Volume: 5, Issue: 3
Page: 15-26
2021

International Journal of
Science and Business
Journal homepage: ijsab.com/ijsb

IJSAB International

Identifying Speakers Using Deep Learning: A review

Lawchak Fadhil Khalid & Adnan Mohsin Abdulazeez

Abstract:

With the advancement of technology and the increasing demand on smart systems and smart applications that provide a quality-of-life improvement, there has been a surge in the demand of more conscious applications, Machine Learning (ML) is considered one of the driving forces behind implementing these types of applications, and one of its implementations is Speaker Identification (SID). Deep Neural Networks (DNNs) and also Recurrent Neural Networks (RNNs) are two main types of Deep Learning that are being used in the implementation of such applications. Speaker Identification is being utilized more and more on daily basis and is being focused on by the research community as a result of this demand. In this paper, a review will be conducted to some of the most recent researches that were conducted in this area and compare their results while discussing their outcomes.

IJSB
Literature Review
Accepted 28 January 2021
Published 30 January 2021
DOI: 10.5281/zenodo.4481596

Keywords: *Machine Learning, Speaker Identification, Deep Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks.*

About Author (s)

Lawchak Fadhil Khalid (corresponding author), Technical College of Informatics - Akre, Duhok Polytechnic University (DPU), Kurdistan Region, Iraq.
Email: lawchak.fadhil@gmail.com

Adnan Mohsin Abdulazeez, Duhok Polytechnic University (DPU), Kurdistan Region, Iraq.

Introduction

The main advantages of using computers are their ability to learn and adapt to data and perform automated tasks based on these learnings, this in turn, will increase the performance of tasks that is performed on computers. This is referred to as Machine Learning (ML) in Computer Science. ML deals with cases that are able to be learnt from testing data, instead of repeating the execution of specific tasks over-and-over (Abdulqader et al., 2020; Jahwar & Abdulazeez, 2020; D. Zeebaree et al., 2020; D. Q. Zeebaree et al., 2019). Recently, the field of Speaker Identification (SID), have seen a surge in research attention due to its promising implementations in the field of security, biometrics, forensics, etc. (Qayyum et al., 2018). SID has two primary modules, feature extraction and an ML model. The model works on recognition of speaker by analyzing features that are extracted from audio content and this has been studied by many researchers (Dahake et al., 2016; Hasan & Rahman, 2004; Nakagawa et al., 2012; Qayyum et al., 2018; Wang & Lawlor, 2017). The perfect modeling of unique features of speech and language for each individual is one of the biggest challenges in SID. These features rely on a range of characteristics, such as gender, pitch, articulation, age and acoustic environment, resulting in the production of unique varieties in the spoken accent and speech pronunciations (Qayyum et al., 2018), (Latif et al., 2018). For the past decade, ML methods and more specifically, Deep Neural Networks (DNNs) has been utilized extensively in the field of Speech Recognition and they have achieved great success (Bargarai et al., 2020; Hinton et al., 2012). These methods have been also seen to achieve greatly in the field of Speaker recognition (Ferrer et al., 2016; Garcia-Romero et al., 2014; Lei et al., 2014) or language recognition (Lopez-Moreno et al., 2014; Matějka et al., 2016; Song et al., 2013). When DNNs are utilized in Speech Recognition, they usually are getting trained for frame-by-frame classification of how the speech sounds (e.g., from phone recordings). Likewise, a DNN is trained directly on frame-by-frame classifications for languages that were successfully used for language recognition in (Lopez-Moreno et al., 2014), however, this approach has only provided a competitive performance to short utterances of speech (Matějka et al., 2016).

Recurrent Neural Networks (RNNs) on the other hand, and more specifically, Long Short-Term Memory (LSTM) model have shown to be more efficient and fairly more successful in the implementations of SID and have proven that the models based on RNNs seem to outperform other models in different lengths and recitations (Qayyum et al., 2018), especially in complex problems such as sound event detection (Rui Lu, Zhiyao Duan, 2017), speech recognition (Graves et al., 2013), machine translation (Adeen et al., 2020; Bahdanau et al., 2016) and speech emotion recognition (Latif et al., 2020). The rest of the paper is organized as following: Section 2 through 5 present theoretical background. Section 6 provides related works that have been performed on the area. Then in Section 7, a brief discussion and comparison of the results is made. Finally, Section 8 provides a conclusion of the review.

2. Machine Learning (ML)

ML is utilized to teach machines how to operate on data in an efficient manner. Sometimes, upon analyzing the data, we are not able to extract information or interpret a pattern from this data. In such cases, we refer back to using ML algorithms (Abdulazeez et al., 2020; Dey, 2016; Richert & Coelho, 2013).

