Multimedia Tools and Applications (2024) 83:24541–24562
https://doi.org/10.1007/s11042-023-16507-2

An optimized enhanced-multi learner approach towards speaker identification based on single-sound segments

Seyed Reza Shahamiri[1]

Received: 21 December 2021 / Revised: 28 July 2023 / Accepted: 9 August 2023 /
Published online: 17 August 2023

Abstract
Speaker Identification (SI) is the task of identifying an unknown speaker of an utterance by comparing the voice biometrics of the unknown speaker with previously stored and known speaker models. Although deep learning algorithms have been successful in different speech and speaker recognition systems, they are computationally expensive and require considerable run-time resources. This paper approaches this issue by proposing an optimized text-independent SI system based on convolutional neural networks (CNNs) that not only delivers accuracies on par with state-of-the-art benchmarks but also demands significantly fewer trainable parameters. The proposed system integrates an Enhanced Multi-Active Learner framework, which distributes the complexity of the learning task among an array of learners, with a novel SI approach in which speakers are identified based on a single sound segment of voice biometrics. Here, experiments were conducted with all 1881 VoxCeleb 1 and TIMIT speakers, and results were compared with the SI systems reported in the literature that were assessed on the same speakers' data. Results indicate that first, the proposed system outperformed the benchmark systems' performances by delivering up to 2.43% better top-1 accuracy, and second, it reduced the number of deep learning trainable parameters by up to 95%. The proposed SI could bring offline, large-scale speaker identification to low-end computing machines without specific deep learning hardware and make the technology more affordable.

Keywords Automatic Speaker Identification · MFCC · Deep neural networks · Optimization

1 Introduction

Speaker Recognition (SR) is the process of identifying speakers based on vocal features (aka voice biometrics) of their given speech samples, whereas speech recognition confines to the content recognition process rather than the speaker [29]. SR tasks include speaker verification, speaker diarization [1], speaker de-identification, etc., in which

✉ Seyed Reza Shahamiri
admin@rezanet.com

[1] Department of Electrical, Computer, and Software Engineering, Faculty of Engineering, University of Auckland, Auckland, New Zealand

using SR to identify an unknown speaker from a set of stored, known speaker models is called Speaker Identification (SI). Mainly, SI is the process of comparing unknown user voice biometrics against many known biometric profiles and finding the best or exact match. Among the most prominent applications of speaker identification are mail automation tasks, automated labeling of speakers of a conversation, user identification and authorization, and acoustic forensics.

Deep learning advances have enabled Deep Neural Networks (DNNs) to produce accurate speech and speaker recognition systems. However, neural network-based models require adjusting and processing many neural weights and bias parameters that make them computationally expensive and complicated, requiring high-end computing machines with powerful processing units [33]. Hence, their applications for lower-capacity smart devices are limited because of the required run-time resources when large-scale applications are intended. One solution is to implement the SI DNN engine on a powerful server and deliver its services via the cloud or over a network (i.e., online SI), or employ other SI engines that require less memory and computation footprint, such as rapid i-vectors [42]. The problems with the first solution are the additional cost of the server, the availability of the network, and security and privacy concerns with respect to transmitting the authentication data over the network. Likewise, the second solution does not benefit from DNNs' strong generalization capabilities. On the other hand, an offline DNN-based SI with a small memory and computation footprint that still delivers high accuracies can address these shortcomings.

Most traditional approaches to SI include modeling all speaker voice biometrics via one classifier (aka *learner*), meaning that the learner is responsible for storing and providing all speaker models [17]. In active learning theory, this is known as the *Single-Learner (SL)* paradigm. A supervised active learning framework called *Multi-Learner* was proposed by [39] that recommended accuracy improvements for pattern recognition tasks comprising two or more *views*. An example of views is describing an image by its visual features (i.e., view_1) and the words surrounding it (i.e., view_2) in a web image retrieval system. Another example is web page classification, where a page can be classified by the words on the page and the words on other pages linked to this page, where the former is considered view_1 and the latter view_2. If there are several learners to approximate the views, the learning theory is known as *Multi-View Multi-Learner (MVML).*

Multi-learner frameworks consider additional views as supplementary data to help improve the performance of the main view (i.e., the first view in the examples above). They employ these additional views as meta-data to provide more information about the main view. It is pertinent to note that the primary objective of multi-learners remains approximating the first view, while the additionally created views are only considered to improve this objective. Nevertheless, incorporating such extra views by multi-learner frameworks makes the approximation task more complicated as there is more information for the learner(s) to process and learn. Considering DNNs as learners means MVML requires additional DNNs in order to model the supplementary views to assist in approximating the main DNN; this increases the required computational resources significantly, but superior approximation accuracy can be achieved in comparison to *Multi-View Single-Learner (MVSL)* models.

An improvement to multi-learner, *Enhanced Multi-Active Learner (EMAL)*, redefines views and describes how they are perceived and modeled via several learners [32]. Contrary to MVML, in which the objective is to improve the approximation of the main view by considering other views and modeling them via supplementary learners, EMAL aims to distribute the complexity of the main view among several learners. Via EMAL, each

learner is responsible for learning an aspect of the main view without increasing the size, number, or complexity of the view or function. EMAL achieves this by breaking down the main view into smaller, less complex views and then distributing the modeling of these new views among the learners.

