is not only suitable for the comparison between the source dataset and the target dataset, but also suitable for two different datasets. The application of small datasets not only reduces the time and cost of dataset generation, but also reduces the training time and the requirement of computing power. From the experimental results, it is quite obvious that the proposed system performed better than the existing speech recognition methods, and its performance is superior in terms of recognition accuracy than the other approaches.

1. Introduction

Hearing loss affects an estimated 360–362 million individuals worldwide [1]. These figures are anticipated to grow by 40% on average by the year 2035. Hearing loss is normally due to two factors, i.e., age and noise. Hearing loss caused by both of these factors, i.e., aging or noise, is gradual and neither treatable, nor reversible. People with serious hearing problems are frequently socially isolated, which can lead to despair and a variety of other harmful outcomes. The most often used technologies for mitigating hearing loss are hearing aids and cochlear implants. Even advanced listening devices, on the other hand, create significant issues for the hearing impaired people, as they typically improve speech audibility but do not necessarily restore intelligibility in loud social circumstances [2]. Humans have been observed using the audio-visual aspect of speech to contextually reduce background noise and concentrate on the target speech in such situations. Furthermore, it is widely recognized that visual information aids in the resolution of acoustical ambiguities. For instance, people use AV cues in speech recognition, to properly interpret the conversation. The McGurk effect [3] shows that most people interpret a visual "ga" with a spoken "ba" as "da." The visual signals, particularly, give information on the location of articulation [4] and muscular movements, which can help distinguish between speeches with identical acoustic sounds.

Speech recognition is the auditory system of machines, which can realize the communication between humans and machines. In general, speech recognition methods are

divided into three types, i.e., based on the channel model and speech knowledge method, template matching scheme, and the use of artificial neural network method [5, 6]. Compared with the traditional speech recognition methods, the artificial neural network (ANN) has a great improvement in modeling ability and speech recognition accuracy. The concept of deep learning (DL) is originated from the neural network system of the human beings [7]. In 2009, DL was first applied to speech recognition task [8]. According to the current development of speech recognition technology, speech recognition algorithms based on DL are mainly divided into long short-term memory (LSTM) network [9], deep neural network (DNN) [8], and convolutional neural network (CNN) [10]. CNN can obtain better robustness by using local filtering and maximum pooling technology. Therefore, CNN has received extensive attention in the field of image, video, and speech recognition in recent years [11, 12]. In recent studies, CNN has been applied in the field of speech recognition and has obtained promising results in terms of accuracy. Compared with previous work, the biggest difference is the use of deep convolution neural network (DCNN) [13, 14]. In speech recognition, there are differences in each person's pronunciation, which can be effectively removed by using DCNN and improves the accuracy of speech recognition systems [15, 16]. DCNN performs better on a large dataset, while, on small dataset, it fails to give better recognition results due to the problem of overfitting. DCNN needs a large dataset in order to prevent the problem of overfitting, so the training and selection of the DCNN architecture are very time-consuming, but an important step. At present, the method used to reduce the dataset size composed of images is the migration learning technique, in which the model architecture is trained on a large database and then tested on a smaller dataset, which is known as target database. The object recognition ability of transfer learning techniques can be found in different studies. Supported by transfer learning, several methods are used for visual identification and are widely used in image classification [17] and medical field [18, 19].

The goal of this study is to demonstrate the breadth of transfer learning's applications and performance on a heterogeneous and sparse database. The initial goal of the research work is to demonstrate that transfer learning is appropriate not just for situations, in which both the source and target database are the same, but also for situations when both databases are quite different. As a result, instead of pictures, a pretrained DCNN is used to learn verbal alphabets from A to Z. The AVICAR dataset has been utilized as a database, and it comprises verbal alphabets from fifty female and male speakers in various driving circumstances [20]. Each letter's audio recordings are converted into a spectrogram using the Fourier transform as part of the preprocessing procedure. The DCNN is trained using the produced pictures of the spectrograms, which are assembled by alphabets. As a result, there are 26 classes in the multiclass categorization problem. Furthermore, the dataset is kept short in order to test the efficiency of a pretrained network model using only the sparse database. The VGG-16 is selected as a pretrained DCNN [21]. It is a 16-layer Convolutional Neural Network developed by the University of Stanford's Visual Geometry Group and pretrained on data from ImageNet, a huge visual dataset for object identification. It is one of the most advanced DNN architectures that have been submitted to the ImageNet problem in recent years. In this article, you will find an overview of DNN models as well as a study of their computational needs, power consumption, and inference time [22].