Figure 1: Different stages of machine learning (Nassif et al., 2019)

The purpose of ML is to learn from data. The process of learning, however, can be in many different ways; For example, we have many researches that shows machines learn on their own (Bowles, 2015; Max Welling, 2010) but also, we have other researches showing teaching of ML through training data (Batista et al., 2004; Bhavsar & Ganatra, 2012). ML is a section of Artificial Intelligence (AI) and it is closely related to (and sometimes even overlapped with) computational statistics, which at the same time, focuses on the use of computer to make predictions (Abdulqader et al., 2020; Xin et al., 2018). ML is commonly utilized in many different fields in order to solve difficult problems that cannot be solved using regular based on computer approaches (Maulud & Abdulazeez, 2020; Murphy, 2012; Domingos, 2016; Shalev-Shwartz & Ben-David, 2014; Zeebaree et al., 2019). The implementations of ML range from medical uses (Abbasi & Goldenholz, 2019; Choy et al., 2018; Peiffer-Smadja et al., 2020; Zeebaree et al., 2019) all the way into generic uses such as education (Bacos, 2020; Fiebrink, 2019; Hodges & Mohan, 2019), sports(Ashley, 2020; Bunker & Thabtah, 2019; Cust et al., 2019; Koseler & Stephan, 2017), finance(Agrawal et al., 2019; Emerson et al., 2019; Ghoddusi et al., 2019) and many more fields. ML is also strongly linked to mathematical optimization, which provides the field with theories, application domains and techniques (Xin et al., 2018). The main types of ML are: (i) supervised learning, (ii) unsupervised learning, (iii) semi - supervised learning, (iv) reinforcement learning and (v) deep learning (Nassif et al. 2019).

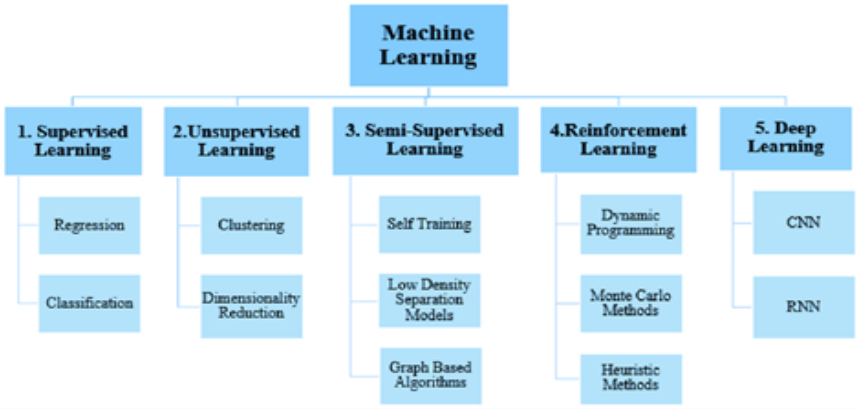

Figure 1: Machine Learning types (Nassif et al. 2019)

3. Deep Learning (DL)

Deep Learning (DL) is the type of ML that we will be focusing on mostly in this review since it is the main type of ML that is used for SID. According to (Gary Marcus, 2018), DL is primarily a statistical technique that is used to classify patterns, depending on the sampled data by utilizing neural networks with multiple layers. In the deep learning literature, neural networks typically consist of a series of input units that stand for objects like pixels or terms. Multiple hidden layers containing hidden units (also referred to as nodes or neurons) and a configured output unit with links between them. Such a network could be trained in a typical application on a large set of handwritten digits (these are the inputs, represented as images) and labels (these are the outputs) that determine the categories to which those inputs belong (Marcus, 2018).

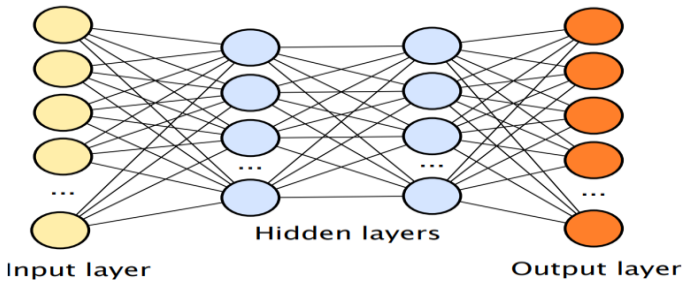

Figure 3: *An example of a Neural Network [54]*

These structures are defined as neural networks since all the nodes which are input, output and hidden can be seen as loosely similar to biological neurons, although they are vastly simplified, and the ties between nodes can be seen as representing the links between neurons in some way (Marcus, 2018). DL consists of two major types: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

3.1 Convolutional Neural Networks (CNNs) are a type of Artificial Neural Network (ANN) that is able to extract local features in a collection of data. It simplifies the network model by assigning weights on singular mapping of features, thus, overall weights can be reduced. These characteristics have pushed CNN to be used widely in pattern recognition area (Fu Jie Huang & LeCun, 2006; Vincent et al., 2008; D. Q. Zeebaree et al., 2018).

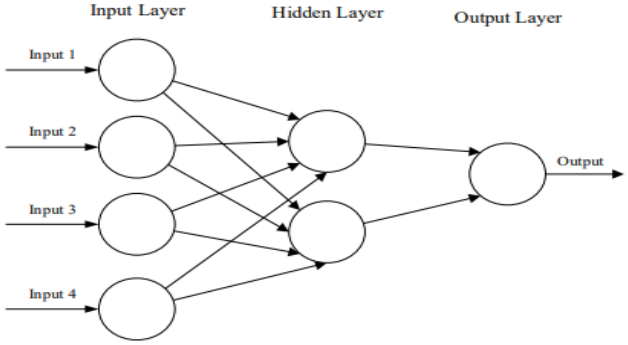

Figure 4: A basic structure of a CNN (O'Shea & Nash, 2015)

3.2 Recurrent Neural Networks (RNNs), differently, RNNs are a second type of ANNs with memory that affects their next forecasts. For forecasts, the sequential knowledge saved in the memory of RNNs is used. The concept of utilizing RNNs instead of the conventional neural network is that it is presumed that not every single input and every single output relies on each other in the traditional neural network. Therefore, the usage of classical neural networks in speech recognition is deemed a poor idea (Amberkar et al., 2018; Yashwanth et al., 2004). There are several different neural networks usable, but RNN is used by them as it is more effective for speech recognition than the others (Amberkar et al., 2018; Saini, P., 2013; Shrawankar & Thakare, 2013; Yashwanth et al., 2004).