In terms of DNNs, the EMAL simplification of the main view may also streamline the DNN structural and computational complexity since the overall simulation/approximation task assigned to each DNN in the network is also simplified; hence, the DNN requires fewer parameters as the result of dealing with less approximation complexity. This architectural simplification can help to rectify issues of long training time and high computational complexity associated with DNNs. Nevertheless, EMAL applications in SI have remained unknown as the traditional MVSL approaches profoundly dominate the SI research community.

Furthermore, the common SI approach employs a combination of acoustic features extracted from sequential sounds to present the speakers' voice biometrics. Notably, 5 to 10 sound segments are stacked to form the input before and after the current segment [28]. Stacking sound segments is particularly important for speech recognition technologies as words are broken down into sounds (or phonemes and phones), and successful identification of each word depends on recognizing the previous sounds. Nevertheless, the discriminative acoustic information used to identify a speaker is mostly embedded in how speech is uttered and not necessarily in the content of speech [15]. Likewise, the objective of text-independent SI is to identify the speaker and not the content of speech. Thus, our proposed system assumes the distinctive speaker features in an acoustic sound segment contain enough speaker-distinguishable information to identify the speaker regardless of the previously uttered sounds. Identifying a speaker using a single-sound segment instead of a stack of sounds can further reduce the complexity of DNN learner(s) as the input dimension can significantly be reduced, which means the classifier requires to learn a smaller number of acoustic features. While our initial study shows the single sound SI approach has merits for smaller scale, straightforward SI tasks [34], it is essential to investigate the effects of such reduction of input dimensions on a larger scale, realistic and challenging SI tasks.

The objectives of this paper are to propose an optimized system for text-independent, closed-set speaker identification based on Convolutional Neural Networks (CNNs) that leverages both EMAL and single sound approach and benchmark its performance and DNN computational complexity. We intend to simplify the complexity of SI yet utilize the powerful knowledge discovery and generalization capabilities of CNNs so that not only an accurate SI is achieved but also the smaller and fewer complex CNNs reduce the computational costs associated with deep learning-based SI systems. The proposed system applies the EMAL approach to SI to improve the efficiency of speaker identification tasks and integrates the single-sound segment SI approach to decrease the SI complexity. Its performance and DNN computational complexity were measured when it was applied to identify 630 speakers whose utterances were provided by Texas Instruments/Massachusetts Institute of Technology (TIMIT) Acoustic-Phonetic Continuous Speech Corpus [14], and 1251 celebrities provided by VoxCeleb [25]. Here, the number of trainable deep learning parameters is considered the metric to measure the computation complexity of the proposed SI system. A comparative performance study with DNN-based SI systems that employed the same datasets is also provided.

The rest of the paper is organized as follows: Section 2 reviews the related SI works. The proposed system is explained in Section 3, and Section 4 describes the experiments. Finally, Section 5 discusses the results and presents the comparative study, while Section 6 concludes the paper.

2 Related work

A typical speaker identification system goes through an enrollment procedure in which speaker models are created and stored by the learner, and a matching procedure that explores the modeled speakers for a match [15]. Both enrollment and matching procedures are initiated by receiving the speech input signals, extracting acoustic features that represent the speech parameters in a form understandable by the system, and then these features are fed to the learner/classifier. Hence, the feature extractor and learner are the two primary components of SI systems.

As an acoustic feature extraction approach, Mel-Frequency Cepstral Coefficient (MFCC) is a practical approach to decomposing an acoustic signal into its phone or sounds and presenting them via frames of acoustic features that can be modeled via various machine learning algorithms. MFCC features have been widely applied in various speech processing tasks, such as impaired speech recognition and heart anomaly detection [19]. Regarding SI, MFCC and its variations are the most frequently applied feature extraction approach. Current speaker identification systems, especially those that leverage deep learning algorithms to store and model the speakers, employ a combination of stacked acoustic features extracted from sequential acoustic frames to present the speakers' voice biometrics, as explained in the previous section.

2.1 i-vector based Speaker Identification

In terms of the SI learner, Gaussian Mixture Models (GMMs) were traditionally one of the most popular methods in different speaker recognition tasks. The first published application of GMM in SI is [27] in which GMMs were used to provide a multi-classification statistical model of speakers' data by modeling the distribution of the speaker's MFCCs. Recent improvements to GMM-based SI systems are i-vector based approaches that employ 0th, first, and second-order statistics produced by GMMs.

Today, one of the most successful classes of learning algorithms in speech processing is deep learning. Recent deep learning implementations for SR highlight that the complexity involved in SR requires special attention compared to traditional pattern recognition problems [28]. In terms of SI, deep learning-based SI approaches have outperformed other traditionally popular ones on large-scale applications, such as i-vector based approaches [40, 43]. This is because deep neural nets learn features discriminatively instead of the generative approach that GMM/i-vector frameworks apply [35]. For example, Chen et al. [8] proposed a bilevel SI framework that used sparse coding with no gaussian assumption to improve i-vectors, and employed a softmax and linear Support Vector Machine. Evaluated on VoxCeleb 1 dataset, the authors achieved the highest top-1 accuracy of 67.2%, which is lower than the DNN performances reported in the literature on the same dataset.

