One Voice is All You Need: A One-Shot Approach to Recognize Your Voice

Priata Nowshin*, Shahriar Rumi Dipto*, Intesur Ahmed*, Deboraj Chowdhury*, Galib Abdun Noor*,
Amitabha Chakrabarty*, Muhammad Tahmeed Abdullah† and Moshiur Rahman*
Department of Computer Science and Engineering, BRAC University*
Department of Robotics & Mechatronics Engineering, University of Dhaka†
Email:(priataoshru, shahriardipto7, intesur.1998, deboraj.chy98, abdunnoorgalib)@gmail.com,
amitabha@bracu.ac.bd, a.tahmeed@yahoo.com, moshiur.radif@gmail.com

***Abstract*—In the field of computer vision, one-shot learning has proven to be effective, as it works accurately with a single labeled training example and a small number of training sets. In one-shot learning, we must accurately make predictions based on only one sample of each new class. In this paper, we look at a strategy for learning Siamese neural networks that use a distinctive structure to automatically evaluate the similarity between inputs. The goal of this paper is to apply the concept of one-shot learning to audio classification by extracting specific features, where it uses triplet loss to train the model to learn through Siamese network and calculates the rate of similarity while testing via a support set and a query set. We have executed our experiment on LibriSpeech ASR corpus. We evaluated our work on N-way-1-shot learning and generated strong results for 2-way (100%), 3-way (95%), 4-way (84%), and 5-way (74%) that outperform existing machine learning models by a large margin. To the best of our knowledge, this may be the first paper to look at the possibility of one-shot human speech recognition on the LibriSpeech ASR corpus using the Siamese network.**

***Index Terms*—One-shot learning, Siamese neural network, audio classification, speaker recognition, triplet loss.**

I. Introduction

Simply described, meta-learning is the notion of learning how to learn. The one-shot learning strategy, in which a neural network can classify a sample by learning only one example of a class, has become a reality with the concept of meta-learning. In this paper, we propose an efficient and faster speaker recognition model compared to the state of the art models. It is of great interest to the user that the machine now can interact with others in the same way as a human would. As a result, there is a rapid increase in the interest of accurately identifying the speaker with whom the machine is conversing just based on their speech in a real-world situation [1], [2]. Traditional speaker identification techniques, on the other hand, necessitate retraining in every case when a new speaker is introduced [3]. This creates the need of storing a huge amount of data and in most of the traditional systems many voice samples at different periods from each speaker are required to accurately detect a speaker [4]. This could be expensive and deploying a speaker recognition system can be complex and time-consuming. To do simple tasks like recognizing audio samples of a speaker, traditional deep learning algorithms are notorious for requiring a huge amount of training data.

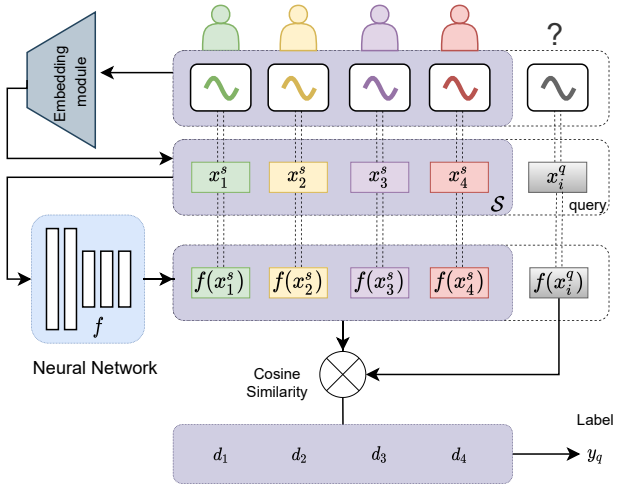

Figure 1: Multiple colors represent multiple people. The audio samples of these classes are then embedded in creating support set samples x_1^s, x_2^s, x_3^s and x_4^s which are then passed through our neural network f. The outcomes are $f(x_1^s), f(x_2^s), f(x_3^s)$ and $f(x_4^s)$. On the other hand, an unknown class which needs to be recognized follows the same procedures creating query sample vector $f(x_i^q)$ which is compared with each support sample vector and cosine similarity (d_1, d_2, d_3, d_4) are calculated. The final classification, y_q is done by finding out the maximum similarity.