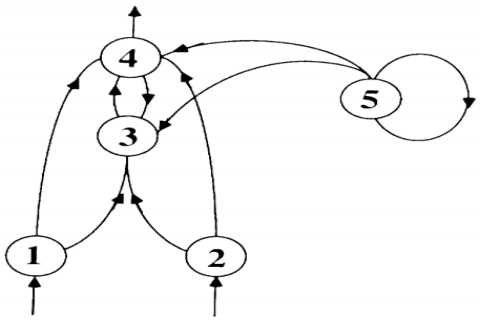

Figure 5: A basic structure of a RNN (Fernando J. Pineda, 1987)

4. Speaker Identification (SID)

In recent years, SID has drawn increasing interest from both the academic and business communities (Campbell, 1997; Hansen & Hasan, 2015; Togneri & Pullella, 2011), and is commonly utilized in implementations such as surveillance[64], discriminative speaker embedding learning (Cai et al., 2018a; Nagrani et al., 2017; Zhang & Koishida, 2017), and speaker diarization (Daniel Garcia-Romero et al., 2017). SID attempts to detect the speaker's identity automatically from an input utterance, given a closed collection of recognized speech models (An et al., 2019; Campbell, 1997; Hansen & Hasan, 2015; Tirumala et al., 2017;

Togneri & Pullella, 2011). The initial attempts at SID were conducted under optimum and highly regulated conditions with very small vocabulary scale, as explored in the analysis (Reynolds, 2002). Lately, using strategies such as UBM-GMM and Joint Factor Analysis (Kenny, 2006), as well as hybrid GMM and help vector machines, several daunting factors such as ambient disturbance, impersonation, and multiple ethnic diversities are being addressed (Togneri & Pullella, 2011). In addition, the advent of daunting datasets such as Speaker In The Wild (SITW) (McLaren et al., 2016) and its expanded variants such as VoxCeleb1 and VoxCeleb2 (Cai et al., 2018b; Chung et al., 2018; Nagrani et al., 2017), with a huge dataset of more than 5,000 speakers and a million utterances, has enabled real-world scenarios to answer SID. As a consequence, many strategies for deep learning have been proposed for this reason, including the models tested in (Chung et al., 2018; Ghahabi & Hernando, 2017; Hinton et al., 2012; Jung et al., 2018; Kenny et al., 2014; Kumar et al., 2018; Lei et al., 2014; Lopez-Moreno et al., 2014; Matejka et al., 2014; Nidadavolu et al., 2019; Park et al., 2018; Rohdin et al., 2019; Shon et al., 2018; Song et al., 2013; Takanori Yamada et al., 2013; J. Wang et al., 2019; Xie et al., 2019; Yun Lei et al., 2014).

The SID procedure entails extracting and defining speech characteristics from a collection of speakers; it is therefore important to choose the most effective methods of extraction of features which provide the most suitable representation of the features of their speech. Getting input utterances polluted with noise is one of the most challenging aspects of SID function extraction (Shahamiri & Binti Salim, 2014). Both layers of DNN remove features at various levels with layer-wise preparation (hierarchically). Deep architecture is a multi-layer hierarchical system, although each layer is self-trained to benefit from the previous layer's performance (Tirumala & Shahamiri, 2016). To overcome this difficulty, DL algorithms were used for hierarchical feature extraction (Kekre et al., 2011) to demonstrate their effectiveness in improving SID efficiency (Ghahabi & Hernando, 2017; McLaren et al., 2015; Richardson et al., 2015). DL has been active in numerous applications involving analytical and comparative function extraction (Dutta et al., 2015; Justin et al., 2015; LeCun et al., 2010; Pobar & Ipsic, 2014; Xie, J. et al., 2012). After a successful feature extraction, a proper model should be prepared to create an embedding of the received utterance by working with a proper aggregation model to identify the speaker's identity with respect to a list of speakers that were previously used for training (Hajavi & Etemad, 2019).

5. Datasets

The datasets which were utilized in order to train and validate the models are shown below in Table I:

TABLE I: *Utilized Datasets*

Source	Dataset Name	Content	Usage
(Nagrani et al., 2017)	VoxCeleb1	VoxCeleb1 contains 153,516 utterances from 1,251 celebrities that have been collected under strict constraints and do not contain any annotation flaws(Chung et al., 2018).	VoxCeleb1 is mostly preferred and used in testing speaker recognition model rather than training them.
(Chung et al., 2018)	VoxCeleb2	This is a newer iteration of its predecessor VoxCeleb1; This dataset contains 1,128,246 utterances from 6,112 different celebrities. This specific dataset, however, contains some annotation flaws (Chung et al., 2018).	Due to its several annotation flaws and its huge size, it is preferred that this dataset is used for training purposed rather than testing.