2.2 Deep learning based speaker identification

An example of employing deep learning in SI is utilizing Restricted Boltzmann Machines (RBM) to assist feature extraction in a hybrid noise-robust SI model proposed by [46]. The study evaluated the model by providing speech data infected by factory noise, destroyer engine room noise, and speech shape noise. Three sets of models based on gammatone features, gammatone frequency cepstral coefficients, and MFCCs were generated for speakers

selected from the NIST Speaker Recognition corpus and reported that MFCC-based models achieved the best benchmarks in most experiments. An investigation into the effects of depth and layered-wise training using deep autoencoders (DAEs) in SI was also conducted in [41].

Recently, deep CNN approaches to identify speakers have proven to be very useful because of CNNs' effectiveness in modeling real-world, noisy data without any requirements of specific features engineering [16, 38]. In this respect, [2] proposed a SI system using a deep CNN that was verified using connected speech samples provided by TIMIT. The network employed 32 and 64 filters for its convolutional layers, each followed by another layer to perform max-pooling, and the outputs of these filters were fed to two dense layers. Similarly, Nagrani et al. [25] adopted VGG-M CNN architecture [7] and designed a closed-set SI model verified on noisy speech samples collected from 1251 celebrities in the VoxCeleb 1 corpus. Another example of large-scale CNN-based SI is [4] in which VGG and Residual Neural Networks were studied.

2.3 Hybrid and ensemble-based speaker identification

Ali and his colleagues [2] presented a hybrid SI approach that employed different learners to perform different tasks in the SI pipeline. Particularly, feature extraction was done by applying a Deep Belief Network (DBN) to extract unsupervised features and then combined with MFCCs. Principal Component Analysis (PCA) was applied for the linear transformation of features. Then, the features were pipelined through multiple learning algorithms, including RBMs, K-Means, and Support Vector Machines (SVMs), and finally mapped to different speaker models. Similarly, a DBN-GMM SI was proposed and verified on a custom corpus by [40]. Another hybrid approach is [36] in which a GMM-DNN SI approach was proposed to improve the performance of identifying speakers' emotions. The authors delivered better performance than conventional perceptron neural networks when they applied their solution to a customized Arabic dataset.

Based on document classification's hierarchical attention network (HAN) [44], Yanper et al. [37] proposed another hybrid SI called H-Vectors. This approach aimed to find which segments of an utterance contribute more towards identifying the speaker. The proposed HAN architecture was composed of three components: (1) a frame-level encoder and attention layer consisting of a CNN, a Gated RNN (GRU) [9], and an MLP, (2) a segment-level encoder that includes another CNN and MLP, and (3) a dense fully connected DNN with two layers. They verified this architecture with NIST SRE 2008 part1 (SRE08), Call Home American English Speech (CHE), and Switchboard Cellular Part 1 (SWBC) datasets achieving accuracies of up to 98.5%, 92.8%, and 86.2% for each dataset respectively.

An ensemble neural networks approach to SI using Probabilistic Neural Networks (PNNs), General Regression Neural Networks (GRNNs), and RBFs was studied by [3] in which each neural net was responsible for probing the data differently to fit the training audio features. In particular, one network was trained on all the data, the other network was only trained on the data showing a margin of error, and the other was trained using the data with no error margin. The model was evaluated on GRID speech corpus and showed improvements in recognition time and accuracy over traditional approaches. Section 5 provides a performance comparison of the SI systems mentioned above.

The literature does not report any EMAL implementation of speaker identification, where the single-learner concept remains the dominant approach among SI researchers. Thus, it is important to investigate whether EMAL benefits are achievable in SI and what advantages EMAL active learning offers.

It is also essential to highlight the difference between SI approaches that use different types and numbers of learners to perform different views or tasks in the SI pipeline

(for example, [2, 27]), or repeat modeling the same view(s) using different variations of learners (such as Bagging ensemble learning [26]), and EMAL based SI systems. EMAL systems use a network of learners to improve the performance of learning the main view by distributing the main view's complexity among several learners. Notably, each learner in an EMAL-based SI has a specific responsibility that is different from other learners (i.e., no redundancy of views) and contributes towards performing a different aspect of the overall task, whereas in bagging ensemble learning each view is modeled several times using a different machine learning algorithm. To put it differently, EMAL stores only one model of each view, but ensemble learning stores several models of each view. The literature usually refers to ensemble methods as the collection of learners that are variations of the same learner. Likewise, a broader category is known as multi-classification systems in which the hybridization of different learners is considered [12].

As an illustration, the example above of using ensemble learning in SI [3] applied three different types of neural networks, while each neural net represented all speaker models (i.e., all of the views); thus, there were three variations of speaker models. Given an utterance, the final SI outcome was calculated by majority voting (bagging), where the speaker model that obtained 2/3 of the votes was selected as the identified speaker. The next section explains EMAL SI in detail.

Regarding single-sound segment SI, we recently investigated whether this approach to SI is feasible and to identify the best parameters and MFCC configuration [34]. We showed that speaker identification systems could operate by relying on the distinctive acoustic features that an individual segment of speech presents (such as an MFCC frame) without relying on previous speech segments. In our previous study, we conducted more than one hundred experiments in which small MVSL SI systems were created using dense, fully connected neural networks considering different SI parameters. The initial results indicated that speaker identification using one sound segment is possible, and results are on par with the traditional stack-based SI approaches where the number of speaker models is small. We applied the previous study's findings to set the acoustic feature parameters in the present study, as stated in the next section. Nevertheless, the neural network used in the previous study and MVSL active learning provided poor results when speech samples of all TIMIT speakers were given. Thus, the present study focused on transferring the knowledge from the previous experience to design an EMAL single-sound segment SI system that can handle complex SI tasks yet reduces DNN trainable parameters.