One-shot learning is a classification problem in which each class is given one example, which is used to train a computer model, which must then make predictions about many unknown examples in the future. The purpose of this paper is to develop a one-shot learning based voice recognition system using a Siamese network that will be able to classify a speaker after hearing his speech once. We look forward to training and testing our dataset using the one shot approach, which is a comparatively new approach for speaker recognition/classification problems. The authentication model developed in this paper is capable of extracting appropriate voice features as well as a mechanism that measures the similarity between the two inputs to identify whether the two recordings belong to the same speaker or not. It can

be computationally costly to learn good features for machine learning applications, and it can be difficult in situations when there is limited data [5]. We are looking forward to presenting a one-shot learning approach that will allow us to train the model faster and with fewer samples. In this paper, we emphasize teaching our system to perform with better accuracy in scenarios when the system has no prior knowledge of the context of the speaker's utterances. Our tests are conducted under difficult but realistic conditions with limited training data. Regardless of the fact that there have been several solutions to the one-shot learning task on human voice data, our research highlights the following contributions:

1) To the best of our knowledge, our work is the first ever approach to evaluate one shot voice recognition task on LibriSpeech ASR human voice corpus [6].
2) Our approach uses computationally light handcrafted features for one shot voice recognition in contrast to features extracted from a deep neural network which is computationally heavy.

II. Literature Review

Artificial intelligence, in combination with cognitive science, is expanding very quickly. It warrants the creation and implementation of a variety of real-time applications, and one of them is speaker recognition. Voice biometrics has recently been used to verify individual identity. Because of its simplicity, distinctiveness, and universality, the human voice is the most useful mode of communication. Speaker identification research has a lengthy history. Researchers began studying speaker identification in the 1940s by the use of spectrogram, but the results were unsatisfactory. Speaker recognition entered the period of artificial intelligence study after numerous computer scientists proposed the concept of artificial intelligence in 1956 [7]. Researchers intended to use computer programs to identify speakers at the time, but speaker identification research did not yield promising results due to limited computer hardware capabilities and the immaturity of related algorithms. The step of computing a sequence of feature vectors that characterize the word is known as features extraction. This feature extraction provides a compact representation of the given speech signal. In automatic speaker recognition systems, Mel Frequency Cepstral Coefficients (MFCCs) are one of the most used techniques especially. It is both widely used for speech and speaker recognition [8].

In the field of computer vision, one-shot learning with Siamese neural organizations has been considered, for instance, in transcribed person recognition [5] and object classification [9]. In any case, this idea has gotten unreasonably minimal public consideration in the field of human voice arrangement, where the expectation is to perceive the marks of the human voice. This is in all likelihood because of the way that profound learning calculations in the space of sound are moderately new, having just been openly accessible for the past ten years. [10]–[12].

Li Fei et al. [13], put forward One-shot learning which utilized a Bayesian strategy to achieve picture arrangement. Following this, Li Fei Fei et al. published a paper that expanded on this topic [9]. Lake et al. [14] examined one-shot learning within the realm of handwritten characters, where the author looked at one-shot learning through the lens of science of the mind. From that point forward, researchers have delivered various papers on deep learning and one-shot learning, and the omniglot dataset has been set as a benchmark for handwritten character identification using one-shot approach [14]–[16].

The use of deep Siamese CNN for image classification was investigated by Koch et al. [5]. Its application was to utilize one-shot learning. They also advised that this methodology be applied to other domains' one-shot learning assignments. This thought was additionally reached out by Google Deep Mind's Vinyals et al. [17], who utilized a more confounded organization engineering and a preparation approach that impersonated the trial case. When compared to other competing algorithms, the authors were able to significantly improve one-shot classification on the Omniglot dataset [16]. With regards to entity relation extraction, one-shot learning with convolutional Siamese networks has been investigated in [18]. This extraction of fine-grained relationships proposed was viewed as a one-shot classification problem, intending to predict unusual relationships joining samples using a single or a few examples. A quadruplet Siamese deep neural network for one-shot learning for the following visual article was recently employed by Dong et al. [19]. They put forward a network design that used an aggregate of four info tests for one-shot learning which differed from previous work.