This study uses DL and transfer learning techniques for speech recognition. The primary contributions of this study are given as follows:

(i) This paper proposes a method by combining DCNN algorithm with transfer learning to realize the speech recognition.

(ii) The use of DCNN algorithm improves the accuracy of speech recognition significantly, and the use of transfer learning technique, i.e., VGG-16, reduces the size of the dataset and helps in increasing the recognition accuracy.

(iii) The simulation results show that the transfer learning method not only reduces the time and cost of dataset generation, but also greatly saves and reduces the training time.

The remainder of this paper is organized as follows: Section 2shows the related work section, Section 3 illustrates the proposed methodology, Section 4 demonstrates the experimental setup and results analysis, while Section 5 concludes the research work.

2. Related Work

Automatic speech recognition (ASR) has evolved into a technology that is increasingly used in our daily lives. Word recognition performance has been proven to reach 100% in noise-free situations [23], but in noisy environments, performance decreases quickly. Model normalization, reliable feature extraction, and classification algorithm, as well as speech augmentation approaches, are some of the strategies that may be used to make ASR more resilient under these situations. Speech augmentation is a common method for this since it requires little to no prior information of the surroundings in order to successfully decrease noise in the voice signal and, as a result, increase identification accuracy. By focusing solely on the acoustic channel, each of these approaches aims to increase the quality of the ASR system. Using visual characteristics taken from the visual movements of a speaker's mouth region in combination with the auditory channel to increase noise resilience has been tried with moderate success. A substantial amount of research has been done in the subject of audio-visual ASR (AVASR) [24]. The comparison of these two dissimilar techniques (speech improvement vs. visual information fusion) to increase the noise resilience of speech recognition in unfavorable settings is of great interest. Due to the paucity of data that can allow such assessments, it has been difficult to determine whether acoustic speech augmentation or visual fusion is preferable, or whether both techniques can be coupled to further boost

resilience in noisy situations. So far, most AVASR research has focused on increasing the visual information quality [24], with the implicit assumption that the use of visual information in the ASR system will enhance its resilience in the presence of noise.

Transfer learning is a popular method for decreasing the size of an image database. In this context, the saved weights and an existing architecture from a comparable problem are used to help learning on a novel developing challenge. Transfer learning involves training the model architecture on a massive training database afore transferring it to a new and considerably limited target database. A research study on the use of transfer learning techniques in different fields was conducted in [25], which aims to show the importance of transfer learning in various applications. Using transfer learning techniques, there are several ways to accomplish the task of visual recognition.

The main objective of this research is to investigate the applications and performance of transfer learning on a diverse and sparse database. The primary objective of the research study is to investigate that transfer learning is suitable not only when both the training and testing dataset are the same, but also when the two datasets are quite dissimilar. This study proposes a speech recognition system based on the combination of DCNN and transfer learning techniques. It is among the most advanced DNN architectures that have been submitted to the ImageNet problem in recent years. In this article, you will find an overview of DNN models as well as a study of their computational needs, power consumption, and inference time.

3. Proposed Methodology

This section of the paper describes the proposed methodology for carrying out the research study. The proposed methodology consists of different steps starting from dataset collection, data preprocessing, and the use of transfer learning techniques along with the deep convolutional neural network.

3.1. Dataset Collection and Preprocessing. AVICAR dataset comes from audio-visual speech corpus assembled from a car using various sensor devices. The University of Illinois scholars acquired and recorded it in 2004. The data was collected using eight microphones installed on the solar shield and four camcorders placed on the dashboard. Separated letters, contact numbers, separated digits, and phrases were the four types of speech that were gathered. All of the classes are captured in English from both 50 female and male speakers each, under five distinct driving circumstances, each with closed and open windows and idling, at the speed of 35 and 55 mph. The data is open to the public and is available without any cost. The audio data of isolated letters is collected in this study in all five driving situations for additional analysis. For every alphabet from A to Z, 200 audio files are selected for training and 50 audio files are selected for testing, respectively. For each letter, audio recordings of both female and male voices were saved in separate places. A total of 13000 audio files are generated, in which 10400 are selected for training, while the rest of the 2600 are used to test the model.