6. RELATED WORK

DL has been utilized in many different implementations of SID, in this section, we will study some of its utilizations and how it has performed under different implementations &

approaches. Qayyum et al. (2018) proposed the utilization of Bidirectional Long Short-Term Memory (BLSTM) which is based on Recurrent Neural Networks (RNNs). RNNs are known to be performing quite well in speech modeling and processing, and in their case for the task of Quranic reciter identification. Their results showed that the BLTSM-based model performed significantly better than other models that were previously used for this purpose and also, computationally less expensive. Cai et al. (2018) In an end-to-end speaker & language framework, the authors investigated the encoding/pooling layer and the loss function. They built a recognition method that recognized variable-length inputs and produced results for the utterance stage. In aggregating variable-length input sequences into a representation of utterance level, the encoding layer played a part. Apart from the simple TAP aggregation feature, to get the utterance level representation, they added a self-attentive pooling layer and a learnable dictionary encoding layer. As for the loss feature, they implemented center loss and angular softmax loss in open-set speaker verification to get the most unequal speaker embeddings. Their experiments demonstrate that an end-to-end learning system's output enhancement is important and ties this to the encoding layer and loss feature they proposed. An et al. (2019) the paper discussed two CNN based methods for SID, which are: Visual Geometry Group (VGG) nets and Residual Neural Networks (ResNet). The authors equip these two strategies with a systematic self-attention system that, instead of depending on maximum or average pooling done by previous DNN-based SID processes, learns a weighted average across all time stamps. Using this organized self-attention layer alongside numerous attention hops, the suggested method becomes acquainted with the handling of fragments of varying duration and may acquire speaker attributes from different facets of the input series at the same time. Cai et al. (2018) explored a deep length normalization strategy in end-to-end SV system. By adding two following layers (L2-normalization and a scale layer) before the output layer, the author managed to make the learned deep speaker embedding normalized in an end-to-end manner. However, the author notates that, the number of the scale parameter is crucial for his approach and specifically when the number of output categories are large. From the experiments of the author, we could note that by setting a proper value for the scale parameter the results were improving significantly. The author finishes by adding that, having a simple inner-product while training a L2-normalized deep embedding system can achieve a state-of-the-art status. Chung et al. (2018) the author created a massive collection of utterances from over 6,000 speakers then continue by developing a CNN based model SID systems and uses this newly acquired collection to train the system. Although that the new collection seems to have some annotation flaws, it still gives an amazing performance for training SID systems, as when the system is tested by using the VoxCeleb1 collection the system seemed to do an efficient job by only having a 4.19% error rate. This turned out to be one of the best results of the number of results that we studied in this paper. Xie et al. (2019) attempted a few different approaches at the subject, although they are using another CNN-based model to develop the SID system (thin-ResNet model), they try a new dictionary-based NetVLAD or GhostVLAD aggregation method that is able to be trained on end-to-end basis. The authors continue by developing their system and training it by using the VoxCeleb2 dataset and test their system by using the VoxCeleb1 dataset. At the same time, the authors tried a different approach to show the effectiveness of their work, they developed a different SID system by using the same CNN-based thin-ResNet model but instead of the dictionary-based GhostVLAD aggregation method they tested a more commonly used TAP aggregation method which resulted in high error rates due to combination of TAP with softmax loss. The authors conclude their research by stating that their new approach has an edge over other approaches since it is using fewer parameters and achieves the amazing performing capabilities on the VoxCeleb1 test. And they finally add that, although that for generic SID procedures short utterances are preferred, their work shows

that in the cases of “in the wild” data, lengthier utterances seems to perform better. Hajavi & Etemad(2019) suggested and actually implements a Fully Convolutional Network (FCN) approach in developing their SID system which they call (UtterIdNet). The researchers train their new system with the infamous VoxCeleb2 dataset and test it using the VoxCeleb1 dataset. Their system showed significant improvements in the short-segment range such as 250 ms, 500 ms and up to 2 seconds. They back their claim of having an efficient and improved system by stating that besides the system’s accuracy improvement, the system’s training took almost half of the other systems that it was compared to. However, their new system was marginally outperformed by some other systems such as the one proposed in (Shon et al., 2018), (Okabe et al., 2018), and they factor this to the choice of a simple aggregation technique for combining the different short segments within the full utterance and they state that this should be investigate this in future works. In the paper by Nagrani et al. (2020), the writers presented a modular approach that was used to automatically produce a dataset called "VoxCeleb1" which "VoxCeleb2" and later became a standard for training and testing purposes in the speech community. The authors investigate the high error rate of traditional models such as (I-Vectors + PLDA) which comparing to other SID models seem to suffer at 8.8% error rate. And they refer this back to the neglection of the softmax loss which was not used with this model. They suggest that by adding the softmax + contrastive loss to another traditional model such as VGG-M the error rate outperforms I-Vectors + PLDA by 1% difference. They conclude their study by saying that while their models are focused on 2D convolutions that are added to spectrogram parameters, their future work may involve exploring alternatives that might be more efficient, like 1D time convolutions organized as input channels with spectrogram frequencies, or ID convolutions directly added on plain waveforms. Okabe et al. (2018) indicated that for the extraction of deep speaker embedding functionality, attentive statistical pooling should be used. They proceed to add that the pooling layer suggested measures weighted means and weighted standard deviations over frame-level characteristics scaled by an attention model that allows the embedding of the speaker to concentrate only on important frames. They further add that log-term differences may be obtained as speaker characteristics in the standard deviations due to this, and that this mixture of focus and standard deviations creates a synergistic impact to provide greater discriminative power to deep speaker embedding. They finally conclude that even though that they have achieved considerable improvements with their X-Vector model in short and long duration conditions, I-Vectors will remain a challenged in longer durations (such as 300s in SRE12 CC2) and that this will be tackled in their future works. Hajibabaei & Dai (2018) the authors in this paper investigate different methods that could potentially improve prediction accuracy of a text-independent SID system. Their results show that the augmentation and time-reversion of the training data can help improve effectiveness of training sets and this, the general power of the trained network and if this augmentation is applied in the testing stage an improvement in prediction accuracy would be seen. The authors suggest that the usage of proposed loss function with independent scale and bias for each class would result in embeddings with much higher identification accuracy. And they conclude by encouraging those who are interested in the field of SID to apply recommended methods to improve the resulting system’s prediction accuracy. Lukic et al. (2016) proposed the use of simple spectrograms as input to a CNN based model and investigated the feasibility to identify speakers by using features generated by a CNN. The researchers concluded that the features that were learned by the CNN were relevant to recognize unknown speakers and also that, their system was designed to remove silence in speeches. The researches also add that, although the system can perform the preprocessing of detecting and removing unvoiced speech, their system is still not completely tuned and there is still a lot of room for improvement. Tirumala & Shahamiri (2017) suggested the use of Deep Auto Encoders (DAEs)