3 The proposed system

3.1 Formulating speaker identification

Suppose the speaker identification task is declared as a function approximating speaker models S by receiving a speech sample $x \in X$, i.e., x is one of the input speech signals from set X. Based on the number of speakers to be identified, S is composed of multiple speaker models:

$$S = \{s_1, s_2, \cdots, s_n\} \tag{1}$$

where n is the number of speakers, and s_i is the ith speaker model ($i = 1$ to n). Given a speech sample x obtained from one of S speakers, speaker identification S^* can be defined as the mappings of X to individual speaker models in Eq. 1 by finding the most probable match as stated by Eq. 2:

$$S^* = \underset{S}{\operatorname{argmax}}\, P(S|X) \tag{2}$$

3.2 Features extraction

The first step in speaker identification is to prepare the input signals X and extract their distinctive acoustic features in a feature extraction process. In either enrolment (aka training) or matching phase (aka inference), the speech samples must be pre-processed to remove speech frames representing silence since such frames may confuse the learners. Silence frames tend to be similar regardless of speakers and do not contain discriminative speaker-dependent data. The proposed system uses 20ms segments (aka frames); thus, any silence data equal to or greater than 20ms should be removed from input utterances. Alternatively, an additional speaker model can be created to model silence segments.

It is important to select a features extraction method that best presents the acoustic characteristics of signals X because S^* refers to these features to associate X with each speaker model s_i. Among different acoustic features extraction methods, MFCC is constructed using frequencies of the vocal track and present acoustic signals in the cepstral domain that employs FFT to represent windowed short signals as the real cepstrum of X. MFCCs are inspired by our natural auditory perception mechanism; hence MFCC frequency bands are spaced equally on Mel scale [5]. Although MFCCs ignore some acoustic information, they still preserve sufficient distinguishable data [10]. This attribute of MFCCs has been widely used in speech and speaker recognition tasks making them one of the most popular acoustic feature extraction methods.

Applying MFCCs to an input speech signal x results in a 2D tensor of Acoustic Features AF where columns are frames representing sound segments, and rows are MFCC coefficients for each frame. Thus, S^* becomes the mapping of AF to speaker models S:

$$S^* = \underset{S}{\operatorname{argmax}}\, P(S|AF) \tag{3}$$

3.3 The single-sound segment approach

Acoustic Features AF in Eq. 3 is a 2-dimensional matrix of sequential sound segments extracted from x that are represented as frames (aka segments) of MFCC features; each segment is a row in AF matrix that may present a sound or phone. Although this 2D representation of utterance x is vital to identify its content and requires processing multiple segments, identifying speech content is not the objective of speaker identification. Thus, the single-sound segment SI approach assumes one frame of speech features (i.e., features corresponding to an individual sound segment) contains enough information to distinguish between speaker models S [34]. Hence, in the proposed system, the feature extraction process presents the speaker's voice biometrics as a single MFCCs frame AF', which means speaker identification S^* relies only on the information provided by AF' to approximate speaker models S without considering the frames before or after AF'. As such, the proposed system redefines speaker identification as given by Eq. 4:

$$S^* = \underset{S}{\operatorname{argmax}} P(S|AF) \tag{4}$$

It is important to note that AF' in Eq. 4 is a *1D* tensor in comparison to *2D* tensor AF, and its dimension is considerably smaller, as stated by Eq. 5:

$$|\mathrm{AF}| = \frac{|\mathrm{AF}|}{\mathrm{k}}, k = \text{the number of } AF \textit{ soundsegments} \tag{5}$$

In this study, each speech frame contains 60 MFCCs, meaning each AF' tensor is composed of 60 coefficients.

3.4 Enhanced multi-active learner

The approximation task in Eq. 4 can be done by one or many learners L:

$$L = \left(l_1, l_2, \cdots, l_m\right) \tag{6}$$

where m is the number of learners. The traditional neural net-based SI approaches employ a single learner (i.e., $m = 1$) to approximate S^* by applying a single-learner paradigm. The complexity of this task requires a neural net with many neurons and parameters to learn the acoustic features of all speaker models s_i. On the other hand, by applying EMAL, the complexity of S^* can be distributed among a network of learners by setting $m > 1$. This distribution decreases the complexity of the required neural network learners and may also improve the classification performance as the number of positive-response acoustic features being modeled by a learner can be reduced. A positive response for the ith speaker refers to any acoustic frame AF' that was extracted from any speech signal x uttered by speaker s_i.

Suppose:

$$AF'_{i,j} = (af'_{i,j_1}, af'_{i,j_2}, \ldots, af'_{i,j_{60}}) \tag{7}$$

where $AF'_{i,j}$ is the jth acoustic segment for speaker model s_i extracted from a speech signal uttered by the ith speaker, and contains the 60 extracted MFCC features. AF' contains many samples of $AF'_{i,j}$ to present how the ith speaker pronounces different sounds, as shown by Eq. 8:

$$D(AF'_i) = D\left(af'_{i,j_1}\right) \times D(af'_{i,j_2}) \times \cdots \times D\left(af'_{i,j_{60}}\right), \quad \textit{for } i = 1 \textit{ to } n, \text{and } j = 1 \text{ to } p \tag{8}$$

where p is the total number of sound samples for the ith speaker, and $D\left(af'_{i,j_y}\right)$ is the number of all possible yth MFCC coefficients ($y = 1$ to 60) extracted from the ith speaker's jth sound segment. In simple terms, $D(AF'_i)$is the number of all MFCC features extracted from all sound samples that S^* requires processing to approximate speaker model s_i in Eq. 4, and $D(AF')$ is the same but for all speaker models S from Eq. 1.