A spectrogram is created for each audio file. The frequency band of a voice is shown by the spectrogram of an audio recording. This technique is used in music, acoustics, radio, and speech identification. The Fourier transform is utilized to produce spectrum from voice files in this study. Figure 1 illustrates the spectrograms of the alphabets from A to D.

As the dataset is sparse, so the data augmentation is applied. Using label-preserving transformations, data augmentation approaches achieve the goal of artificially enlarging the dataset. Because of the existence of sparse datasets, data expansion is carried out, and label preserving transformation is used to realize the manual expansion of datasets. In order to enlarge the data, there is no need to generate new images, and the existing dataset is slightly modified, using different augmentation techniques like flipping, rotation, and translation. When these images are given to the neural networks, they consider it as distinct images. In this paper, different enhancement adjustments are tested, and the best results are obtained through random rotation and random width movement. The converted image is generated from the original image, which is produced on the CPU during the training of the last batch that is not required to be stored.

3.2. Transfer Learning Techniques. Training a DCNN on a short dataset, even when augmented with data, as demonstrated in many research papers [10, 11], does not yield sufficient results, ss discussed earlier that training the DCNN on small datasets gives unsatisfactory results. The results produced by this approach are different from the theory, so transfer learning techniques are used to solve this problem. In addition to the pretrained weights, there are different architectures that can be freely used for identification, fine-tuning, and feature-extraction. In this paper, the VGG-16 model is used for further work, because compared with other available models, the testing accuracy results of the VGG-16 model are much better than those of the other models. Figure 2 shows the VGG-16 architecture, which starts with an input image of size $244 \times 244 \times 3$ and then adds convolutional layer having 3×3 field size, a stride size of 1, as well as 5 Max-pooling layers having 2×2 window size. Next, 3 fully connected layers are used, and finally softmax is used as the activation function at last. For all the hidden layers, the rectified linear activation unit (ReLU) and activation function are used. This architecture is trained on ImageNet dataset. The ImageNet is an image data set, composed of the above 14M images. These images are manually categorized to identify the objects in the image dataset. A subset of over 1 million pictures is utilized to train the VGG-16 model, and the images are categorized into 1000 item categories to generate a wide range of image feature representations. Pretraining has the benefit of recognizing the correlations between objects and creating a structure, and forming classifications on a large and varying dataset that may be

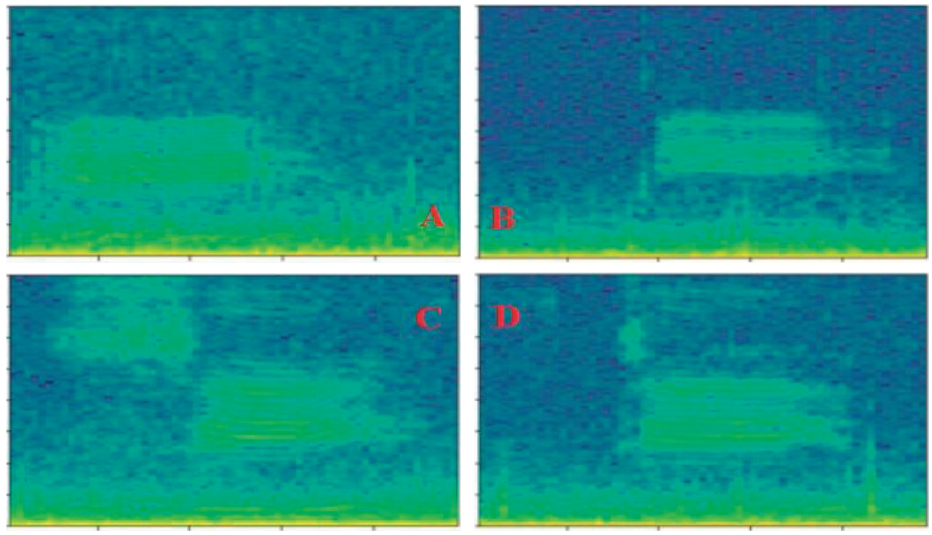

Figure 1: Spectrogram transformation of the alphabets from A to D using Fourier transformation.