An architecture named Siamese convolutional neural network (SCNN) was designed for a one-shot speaker identification which was proposed by Velez et al. [3]. The author's main objective was to develop a model to identify whether two audio recordings had been created by the same speaker. Using the notable VGG [20] and ResNet-50 [21] designs, these three portrayals accomplished with a serious level of accuracy of 81.2% in certifiable ecological assessment errands, demonstrating that one-shot learning with Siamese organizations can be stretched out to true sound classification issues. Elof et al. [22] investigated the topic of multimodal one-shot learning by using pairs of spoken and visual numbers to determine whether a Siamese representation exists which could do one shot classification on this input data in multiple formats.

With the use of Siamese networks, Manocha et al. [23] suggested a new method for sound collection based on content. In a recent 2019 study, Zhang et al. [24] used Siamese CNNs in the framework of the audio quest. For their method, they have used a Siamese framework that can extract characteristics as well as predict similarities among voice limitations as well as actual audio. For this topic, the authors presented two systems: symmetric IMINET and asymmetric TLIMINET. The discoveries showed that the two variations of the framework outflanked an advanced framework and that utilizing the idea of move learning for a Siamese organization altogether further developed sound recuperation effectiveness [25].

III. Methodology

In this section, we define the problem statement of one-shot voice recognition and describe how our proposed method can tackle the problem.

A. Problem Statement

The dataset is divided into three segments: the training set ($\mathcal{T}$), the support set ($\mathcal{S}$) , and the query set ($\mathcal{Q}$). The set $\mathcal{T}$ is used by the model to learn to compare voice samples. This training is done in a meta-learning setup. The sets $\mathcal{S}$ and $\mathcal{Q}$ use the same set of classes which are disjoint from $\mathcal{T}$. The training set is formed of arbitrary number pairs of voice samples and labels from K_T number of classes. Let the number of samples in $\mathcal{T}$ be N_T. So the training set, $\mathcal{T} = \{(x_i^t, y_i^t)\}_{i=1}^{N_T}$.

In one-shot learning, the set $\mathcal{S}$ has K_S classes, each representing one person and for each class there is only one pair of voice sample and label, $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{K_S}$. Thus, the number of samples in $\mathcal{S}$ is $N_S = K_S$.

The query set is just a set of N_Q voice samples from K_S classes, $\mathcal{Q} = \{x_i^q\}_{i=1}^{N_Q}$. The problem of this paper is to assign a label $y_q \in \mathbf{C}$ where $\mathbf{C} = \{c_i\}_{i=1}^{K_S}$ for any x_i^q.

B. Proposed Model

To train our model which learns to compare voice samples, we generate a set of meta-samples from $\mathcal{T}$. A meta-sample $((x_a, y_a), (x_n, y_n), (x_p, y_p))$ is composed of three samples named Anchor, Negative, and Positive. The anchor audio samples serve as the foundation for the instruction. It is a single audio sample of a class or individual. And the positive samples come from the same class. The third sample is a negative sample, which does not fall into the same category as the first two. The training set $\mathcal{T}$ is redesigned to form set of N_T' such tuples. The new training set is $\mathcal{T}' = \{((x_{a_i}^t, y_{a_i}^t), (x_{n_i}^t, y_{n_i}^t), (x_{p_i}^t, y_{p_i}^t))\}_{i=1}^{N_T}$.