in the implementation of SID systems. The experiments that were held in the research were using data from 84 speakers. Their experiments carried out found that a DAE network which has three autoencoders over conventional neural network architectures was able to achieve a recognition accuracy of 98.8 percent. As illustrated in previous DL reports, the analysis confirms the value of scope, and especially with its variations in accuracy between normal back propagation and layer-wise exercise. Zhao et al. (2014) focused on SID implementation in a noisy environment which as they argue, is rarely studied in the field. Their analysis approaches the problem in two steps, with the first step utilizing a DNN classifier to remove the noise by binary masking. After that, on the basis of direct masking and restricted marginalization, they conduct SID with models that are conditioned in chosen reverberant environments. Their observation reveals that SID output over similar systems in a broad variety of signal-to-noise ratios and reverberation periods is greatly enhanced by the proposed method.

Nicolson & Paliwal (2020) introduced a sum-product networks (SPNs) based system for robust speech processing by utilizing a simple robust speaker identification task. They discussed that, although that current SPN toolkit and learning algorithms are still in their infancy, their aim was to show that SPNs are actually capable of becoming a useful tool to be used in speech-processing field in the future. The conclusion of their research shows that, in terms of SID accuracy, the SPN models seemed to be more robust that the two CNN based models that they were compared with. Additionally, the SPN models consisted of much lower parameter counts that their CNN-based counterparts. The result indicated that SPN models could be a more robust, parameter efficient alternative for regular CNN-based models in SID. Ravanelli & Bengio (2019) indicated that Mutual Information (MI) or related measurement forms are promising resources for unsupervised learning of representations, although it is difficult to quantify reciprocal information between two random variables, especially in high-dimensional spaces. Some recent experiments seem to accomplish an implied optimization of MI with the design of encoder-discriminator, which is close to that of Generative Adversarial Networks, the authors add (GANs). By optimizing shared knowledge between an encoded representation of chunks of speech, which is sampled randomly from the same sentence, they experiment with the capturing of speaker identities. The researchers suggest that this method appears to successfully acquire valuable tasks for recognizing and checking speakers, as the tests involve both unsupervised and semi-supervised learning conditions and equate the performances obtained with various objective roles.

TABLE II: *Comparison among different SID approaches and their error rates*

Source	*Model used*	*Feature extraction method*	*Loss*	*Dataset*	*Error Rate %*
(Nagrani et al., 2020)	I-Vector	PLDA	-	VoxCeleb1	8.80
(Cai et al., 2018b)	ResNet34	SAP	A-softmax+PLDA	VoxCeleb1	4.40
(Cai et al., 2018b)	ResNet34	LDE	A-softmax+PLDA	VoxCeleb1	4.48
(Okabe et al., 2018)	TDNN (X-Vector)	TAP	softmax	VoxCeleb1	3.85
(Hajibabaei & Dai, 2018)	ResNet20	TAP	AM-softmax	VoxCeleb1	4.30
(Chung et al., 2018)	ResNet50	TAP	softmax + Contrastive	VoxCeleb2	4.19
(Cai et al., 2018a)	LDE-ASoftmax	LDE	A-softmax	VoxCeleb1	4.56
(Xie et al., 2019)	Thin ResNet34	TAP	softmax	VoxCeleb2	10.48
(Xie et al., 2019)	Thin ResNet34	GhostVlad	softmax	VoxCeleb2	3.22
(An et al., 2019)	ResNet-18+Self-Attention	-	-	VoxCeleb1	9.20
(Hajavi & Etemad, 2019)	UtterIdNet	TDV	softmax	VoxCeleb2	4.26