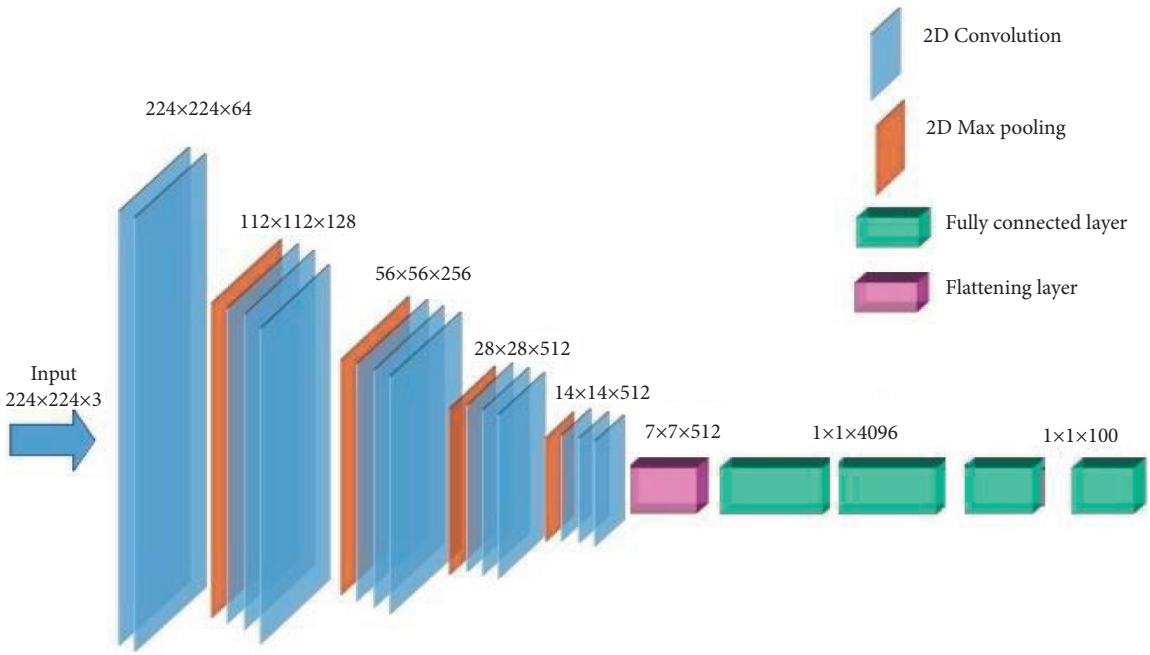

Figure 2: Schematic diagram of the VGG-16 transfer learning technique.

adjusted and gathered into new tasks to complete the redesigned tasks. Literally, it shifts learning progress to the current topic.

When utilizing limited datasets, the benefits of transfer learning for visual identification are widely documented. For small datasets and visual recognition, transfer learning is very suitable, especially in medical image analysis, in which usually a very small dataset is used, and DCNN is the preferred method of analysis of medical images. The pre-training of large data sets extracts useful features from the data, applies those features to the subsequently given tasks of small datasets, and proposes improvements in learning the sparse data.

4. Experimental Setup and Results Analysis

The experimental setup and simulation results are discussed in this section. The experimental setup includes a laptop system having the specifications of: core *i*7, 7th generation, RAM of 16 GB, Hard disk of 500 GB, 128 GB of SSD, and 2.7 GHz processor operating on Windows 10. The IDE used for carrying out the simulations is Spider and *Python* is the language used for the implementation of the algorithms. For the implementation of AVICAR dataset Keras deep learning (DL) framework is used, which uses TensorFlow at the backend. In addition to various pretrained DCNN models, Keras is also used to implement the VGG-16 model as

described earlier. The model comes with weights that have been pretrained and may be used for forecasting, extraction of features, and fine-tuning. In this method, fine-tuning is used to develop the proposed model. As shown in Figure 3, the pretrained VGG-16 model is altered by mangling the final fully connected layer prior to the last maximum pooling layer, expanding the global spatial average pooling layer (GAP) and two fully connected layers. GAP decreases the number of model parameters, eases the spatial dimension, and ensures that the model will not overfit. After that, a fully connected layer with a size of 11512, a linear predictor (ReLu), and a subsequent fully connected softmax layer with 26 classes follow the GAP. This is the number of classes required for the experimentation.