The deep neural network $f(.)$ takes a sample's audio features as input gives a feature vector as output, which is used for comparing two samples shown in Figure 2.

$$d_i+ = \phi(f(x_{a_i}^t), f(x_{p_i}^t))$$
$$d_i- = \phi(f(x_{a_i}^t), f(x_{n_i}^t))$$

Where the function ϕ calculates the cosine similarity of two vectors.

$$\phi(\vec{p}, \vec{p}) = \frac{\vec{p}.\vec{q}}{|\vec{p}| \times |\vec{q}|}$$

The unique training loss called the triplet loss [26] is the main essence of the Siamese network. The triplet loss function takes three inputs per training (the anchor, positive, and negative samples). The loss function is designed in such a way that the distance of the anchor with the positive sample is minimized and the distance with the negative sample is maximized.

$$\mathcal{L}(x_{a_i}^t, x_{p_i}^t, x_{n_i}^t) = max(|d_i+|^2 - |d_i-|^2 + \alpha, 0)$$

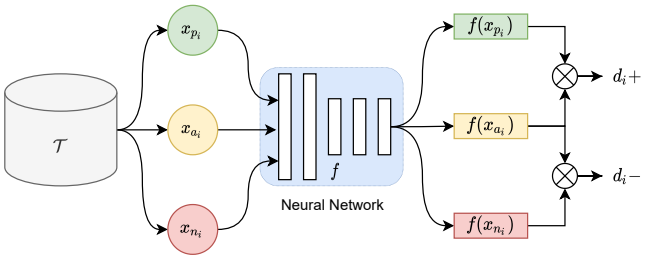

Figure 2: Three samples are x_{p_i}, x_{a_i} and x_{n_i} are carefully picked from training set $\mathcal{T}$ and passed through deep neural network. The output $f(x_{a_i})$ is compared with $f(x_{p_i})$ and $f(x_{n_i})$ to calculate the positive distance ($d+$) and the negative distance ($d-$) respectively.

In the loss function, a positive tuning hyper-parameter, alpha(α), is employed to ensure that the negative distance is substantially bigger than the positive distance.

C. Testing

Here the query set contains multiple samples of multiple classes that have never been trained before and the support set contains one sample of multiple classes, sharing the same label space. The bigger the label space, the difficult it is for the model to compare and predict. The query sample is feed forwarded through the neural network to extract features, which is then compared to each of the support set class samples whose features are also extracted by the network. Finally, the model chooses the support class with the highest similarity for the query sample as the predicted output. Using the same technique, all remaining query samples are also evaluated with the support set samples. The testing phase is well illustrated in Figure 1.

D. Feature extraction

Mel Frequency Cepstral Coefficients (MFCCs) are one of the most sophisticated methods for extracting data from audio samples providing state-of-the-art performance in audio recognition tasks. Multiple types of speech recordings are evaluated in the cepstral domain using MFCC to distinguish between individuals with Parkinson's disease and healthy people in the study [27]. According to [28], the following are the steps to calculate MFCCs for a given audio sample:

1) Shorten the signal by slicing it into shorter time frames.
2) Calculate the power spectrum periodogram estimate for each frame.
3) On total the energy in each filter, apply the mel filterbank to the power spectra.
4) Take the log filterbank energies and apply the discrete cosine transform (DCT) to them.
   The Discrete Fourier Transform (DFT) $S_i(k)$ of the i^{th} frame of time domain signal $S_i(n)$ can be written as

$$S_i(k) = \sum_{n=1}^{N} S(n)h(n)e^{-j2\pi kn/N} \qquad 1 < k < K \qquad (1)$$

Subsets	Speaker	Phase	Arrays	Samples
Train-clean-360	921	Training	Anchor	33675
			Negative	33675
Train-clean-100	255		Positive	33675
Test clean	40	Testing	Support set	1 for each class
			Query set	n for each class

Table I: Number of samples in each sets. Train-clean-360 and train-clean-100 have 921 and 255 speakers respectively, yielding 101,025 samples, each separated into three sections: anchor, positive, and anchor.