In an EMAL approach to SI, a network of learners L is used to learn AF' where each learner l_i is responsible for approximating s_i. Hence, l_i only needs to learn features presented in AF'_i for its positive responses, whereas single-learner SI systems use one learner l_1 to learn AF'. It is pertinent to note that $D(AF'_i)$ for positive responses is considerably smaller than $D(AF')$ because the number of learnable speech features per s_i is less than S as denoted by Eq. 9:

$$D(AF'_i) = \frac{D(AF')}{\mathrm{o}}, \quad o = \text{the number of sounds samples associated with } s_i \tag{9}$$

To put it differently, each learner in an EMAL-based speaker identification system only learns acoustic features directly associated with one speaker instead of all speaker features. Combining the reduction in the number of learnable coefficients provided by Eqs. 5 and 9, it can

be concluded that each EMAL learner l_i needs to process significantly fewer speaker-dependent features belonging to its class than a single-learner SI. In particular, each neural network l_i acts as a binary classifier that decides whether a given AF' vector is associated with speaker s_i (that is responsible to model) since it only needs to map the input frame vector to s_i instead of S.

3.5 The single-sound EMAL-based system

Figure 1 depicts the proposed system in which $n = m$ (n is the number of speaker models, and m is the number of learners), meaning each speaker is modeled by an individual learner. After silence segments are removed from X (when necessary), each utterance is presented by several sound segments of 60-dimensional MFCCs indicated by AF'. Each $AF'_{i,j}$ (Eq. 7) needs to be appropriately labeled with its associated speaker s_i and stored to be used for training. The proposed system employs a network of CNN learners L to learn the voice biometrics of speakers where each l_i associates itself with one of s_i speakers, which means for n number of speaker models, n learners are required.

Since each learner performs a binary classification, it only needs one sigmoid output neuron that calculates the probability of the given $AF'_{i,j}$ representing speaker model s_i. During the training phase, each $AF'_{i,j}$ is individually given to all learners; the learner that is responsible for the speaker in which her voice biometrics frame is given has its output neuron set to one (y_{max}) while the remaining learners receive a zero (y_{min}) for the output neuron to show that the given sound segment does not belong to the speaker they represent.

Particularly, let us assume $i = 1$ in Fig. 1, which means $AF'_{1,j}$ is one of the sound segments representing the first speaker model s_1. In this case, the target vector for l_1 is set to

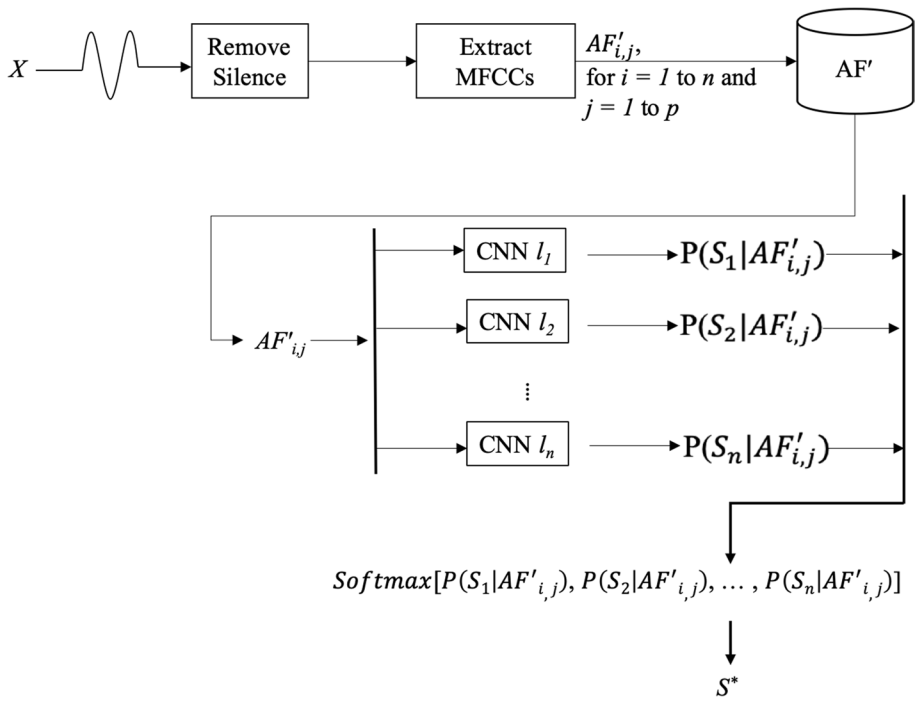

Fig. 1 The proposed system

one (positive response), while the rest of the l_i learners ($i \neq 1$) receive a zero for their target vectors (negative response) while speaker enrollments are being performed. Next, $i = 2$ meaning that $AF'_{2,j}$ is a sound associated with s_2, so l_2 receives a one for its target vector and zero for the remaining learners. This process continues by providing each sound sample for each i and j per training epoch.