The upper layers of the VGG-16 architecture are trained using spectrograms produced by the data augmentation techniques, which improves the performance by artificially expanding the size of the data. The stochastic gradient descent optimizer (SGD) having a modest learning rate of 0.0004 is selected as an optimizer. The model is trained with a batch size of 8 for 25, 50, 100, and 200 epochs using audio data of both female and male voices separately, and a combination of both female and male voices. The pretraining class hours should approach to or within the range of training capacity. Table 1 shows the training results of accuracy in percentage.

From Table 1, it is quite obvious that a data set is composed of 5200 male or female voice files, and the voice test results of each gender can be received in an effective way. Training cannot produce similar results for the mix of male and female voices using a dataset of 5200 files. However, a dataset that contains double files, i.e., 10400 files, not only achieves the results of individual training, but it is even better than the results of individual training. It can be observed that, for the total test cases, the results that cannot be completed after more than 25 class hours of training can be obtained through 50 or more class hours of training.

Fine-tuning can be done in the second phase. Only the upper layers are used for training, while the lower layers are kept constant. The outcomes of the proposed design are investigated by altering the percentage fraction of frozen layers from 10%–90% in this study. Further, the outcomes of the training of female and male voices are examined independently, as well as with a dataset that includes both female and male voices. There are 5200 files in total across all datasets. Since a larger learning rate might cause misrepresentation of the pretrained weights, which are supposed to reflect excellent outcomes for the new model, a SGD optimizer with a modest learning rate of 0.0002 is employed for fine-tuning.

Table 2 illustrates the speech recognition accuracy after applying the fine tuning of the frozen layer.

It can be seen from Table 2 that female and male voices are tested individually as well as in combination. The results of overall accuracy are comparable with those of pretraining for female and male voices trained individually and in combination, and the accuracy of the combined male and female voices is lower as compared to the separated one. It can be seen that, under all experimental conditions, the best

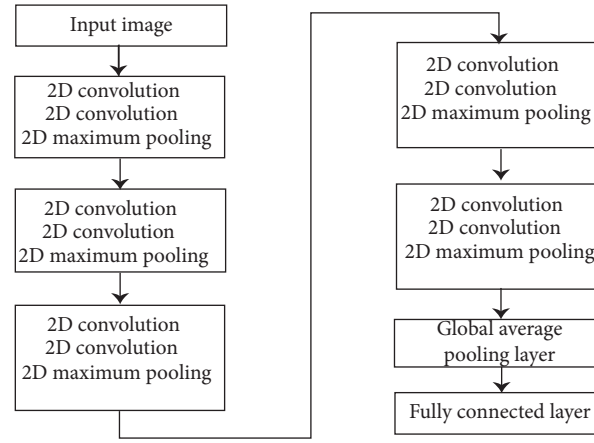

Figure 3: Improved VGG-16 model.

Table 1: Speech recognition accuracy under different training hours.

Research objects	Training hours			
	25	50	100	200
Female sex	29.30	36.19	40.60	41.30
Male	30.42	35.89	41.32	42.18
Female/male	22.24	29.32	29.80	30.87
Female/male (double)	37.82	43.21	45.26	46.32

Table 2: Speech recognition accuracy after applying the fine-tuning of frozen layer.

Research objects	Class hours	Percentage of frozen fine adjustment layer				
		10	30	50	70	90
Female/Male	25	64.29	61.29	61.52	54.30	23.60
	50	63.32	62.94	59.31	54.10	29.11
	100	63.89	63.62	61.36	53.29	34.30
	200	62.22	61.42	57.10	53.80	35.64
Female sex	25	76.25	77.20	74.61	65.14	35.58
	50	71.96	76.56	72.58	65.61	41.20
	100	76.62	76.36	74.25	65.72	39.76
	200	74.90	75.28	71.58	66.40	46.50
Male	25	73.10	78.66	76.79	71.24	25.10
	50	77.32	76.74	74.88	65.70	39.78
	100	77.80	76.87	74.71	66.68	43.86
	200	75.32	76.02	71.66	65.56	39.50

results can be obtained when the percentage of frozen layer is 10%–50%. When 90% of the layers are frozen, there is no training result at all. When the pretraining dataset is unrelated to the real dataset, freezing most of the layers and just training the final remaining layers is pointless, because feature fitness is insufficient, which may be increased by reducing frozen layers until a specific point.