Here, $h(n)$ is the Hamming Window and K is the length of the DFT in Equation 1. The power spectrum of frame i is determined using a periodogram in Equation 2.

$$P_i(k) = \frac{1}{N}|S_i(k)|^2 \tag{2}$$

$$m = 1125\ln(1 + \frac{f}{700}) \tag{3}$$

$$m^{-1}(m) = 700 * (\exp\left(\frac{m}{1125}\right) - 1) \tag{4}$$

Equations 3 and 4 are used to convert frequency to mel and mels to frequency, respectively.

$$H_m(k) = \begin{cases} 0 & k¡f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \le k \le f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) \le k \le f(m+1) \\ 0 & k\ f(m+1) \end{cases} \tag{5}$$

The first filter bank in Equation 5 will start at the very first point, peak at the next, and then go back to its initial point at the final point. The second filter bank will start at the second point, peak at the third point, then drop to zero at the fourth point, and so on. The formula for estimating these is as follows. Here's a list of M+2 Mel-spaced frequencies, along with the set of filters that is needed.

IV. Experiment

A. Setup

Dataset: The LibriSpeech ASR corpus [6] was utilized. There are 12 subsets of voice samples, text files, and other data in the collection. Subsets having the word "clean" in their name are said to be "cleaner". Based on our requirements, we worked on three of the subsets, "Train-clean-100", "Train-clean-360", and "Test-clean". There are 921 speakers in Train-clean-360, 255 speakers in Train-clean-100, and 40 speakers in Test-clean shown in Table I. As a result, we get 1176 unique speakers in total in the training set. For each anchor and positive array, we took a total of 33675 samples along with 33675 samples for the negative array. In our research, a total of 101,025 samples were trained. According to our research, LibriSpeech ASR corpus has performed well with traditional approaches. As it is supposedly cleaner than other most used datasets and also works better on almost all the state of the art models, we chose this dataset to proceed our work with. However, in order to adapt in our one-shot learning process, it must be partitioned. The training set and test set classes or categories in our suggested model cannot be repeated.

K-way	Accuracy	f1-Score	Precision	Recall
2-way	1.00	1.00	1.00	1.00
3-way	0.95	0.95	0.95	0.95
4-way	0.84	0.84	0.84	0.84
5-way	0.74	0.74	0.78	0.74

Table II: Performance of our proposed model on 2,3,4,5-way one shot classification.

Evaluation: The assessment metrics accuracy, precision, recall, and F1-score were used in this research. We combined three different training sets (anchor, positive, and negative) into a single dataset and compared the positive and negative similarity during training phase. Positive similarity is a metric that determines the percentage of a sample that is matched to its correct owner, whereas negative similarity indicates that a sample is unmatched to those who are not its owner. We compared our method with traditional classifiers used in supervised learning setup: K-Nearest Neighbour (KNN), Random Forest Classifier, and Multilayer Perceptron.

Implementation details: We have used the Siamese network to measure the similarity and hence classify the data. For this, we have taken samples of 2 classes, 3 classes, 4 classes and 5 classes per testing. The dense layer had size two layers of size 80 followed by three layers of size 40, and we utilized the activation function "relu.". The Adam optimizer was used with a value of 0.001. We completed 50 epochs in all, with a final loss of 2.4061.

B. Result

According to our study of training data, positive similarity was 98.4%, while negative similarity was 95%. Our model's training time was 6.43 seconds. The confusion matrix for 2 to 5 way one-shot classification are shown in Figure 3. For each of the 2, 3, 4 and 5 way one-shot classification, we achieved accuracy of 100%, 95%, 84% and 74% respectively. The performance of our proposed model is demonstrated with evaluation metrics in Table II.

V. Discussion

A. Combined analysis for other models

The aggregate categorization report for all models implemented is shown in the Figure 4. It is made up of accuracy for all classes. We may infer from this figure that the proposed model, one-shot learning with Siamese network, produces the greatest scores among the others. When we compare accuracy for two classes, we obtain 76%, 29% and 22% for KNN classifier, Random Forest classifier, and MLP classifier respectively. Our model, on the other hand, provides exact accuracy 100%. Similarly for 3, 4, and 5 way one-shot classification, our model outperforms other traditional supervised learning models by a great margin. The accuracy graph for all the models combinedly showed in Figure 4.