Throughout this review, we have discussed many different approaches of Speaker Identification implementations and each have shown different performance rates. In Table II the most significant and comparable results are compiled from all the different types of models, aggregation methods and different loss. From the table we can see that the authors of (Xie et al., 2019) have clearly outperformed all the other model types with an error rate of 3.22 percent; this result is clearly produced by their use of GhostVlad aggregation method, where in their second test of using the basic Temporal Aggregation Pooling (TAP) method, their system's performance degraded down to 10.48 percentage of errors. The authors in (Okabe et al., 2018) are closely scoring to the outperforming model by scoring 3.85 percent error rate, in order to apply various weights to different frames by using an attention mechanism and to produce weighted standard deviations within weighted ways. This approach shows that the error rates were significantly reduced from the conventional methods that were used in earlier researches. The rest of the approaches are scoring marginal results except of the results in (Nagrani et al., 2020) and (An et al., 2019), which are using traditional method (Nagrani et al., 2020) and no use of loss functions by An et al. (2019) which resulted in 8.8 and 9.2 percent error rates respectively.

8. Conclusion

In this paper we have reviewed the basics and clarified the progress that has been made in the past couple of years in the Speaker Identification field. It is concluded that the SID can be safely implemented in systems that requires its presence to function, since it is in a state that can be effectively utilized, however, there is still a space for improvements. Although most of the utilized models such as (UtterIdNet proposed in (Hajavi & Etemad, 2019), ResNet20 (Hajibabaei & Dai, 2018), TDNN(X-Vector) (Okabe et al., 2018) and I-Vector (Nagrani et al., 2020)) are trained using the two most prominent datasets (VoxCeleb1 and VoxCeleb2), there could always be more datasets to help and push the field forward.

References

Abbasi, B., & Goldenholz, D. M. (2019). Machine learning applications in epilepsy. *Epilepsia*, *60*(10), 2037–2047. https://doi.org/10.1111/epi.16333

Abdulazeez, A., Salim, B., Zeebaree, D., & Doghramachi, D. (2020). *Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol* (pp. 157–177). International Association of Online Engineering. https://www.learntechlib.org/p/218341/

Abdulqader, D., Mohsin Abdulazeez, A., & Zeebaree, D. (2020). *Machine Learning Supervised Algorithms of Gene Selection: A Review*.

Adeen, N., Mohsin Abdulazeez, A., & Zeebaree, D. (2020). *Systematic Review of Unsupervised Genomic Clustering Algorithms Techniques for High Dimensional Datasets*. https://www.researchgate.net/profile/Diyar_Zeebaree/publication/341119552_Systematic_Review_of_Unsupervised_Genomic_Clustering_Algorithms_Techniques_for_High_Dimensional_Datasets/links/5eafbe5c299bf18b9594945e/Systematic-Review-of-Unsupervised-Genomic-Clustering-Algorithms-Techniques-for-High-Dimensional-Datasets.pdf

Agrawal, A., Gans, J., & Goldfarb, A. (Eds.). (2019). *The economics of artificial intelligence: An agenda*. The University of Chicago Press.

An, N. N., Thanh, N. Q., & Liu, Y. (2019). Deep CNNs With Self-Attention for Speaker Identification. *IEEE Access*, *7*, 85327–85337. https://doi.org/10.1109/ACCESS.2019.2917470

Ashley, K. (2020). *Applied Machine Learning for Health and Fitness: A Practical Guide to Machine Learning with Deep Vision, Sensors and IoT*. Apress. https://doi.org/10.1007/978-1-4842-5772-2

Bacos, C. A. (2020). Machine Learning and Education in the Human Age: A Review of Emerging Technologies. In K. Arai & S. Kapoor (Eds.), *Advances in Computer Vision* (Vol. 944, pp. 536–543). Springer International Publishing. https://doi.org/10.1007/978-3-030-17798-0_43

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv:1409.0473 [Cs, Stat]*. http://arxiv.org/abs/1409.0473

Bargarai, F., Abdulazeez, A., Tiryaki, V., & Zeebaree, D. (2020). *Management of Wireless Communication Systems Using Artificial Intelligence-Based Software Defined Radio* (pp. 107–133). International Association of Online Engineering. https://www.learntechlib.org/p/217853/

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Bhavsar, H., & Ganatra, A. (2012). A Comparative Study of Training Algorithms for Supervised Machine Learning. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*.

Bowles, M. (2015). *Machine Learning in Python: Essential techniques for predictive analysis*. John Wiley & Sons, Inc.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Cai, W., Chen, J., & Li, M. (2018a). Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System. *ArXiv:1804.05160 [Cs, Eess]*. http://arxiv.org/abs/1804.05160

Cai, W., Chen, J., & Li, M. (2018b). Analysis of Length Normalization in End-to-End Speaker Verification System. *ArXiv:1806.03209 [Cs, Eess]*. http://arxiv.org/abs/1806.03209

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., Geis, J. R., Pandharipande, P. V., Brink, J. A., & Dreyer, K. J. (2018). Current Applications and Future Impact of Machine Learning in Radiology. *Radiology*, *288*(2), 318–328. https://doi.org/10.1148/radiol.2018171820

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep Speaker Recognition. *Interspeech 2018*, 1086–1090. https://doi.org/10.21437/Interspeech.2018-1929

Cust, E. E., Sweeting, A. J., Ball, K., & Robertson, S. (2019). Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. *Journal of Sports Sciences*, *37*(5), 568–600. https://doi.org/10.1080/02640414.2018.1521769