Because of the inadequate findings in Table 1, the hypothesis that a pretraining above 25 epochs would be insufficient could not be asserted. In the context of fine tuning,

Table 3: Speech recognition accuracy after applying the fine-tuning of frozen layer.

Research objects	Class hours	Percentage of frozen fine adjustment layer				
		10	30	50	70	90
Male/Female (double)	25	78.76	80.42	78.21	71.89	41.40
	50	78.28	78.77	78.75	72.81	50.58
	100	79.69	80.40	79.37	70.88	48.60
	200	80.20	78.81	78.86	78.33	50.70
Female/Male	25	64.50	61.22	61.52	54.52	23.72
	50	63.31	62.90	59.31	53.19	29.19
	100	63.68	63.60	61.43	54.43	34.43
	200	62.21	61.50	57.23	53.80	35.56

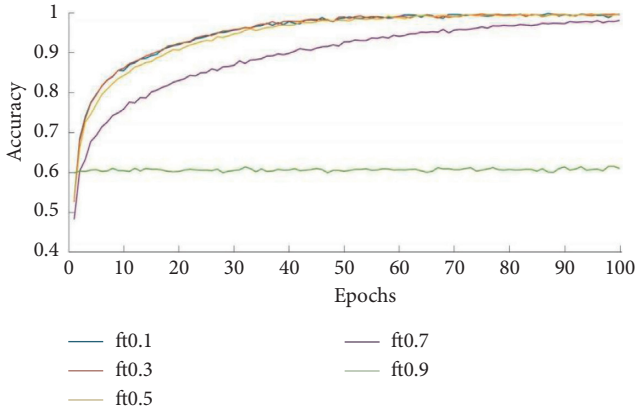

Figure 4: Training accuracy of 10400 male and female voice file sets with different number of frozen layers.

only 25 class hours of pretraining produces similar results as other test cases. Table 3 demonstrates that doubling the dataset improves training outcomes for both male and female voices, resulting in an accuracy of almost 80%, despite the dataset's limitation of just 10400 files.

It can be seen from the saturation of learning rate and accuracy in Figure 4 that more than 50 class hours of training is enough for fine-tuning, and the proportion of frozen layer is less than 50%, as assumed by the results in Table 2.

5. Conclusion

Automatic speech recognition (ASR) is a technique that is becoming more important with the passage of time and is being used in our daily lives. This paper mainly uses the combination of DCNN and transfer learning techniques, i.e., VGG-16, for speech recognition. The main goal of this study is to use transfer learning techniques for the speech recognition and to increase the recognition accuracy of audio speech files. Although utilizing a totally different dataset, the primary goal is to use transfer learning. Although different datasets are used, the simulation results show that the pretraining features are generally applicable even if there is a difference between the target database and the source dataset of the pretraining model. The other principal aim was to investigate transfer learning in the context of spoken letter identification on a limited dataset. In the application of speech letter recognition, transfer learning is used on small data sets. The simulation results show that, even for a very small dataset, it can detect voice letters significantly, but the recognition accuracy is slightly lower than that of the other methods using large datasets. However, only using the data set of 10400 male and female voice files, even if part of the audio data is recorded under noise conditions, the accuracy reaches nearly 80%. The application of small datasets reduces the time and cost of datasets generation and also reduces the time of training the model and the demand for computing power. The future work of this paper is to use more transfer learning techniques along with the ML and DL algorithms in order to improve the speech recognition accuracy using both the small and large datasets.

Data Availability

The data used to support the findings of this study are included within the article.

Disclosure

The paper extends the conference paper "Spoken Letter Recognition using Deep Convolutional Neural Networks on Sparse and Dissimilar Data" (https://ieeexplore.ieee.org/document/8702300) in 2019 IEEE International Symposium on Circuits and Systems (ISCAS).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] A. C. Davis and H. J. Hoffman, "Hearing loss: rising prevalence and impact," *Bulletin of the World Health Organization*, vol. 97, no. 10, pp. 646–646A, 2019.