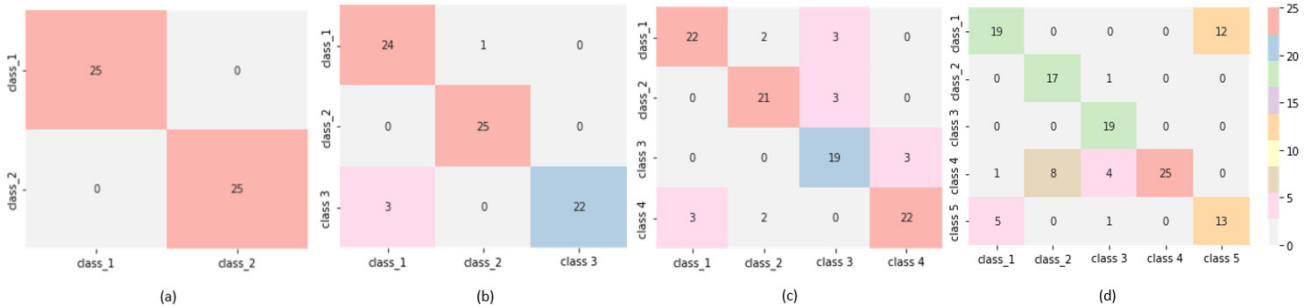

Figure 3: Confusion matrix of (a) 2-way, (b) 3-way, (c) 4-way and (d) 5-way one-shot voice recognition task using our proposed model.

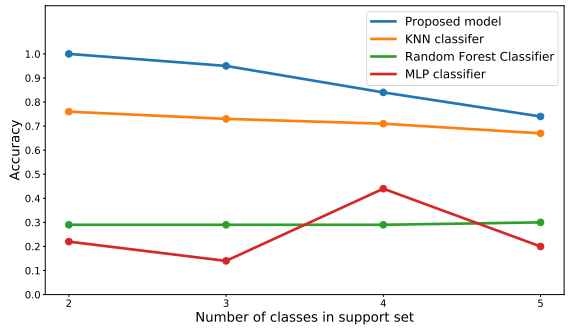

Figure 4: The change of accuracy for increasing number of support set for different classifiers and our proposed model.

B. *Feature analysis*

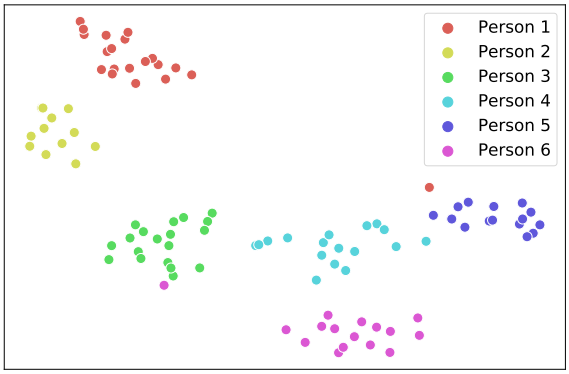

Figure 5: T-SNE visualization of MFCCs features of audio sample of 6 persons.

The goal of MFCCs and mel spectrogram is to transform a time domain signal into a frequency domain signal so that we can grasp all of the data in voice sounds. Mel-frequency spectrograms are similar to normal spectrograms, except they use Mel-frequency spacing. MFCCs need us to compute the Cepstrum or power spectrum, which is a "spectrum of a spectrum". Cepstrum outperforms the mel spectrogram because of its unique calculation. Filtering in the time domain is similar to frequency domain multiplication, which is then added in the cepstrum. This enables better separation and investigation of the source and transmission channel. MFCCs are created without much pre-processing from the raw audio input. The efficacy of the generated features is shown in Figure 5 in a TSNE plot.