Nevertheless, this labeling strategy results in an imbalanced training dataset used to train CNN l_i because there are only $\frac{1}{n}$ positive training samples against $\frac{n-1}{n}$ negative samples (n is the number of speaker models). This problem can be resolved by increasing the class weight of positive-response training samples and assigning smaller weights to other samples during speaker enrollment (i.e., training), as explained in [11]. This approach instructs the l_i optimizer to pay increased attention to sound segments extracted from speech samples uttered by the ith speaker.

Once the training procedure is complete, given an unforeseen speech frame (with all silence traces removed, if required), the EMAL-SI should be able to highlight speaker-distinguishable acoustic features presented by an unforeseen sound segment and relate them to one of the speaker models. This is done by feeding the unidentified sound segment to all learners and querying them to relate the data to the speaker model they present. Each CNN output is the likelihood that the given sound segment is uttered by the speaker the CNN is responsible for. The learners' outputs are then given to a softmax function to be squashed as probability distributions. By finding the maximum of the softmax function results, it can be determined which CNN provides the highest probability.

Given the proposed system is a closed-set speaker identification system, it requires all speakers/users to be enrolled in the system in advance. In case there are new speakers, new learners have to be added to L to match the number of learners with the number of speakers, and all learners need to be re-trained following the process explained here.

4 Experimental setup and results

This section describes the experimental setup, evaluation methodology, and results.

4.1 Datasets

Our experiments were conducted over two corpora. The first corpus was TIMIT which contains phonetically rich, clean speech samples obtained from 630 speakers in which ten utterances of each speaker are provided. The research community has widely used the dataset in different speech processing tasks. We considered speech materials from all 630 TIMIT speakers to verify the proposed system. All ten utterances per speaker were employed, of which eight were used to provide the training sound segments and the remaining two for extracting the test segments. Conducting 10-fold repeated random sub-sampling validation [6], the training and testing utterances were changed for each fold randomly.

The second dataset, VoxCeleb 1 [25], contains more than 153 K gender-balanced utterances collected from 1251 celebrities captured from videos uploaded to YouTube. Overall, the dataset contains 352 h of speech. The participants in this dataset were 55% male, and speech samples were collected from speakers with a diverse range of ethnicities, professions, ages, and accents. The audio samples reflect real-world environments, including

utterances with background chatter and music, room reverberations, laughter, etc. There are an average of 116 utterances per speaker, and the average length of each sample is 8.2 s.

Similar to the baseline experiment conducted on this dataset [9], we used around 145 K utterances for training and 8 K for testing, including all speakers provided in the corpus. We did not consider any cross-validation procedure similar to the baseline system since any experiment considering VoxCeleb full dataset is significantly resource-consuming.

Comparing TIMIT and VoxCeleb 1, the latter increases the complexity of speaker identification because (1) VoxCeleb audio samples include different types of background noise profiles, and (2) the number of speakers is almost twice larger than TIMIT.

4.2 Evaluation criteria

The performance of the proposed system was measured using two criteria. The first criterion was speaker identification accuracy (aka top-1 accuracy), as the proportion of correct identification of speakers based on the testing sound segment data. Accuracy conveys the practicality of S^* to identify speakers based on the given acoustic sound segment during testing.

The second criterion was Normalized Root Mean Squared Error (NRMSE):

$$NRMSE = \frac{Root\ MSE}{y_{max} - y_{min}} = \frac{\sqrt{MSE}}{1 - 0} \tag{10}$$

where y_{max} and y_{min} were one and zero, respectively, as explained in Section 3. NRMSE was considered to measure the system's error rate in terms of how close the results generated by the SI were to the target results. In particular, lower NRMSE implies S^* is more capable of making precise predictions. It is pertinent to note that NRMSE was calculated based on the results obtained from all sigmoid output neurons and before the softmax function was applied.

4.3 Experimental setup

All experiments were implemented in Python. Feature extraction was done via Python Speech Feature Extraction library [24], and the CNN learners were implemented on Google's TensorFlow framework.

For TIMIT experiments, another Python library called Pydub [11] was used to automatically remove any trace of silence from speech utterances before feature extraction was performed. The silence threshold was set according to Decibels Relative to the Full Scale of each utterance. Silence removal was not necessary for the VoxCeleb experiment as the audio samples in the dataset did not include traces of silence.

4.4 CNN architecture

The convolution setup of the CNNs was inspired by [23], also used in [23]. Each CNN l_i comprised two convolutional layers with 32 and 64 filters, respectively, followed by a max-pooling layer for down-sampling the feature maps after the individual convolution

layers. Nevertheless, we did not apply the standard 2D windows on feature maps since the input sound segments were 1D tensors of 60 MFCCs rather than the 2D tensors of multi-frame MFCCs. In particular, both convolution layers applied a 3·1 window to the feature maps, and down-sampling was done by a kernel of size 4·1. The convolutional layers had no strides (i.e., 1·1), while max-pooling strides were 2·2.

Identifying the dense layers hyperparameters, initial experiments on a small subset of the dataset (20 TIMIT speakers - ten females, ten males) were conducted in advance. Next, the hyperparameters were tuned and selected by a grid search algorithm [20], where 2 to 4 dense layers with 32, 64, and 128 neurons and different activations were trialed. Then, the EMAL CNN architecture that provided the best accuracy was selected and applied to the full datasets. As a result of the grid search algorithm, each CNN l_i architecture was selected, as shown in Fig. 2 - the remaining hyperparameters are provided in Table 1.