[2] N. A. Lesica, "Hearing aids," *The Hearing Journal*, vol. 71, no. 5, pp. 43–46, 2018.

[3] L. D. Rosenblum, M. A. Schmuckler, and J. A. Johnson, "The McGurk effect in infants," *Perception & Psychophysics*, vol. 59, no. 3, pp. 347–357, 1997.

[4] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, vol. 335, no. 1273, pp. 71–78, 1992.

[5] R. Rotili, E. Principi, S. Squartini, and B. Schuller, "A real-time speech enhancement framework in noisy and reverberated acoustic scenarios," *Cognitive Computation*, vol. 5, no. 4, pp. 504–516, 2013.

[6] S. Abolfazli, Z. Sanaei, A. Gani, F. Xia, and L. T. Yang, "Rich mobile applications: genesis, taxonomy, and open issues," *Journal of Network and Computer Applications*, vol. 40, no. 7, pp. 345–362, 2014.

[7] R. D. Pea, M. I. Mills, E. Hoffert et al., "Methods and apparatus for interactive point-of-view authoring of digital video content," US7823058B2, 2006.

[8] M. Teresa Pazienza and A. Stellato, *Semi-automatic Ontology Development: Processes and Resources*, Information Science Reference, 2012.
[9] S. Sun, "Evaluation of potential correlation of piano teaching using edge-enabled data and machine learning," *Mobile Information Systems*, vol. 2021, pp. 1–11, 2021.
[10] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," *Complex Networks and Their Applications VIII*, vol. 10, pp. 928–940, 2019.
[11] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 5859–5866, 2020.
[12] M. N. Stolar, M. Lech, R. S. Bolia, and M. Skinner, "Real time speech emotion recognition using RGB image classification and transfer learning," in *Proceedings of the International Conference on Signal Processing & Communication Systems*, pp. 1–8, IEEE, Surfers Paradise, Australia, December 2017.
[13] Y. Xu, "Systematic study on expression of vocal music and science of human body noise based on wireless sensor node," *Mobile Information Systems*, vol. 2021, pp. 1–9, 2021.
[14] L. Liu, "Moving object detection technology of line dancing based on machine vision," *Mobile Information Systems*, vol. 2021, pp. 1–9, 2021.
[15] D. P. Liu and Q. S. Zhu, "Real-time speaker detection in conversational speech," *Journal of Chinese Computer Systems*, vol. 8, 2008.
[16] R. D. Pea, M. I. Mills, and J. H. Rosen, "Interactive point-of-view authoring of digital video content using a resizable overlay window and a cylindrical layout," US8972861[P], 2015.
[17] M. E. A. Elshaer, S. Wisdom, and T. Mishra, "Transfer learning from sound representations for anger detection in speech," arXiv preprinthttp://arxiv.org/abs/1902.02120, 2019.
[18] P. Arora and R. Haeb-Umbach, "A study on transfer learning for acoustic event detection in a real life scenario," in *Proceedings of the International Workshop on Multimedia Signal Processing*, October 2017.
[19] D. Cevher, S. Zepf, and R. Klinger, "Towards multimodal emotion recognition in German speech events in cars using transfer learning," 2019, https://arxiv.org/abs/1909.02764.
[20] K. Kalischewski, D. Wagner, J. Velten, and A. Kummert, "Spoken letter recognition using deep convolutional neural networks on sparse and dissimilar data," in *Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, IEEE, Sapporo, Japan, May 2019.
[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprinthttps://arxiv.org/abs/1409.1556, 2014.
[22] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," arXiv preprinthttps://arxiv.org/abs/1605.07678, 2016.
[23] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice hall PTR, Hoboken, New Jersey, US, 2001.
[24] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: an overview," *Issues in visual and audio-visual speech processing*, vol. 22, p. 23, 2004.
[25] Y. Muhammad, M. D. Alshehri, W. M. Alenazy, T. Vinh Hoang, and R. Alturki, "Identification of pneumonia disease applying an intelligent computational framework based on deep learning and machine learning techniques," *Mobile Information Systems*, vol. 2021, pp. 1–20, 2021.