VI. Conclusion

In this paper, we explored the challenges of speaker recognition with limited training samples. We developed a one-shot learning strategy based on triplet loss to deal with the data shortage. We achieved a remarkable difference by employing subsets of the LibriSpeech ASR corpus, indicating that our suggested model outperforms alternative state-of-the-art architectures. Traditional methodologies, however, have a significant latency when the dataset is small and untrained. Our technique identifies the issue and seeks to improve the chances of recognizing someone whose voice has never been trained. As our supplied analysis and research statement demonstrate, the speaker recognition task can be completed with extremely few resources. This methodology will be useful in a variety of security situations when obtaining several samples is difficult. In the future work, we will assess the model's performance with a few additional modifications and tunings for larger numbers of classes. A contractive auto-encoder may also be used to extend the Siamese network by learning a generalized embedding. The embeddings computed are utilized to train a prototype embedding using prototypical loss. We will also test it out on an embedded system to see how it performs in real-world scenarios.

References

[1] F. Grondin and F. Michaud, "Wiss, a speaker identification system for mobile robots," in *2012 IEEE International Conference on Robotics and Automation*, IEEE, 2012, pp. 1817–1822.

[2] K. Youssef, S. Argentieri, and J.-L. Zarader, "Binaural speaker recognition for humanoid robots," in *2010 11th International Conference on Control Automation Robotics & Vision*, IEEE, 2010, pp. 2295–2300.

[3] I. Vélez, C. Rascon, and G. Fuentes-Pineda, "One-shot speaker identification for a service robot using a cnn-based generic verifier," *arXiv preprint arXiv:1809.04115*, 2018.

[4] M. Alsulaiman, G. Muhammad, Y. Alotaibi, A. Mahmood, and M. A. Bencherif, "Building a speaker recognition with one sample," in *Proceedings of the Second Symposium International Computer Science and Computational Technology (ISCSCT'09) Huangshan, People's Republic of China*, 2009, pp. 26–28.

[5] G. Koch, R. Zemel, R. Salakhutdinov, *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, Lille, vol. 2, 2015.

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[7] M. Minsky, "Steps toward artificial intelligence," *Proceedings of the IRE*, vol. 49, no. 1, pp. 8–30, 1961.

[8] S. Engelberg, *Digital signal processing: an experimental approach*. Springer Science & Business Media, 2008.

[9] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[10] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *Advances in neural information processing systems*, vol. 22, pp. 1096–1104, 2009.

[11] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, IEEE, 2015, pp. 1–6.

[12] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *2014 22nd European signal processing conference (EUSIPCO)*, IEEE, 2014, pp. 506–510.

[13] M. Brown, D. G. Lowe, *et al.*, "Recognising panoramas.," in *ICCV*, vol. 3, 2003, p. 1218.

[14] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, 2011.

[15] B. Lake, R. Salakhutdinov, and J. Tenenbaum, "Concept learning as motor program induction: A large-scale empirical study," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, 2012.

[16] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[17] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[18] J. Yuan, H. Guo, Z. Jin, H. Jin, X. Zhang, and J. Luo, "One-shot learning for fine-grained relation extraction via convolutional siamese neural network," in *2017 IEEE International Conference on Big Data (Big Data)*, IEEE, 2017, pp. 2194–2199.

[19] Y. Wang, X. Deng, S. Pu, and Z. Huang, "Residual convolutional ctc networks for automatic speech recognition," *arXiv preprint arXiv:1702.07793*, 2017.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[22] R. Eloff, H. A. Engelbrecht, and H. Kamper, "Multimodal one-shot learning of speech and images," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 8623–8627.

[23] P. Manocha, R. Badlani, A. Kumar, A. Shah, B. Elizalde, and B. Raj, "Content-based representations of audio using siamese neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 3136–3140.

[24] Y. Zhang, B. Pardo, and Z. Duan, "Siamese style convolutional neural networks for sound search by vocal imitation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 429–441, 2018.

[25] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *arXiv preprint arXiv:1701.02720*, 2017.

[26] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.

[27] A. Benba, A. Jilbab, and A. Hammouch, "Analysis of multiple types of voice recordings in cepstral domain using mfcc for discriminating between patients with parkinson's disease and healthy people," *International Journal of Speech Technology*, vol. 19, Sep. 2016. DOI: 10.1007/s10772-016-9338-4.

[28] B. Logan, "Mel frequency cepstral coefficients for music modeling," *Proc. 1st Int. Symposium Music Information Retrieval*, Nov. 2000.