During the TIMIT experiment, the training data were shuffled after every run to apply cross-validation, and batch training was stopped when a loss lower than 0.05 or 300 epochs were achieved. EMAL SI needed 630 of the CNNs shown in Fig. 2 to model TIMIT and

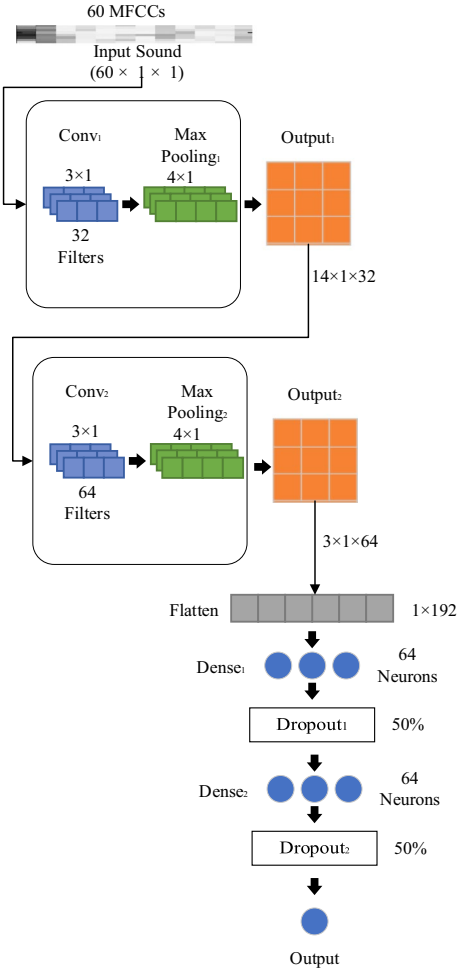

Fig. 2 CNN *li* architecture

Table 1 DNN architecture and hyperparameters

Layers	Activation Function	Dropout	Other parameters
Conv layers	Relu	NA	*Optimizer*: Adam *Learning rate* α: 0.0001 *Loss Function*: Binary Cross Entropy *Mini-batch*: 256
Dense 1: 64 neurons	Elu	50%	
Dense 2: 64 neurons	Elu	50%	
Output: one neuron	sigmoid (to perform binary classification)	NA	

1251 CNNs to model VoxCeleb 1, one per s_i, and they all followed the same architecture and hyperparameters. Accordingly, each CNN needed to adjust 22,913 trainable parameters. In order to avoid overfitting, dropout regularization with a relatively high drop rate of 50% was applied to each fully connected layer of the CNNs.

4.5 Results

The evaluation results of experimenting with the proposed SI system are shown in Table 2. Please be noted that the training loss in each row is the average value obtained from all 630 CNNs in that fold for the TIMIT experiments.

5 Discussion

Accuracy and NRMSE standard deviations for the TIMIT experiments were 0.26% and 0.94%, respectively. According to Table 2, the accuracies and NRMSEs obtained during these experiments followed normal distribution since 70% of the observations were between one standard deviation above or below the mean, and all observations fell between

Table 2 The proposed SI system experimental results

Corpus	Folds	Training Loss (cross-entropy)	Testing accuracy (%)	Testing NRMSE (%)
TIMIT	1	0.0476	98.46	11.02
	2	0.0661	97.80	13.31
	3	0.0565	98.21	11.89
	4	0.0653	97.82	13.19
	5	0.0429	98.48	10.91
	6	**0.0374**	**98.61**	**10.40**
	7	0.0567	98.23	12.29
	8	0.0506	98.30	11.61
	9	0.0399	98.52	10.79
	10	0.0496	98.31	11.55
	Mean	*0.0452*	*98.27*	*11.69*
	Median	*0.0452*	*98.30*	*11.58*
VoxCeleb	NA	0.1092	82.93	18.26

Italic are mean and medain values, and bold are the best results achieved

positive or negative two standard deviations concerning the mean. Normal distribution confirms the reliability of the results obtained.

The low NRMSEs in Table 2 show the proposed SI predictions were close to the targets that imply the robustness of using a single sound segment to identify speakers. Similarly, the high accuracies indicate the applicability of using a single sound to perform complex SI tasks.

5.1 Comparative study

To highlight the advantages of the proposed system, a comparative study with state-of-the-art neural network-based speaker identification systems published in the literature is presented in Table 3. Additionally, we selected the SI systems explained in [23] and [25], highlighted in Table 3, for direct benchmarking since they both adopted a deep CNN-based approach, reported results based on the same corpora, and considered the same number of speakers we used in our experiments. This selection enables us to benchmark our proposed SI system with theirs directly. The rest of the shown SI systems employed a significantly smaller number of speakers or were applied to different datasets. Reducing the number of speakers decreases the complexity of the SI task; hence a direct accuracy comparison of large- and small-scale SI systems is not fair since they belong to different complexity classes.

It is pertinent to note that deep learning-based speaker verification (SV) systems were not included in this table for two reasons. First, the objectives of SV and SI are different. Mainly, SV's objective is to apply voice biometrics to verify users' claimed identity, while SI intents to find the closest match of an unlabeled speech sample to a set of stored speaker models [15]. Although the enrollment process is similar in both tasks, testing is different, as explained in [42]. Second, a direct comparison of SV and SI systems is impossible as they use different performance indicators. In particular, using Equal Error Rate (EER), a metric that looks at false positive and negative ratios, is mostly considered to measure the performance of speaker verification systems, whereas accuracy is the typical performance indicator for speaker identification systems [15].