Dahake, P. P., Shaw, K., & Malathi, P. (2016). Speaker dependent speech emotion recognition using MFCC and Support Vector Machine. *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, 1080–1084. https://doi.org/10.1109/ICACDOT.2016.7877753

Dey, A. (2016). *Machine Learning Algorithms: A Review*. /paper/Machine-Learning-Algorithms-%3A-A-Review-Dey/56e8863838b4dcc4790108cd1e7e680a104a7c30

Emerson, S., Kennedy, R., O'Shea, L., & O'Brien, J. (2019). *Trends and Applications of Machine Learning in Quantitative Finance* (SSRN Scholarly Paper ID 3397005; Issue ID 3397005). Social Science Research Network. https://papers.ssrn.com/abstract=3397005

Evaluating Data Mining Classification Methods Performance in Internet of Things Applications | Journal of Soft Computing and Data Mining. (n.d.). Retrieved January 10, 2021, from https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/7127

Fernando J. Pineda. (1987). *Generalization of Back-Propagation to Recurrent Neural Networks*.

Ferrer, L., Lei, Y., McLaren, M., & Scheffer, N. (2016). Study of Senone-Based Deep Neural Network Approaches for Spoken Language Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(1), 105–116. https://doi.org/10.1109/TASLP.2015.2496226

Fiebrink, R. (2019). Machine Learning Education for Artists, Musicians, and Other Creative Practitioners. *ACM Transactions on Computing Education*, *19*(4), 1–32. https://doi.org/10.1145/3294008

Fu Jie Huang, & LeCun, Y. (2006). Large-scale Learning with SVM and Convolutional for Generic Object Categorization. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, *1*, 284–291. https://doi.org/10.1109/CVPR.2006.164

Garcia-Romero, D., Zhang, X., McCree, A., & Povey, D. (2014). Improving speaker recognition performance in the domain adaptation challenge using deep neural networks. *2014 IEEE Spoken Language Technology Workshop (SLT)*, 378–383. https://doi.org/10.1109/SLT.2014.7078604

Gary Marcus. (2018). *Deep Learning: A Critical Appraisal*.

Ghoddusi, H., Creamer, G. G., & Rafizadeh, N. (2019). Machine learning in energy economics and finance: A review. *Energy Economics*, *81*, 709–727. https://doi.org/10.1016/j.eneco.2019.05.006

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

Hajavi, A., & Etemad, A. (2019). A Deep Neural Network for Short-Segment Speaker Recognition. *ArXiv:1907.10420 [Cs, Eess, Stat]*. http://arxiv.org/abs/1907.10420

Hajibabaei, M., & Dai, D. (2018). Unified Hypersphere Embedding for Speaker Recognition. *ArXiv:1807.08312 [Cs, Eess]*. http://arxiv.org/abs/1807.08312

Hasan, R., & Rahman, S. (2004). *SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS*. Undefined. /paper/SPEAKER-IDENTIFICATION-USING-MEL-FREQUENCY-CEPSTRAL-Hasan-Rahman/32c4db25607bd52a6d0aeb5498ec9c8d564e6d2e

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hodges, J., & Mohan, S. (2019). Machine Learning in Gifted Education: A Demonstration Using Neural Networks. *Gifted Child Quarterly*, *63*(4), 243–252. https://doi.org/10.1177/0016986219867483

Jahwar, A. F., & Abdulazeez, A. M. (2020). META-HEURISTIC ALGORITHMS FOR K-MEANS CLUSTERING: A REVIEW. *PalArch's Journal of Archaeology of Egypt / Egyptology*, *17*(7), 12002–12020.

Koseler, K., & Stephan, M. (2017). Machine Learning Applications in Baseball: A Systematic Literature Review. *Applied Artificial Intelligence*, *31*(9–10), 745–763. https://doi.org/10.1080/08839514.2018.1442991

Latif, S., Rana, R., Qadir, J., & Epps, J. (2020). Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study. *ArXiv:1712.08708 [Cs, Eess, Stat]*. http://arxiv.org/abs/1712.08708

Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1695–1699. https://doi.org/10.1109/ICASSP.2014.6853887

Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., Gonzalez-Rodriguez, J., & Moreno, P. (2014). Automatic language identification using deep neural networks. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5337–5341. https://doi.org/10.1109/ICASSP.2014.6854622

Lukic, Y., Vogt, C., Durr, O., & Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. https://doi.org/10.1109/MLSP.2016.7738816

Matějka, P., Glembek, O., Novotný, O., Plchot, O., Grézl, F., Burget, L., & Cernocký, J. H. (2016). Analysis of DNN approaches to speaker identification. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5100–5104. https://doi.org/10.1109/ICASSP.2016.7472649

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147. https://doi.org/10.38094/jastt1457

Max Welling. (2010). *A First Encounter with Machine Learning*.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, *60*, 101027. https://doi.org/10.1016/j.csl.2019.101027

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. *Interspeech 2017*, 2616–2620. https://doi.org/10.21437/Interspeech.2017-950

Nakagawa, S., Wang, L., & Ohtsuka, S. (2012). Speaker Identification and Verification by Combining MFCC and Phase Information. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1085–1095. https://doi.org/10.1109/TASL.2011.2172422

Nassif, A., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2019.2896880

Nicolson, A., & Paliwal, K. K. (2020). Sum-Product Networks for Robust Automatic Speaker Identification. *Interspeech 2020*, 1516–1520. https://doi.org/10.21437/Interspeech.2020-1501

Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive Statistics Pooling for Deep Speaker Embedding. *Interspeech 2018*, 2252–2256. https://doi.org/10.21437/Interspeech.2018-993

O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. *ArXiv:1511.08458 [Cs]*. http://arxiv.org/abs/1511.08458

P. Domingos. (2016). A Few Useful Things to Know about Machine Learning. *Communications of the ACM*, *55*(10), Article 10. https://pdfs.semanticscholar.org/c3b6/0802b56eeec611e9def0fdfbcaf42b851b99.pdf

Peiffer-Smadja, N., Rawson, T. M., Ahmad, R., Buchard, A., Georgiou, P., Lescure, F.-X., Birgand, G., & Holmes, A. H. (2020). Machine learning for clinical decision support in infectious diseases: A narrative review of current applications. *Clinical Microbiology and Infection*, *26*(5), 584–595. https://doi.org/10.1016/j.cmi.2019.09.009

Qayyum, A., Latif, S., & Qadir, J. (2018). Quran Reciter Identification: A Deep Learning Approach. *2018 7th International Conference on Computer and Communication Engineering (ICCCE)*, 492–497. https://doi.org/10.1109/ICCCE.2018.8539336

Ravanelli, M., & Bengio, Y. (2019). Learning Speaker Representations with Mutual Information. *ArXiv:1812.00271 [Cs, Eess]*. http://arxiv.org/abs/1812.00271

Richert, W., & Coelho, L. P. (2013). *Building machine learning systems with Python: Master the art of machine learning with Python and build effective machine learning systems with this intensive hands-on guide*. Packt Publ.

Rui Lu, Zhiyao Duan. (2017). BIDIRECTIONAL GRU FOR SOUND EVENT DETECTION. *Detection and Classification of Acoustic Scenes and Events*.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shon, S., Tang, H., & Glass, J. (2018). Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model. *ArXiv:1809.04437 [Cs, Eess]*. http://arxiv.org/abs/1809.04437

Song, Y., Jiang, B., Bao, Y., Wei, S., & Dai, L.-R. (2013). I-vector representation based on bottleneck features for language identification. *Electronics Letters*, *49*, 1569–1570. https://doi.org/10.1049/el.2013.1721

The intellectual challenge of CSCW: the gap between social requirements and technical feasibility: Human-Computer Interaction: Vol 15, No 2. (n.d.). Retrieved January 10, 2021, from https://dl.acm.org/doi/10.1207/S15327051HCI1523_5

THE USE OF MACHINE LEARNING IN HIGHER EDUCATION. (2019). *Issues In Information Systems*. https://doi.org/10.48009/2_iis_2019_56-61

Tirumala, S. S., & Shahamiri, S. R. (2017). A Deep Autoencoder approach for Speaker Identification. *Proceedings of the 9th International Conference on Signal Processing Systems - ICSPS 2017*, 175–179. https://doi.org/10.1145/3163080.3163097

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, 1096–1103. https://doi.org/10.1145/1390156.1390294

Wang, Y., & Lawlor, B. (2017). Speaker recognition based on MFCC and BP neural networks. *2017 28th Irish Signals and Systems Conference (ISSC)*, 1–4. https://doi.org/10.1109/ISSC.2017.7983644

Xie, W., Nagrani, A., Chung, J. S., & Zisserman, A. (2019). Utterance-level Aggregation for Speaker Recognition in the Wild. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5791–5795. https://doi.org/10.1109/ICASSP.2019.8683120

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., & Wang, C. (2018). Machine Learning and Deep Learning Methods for Cybersecurity. *IEEE Access*, *6*, 35365–35381. https://doi.org/10.1109/ACCESS.2018.2836950

Zeebaree, D., Abdulazeez, A., Zebari, D., Haron, H., Nuzly, H., & Malaysia, T. (2020). Multi-Level Fusion in Ultrasound for Cancer Detection Based on Uniform LBP Features. *Computers, Materials and Continua*, *66*, 3364–3382. https://doi.org/10.32604/cmc.2021.013314

Zeebaree, D., Haron, H., Mohsin Abdulazeez, A., & Zebari, D. (2019). *Machine learning and Region Growing for Breast Cancer Segmentation* (p. 93). https://doi.org/10.1109/ICOASE.2019.8723832

Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M. (2018). Gene Selection and Classification of Microarray Data Using Convolutional Neural Network. *2018 International Conference on Advanced Science and Engineering (ICOASE)*, 145–150. https://doi.org/10.1109/ICOASE.2018.8548836

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019). Trainable Model Based on New Uniform LBP Feature to Identify the Risk of the Breast Cancer. *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 106–111. https://doi.org/10.1109/ICOASE.2019.8723827

Zhao, X., Wang, Y., & Wang, D. (2014). Robust Speaker Identification in Noisy and Reverberant Conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(4), 836–845. https://doi.org/10.1109/TASLP.2014.2308398

Cite this article:

Lawchak Fadhil Khalid & Adnan Mohsin Abdulazeez (2021). Identifying Speakers Using Deep Learning: A review. ***International Journal of Science and Business****, 5*(3), 15-26. doi: https://doi.org/10.5281/zenodo.4481596

Retrieved from http://ijsab.com/wp-content/uploads/682.pdf

Published by