In the first baseline system [23], TIMIT speech data were presented to a CNN as one-second spectrograms of $128 \cdot 100$ pixels. The CNN consisted of two convolutional layers with 32 and 64 filters, respectively (similar to the CNNs used in our study), each followed by a max-pooling layer (pooling size was $4 \cdot 4$ and stride was $2 \cdot 2$). The convolutional layers were stacked into two dense layers with 6300 and 3150 neurons. The output layer consisted of 630 softmax neurons (one neuron per speaker). Similar to our study, 20% of each speaker's data was used for testing and the rest for training. The CNN delivered 97% accuracy, but no cross-validation was applied. Such CNN requires more than 279 million trainable parameters with large memory and computational footprint. In comparison, as shown in Table 2, our proposed EMAL SI achieved a minimum accuracy of 97.80% and a maximum of 98.61% on TIMIT cross-validation folds 2 and 6 experiments, respectively. This is an improvement of up to 1.61% that was achieved via significantly less complicated CNNs, as explained in the next section.

The CNN used in the second baseline system [25] adopted a different convolutional architecture than [23]. VoxCeleb provides significantly more training data than TIMIT and imposes more complexity, includes more speakers, and contains different types of background noises, which resulted in the lower accuracy reported for this dataset. The VoxCeleb 1 CNN was composed of five convolutional layers, four pooling layers, and two dense layers with 4096 and 1024 neurons, respectively, that received speech spectrograms

Table 3 DNN-based speaker identification systems comparative study

Reference	Neural Net	Dataset	Number of Speakers	Input	Highest accuracy (top-1)	Comments
[21]	Convolutional Deep Belief Network	TIMIT	168	Multi frames of MFCCs and spectrograms plus average, max, or standard deviation over all frames.	100%	Complex DNN structure although the detailed DNN architecture was not provided Large memory and computation footprint Multi-frame Multi-View Single-Learner
[2]	DBN + PCA + RBM + K-Means + SVM	The Urdu dataset	10	Multi frames of 36 MFCCs + spectrograms. Input dimension was 256 per frame.	92.60%	High complexity as the result of combining different algorithms. Small dataset Multi-frame Multi-View Single-Learner No cross-validation was applied
[13]	Golden Ratio-aided MLP	RAVDESS	24	MFCCs and Spectral information	99.3%	Small dataset Multi-frame information not provided Multi-View Single-Learner Complex features extraction with 1000 features
[18]	CNN	TIMIT	64	Mel Spectograms	99.56%	Small dataset Multi-frame information not provided Multi-View Single-Learner

Table 3 (continued)

Reference	Neural Net	Dataset	Number of Speakers	Input	Highest accuracy (top-1)	Comments
[40]	DBN + GMM	Custom	10	48 MFCCs + bottleneck features	98.10	Small dataset Multi-frame information not provided Multi-View Single-Learner Better performance than deep i-vector models was obtained
[45]	DAE + Bottleneck Feature + DBN	CENSREC-4	100	Multiple frames of 25 MFCCs	91.94%	Distance talking SI Digit utterances only High complexity as the result of combining different algorithms Multi-frame Multi-View Single-Learner No cross-validation was applied
[23]	CNN	TIMIT	630	One second spectrogram, 128×100 pixels	97.0%	Large memory and computation footprint
[25]	CNN	VoxCeleb 1	1251	Three second spectrogram, $512 \cdot 300$ pixels	80.5%	Multi-frame Multi-View Single-Learner No cross-validation was applied
[41]	DAE ANN	Census (AN4) speech	84	Multiple frames of 16 MFCCs Multiple frames of 16 MFCCs	98.8% 83.7%	Multi-frame Multi-View Single-Learner

Table 3 (continued)

Reference	Neural Net	Dataset	Number of Speakers	Input	Highest accuracy (top-1)	Comments
[30]	Radial Neural Network PNN	Arabic	20	Signal-image features extracted from min eigenvalues of Töeplitz matrix	97.18% 98.54%	Multi-View Single-Learner Small dataset Digit utterances only No cross-validation was applied
[3]	PNN and RBF and GRNN	GRID	34	Multiple frames of 20 MFCCs	97.5%	Ensemble learning Multi-frame Small dataset
[9]	CNN + GRU + MLP + CNN + MLP + DNN	SRE 2008 CHE SWBC	1336 120 254	Multiple frames of 20 MFCCs	98.5 92.8 86.2	Large memory and computation footprint Multi-frame No cross-validation was applied
The Proposed System	CNN	TIMIT VoxCeleb 1	630 1251	Single frame mono tensor of MFCCs 60 × 1 pixels	98.61% 82.93%	Enhanced Multi-Active Learning Small input dimension Small memory and computation footprint

of size 512·300 pixels as input. This neural network architecture translates into approximately 106 million trainable parameters. Similar to our experiment on the same corpus, the authors used around 8 K of the utterances for testing and the rest for training. Our proposed EMAL SI delivered 2.43% better accuracy over this benchmark.

5.2 Optimization and parameter explosion

Using a single sound facilitates the SI task as the dimension of the data DNN learners process decreases significantly; this means deep learning models with fewer parameters are required compared to traditional SI approaches that use a chain of acoustic sounds. Similarly, training SI systems using a sound segment is more convenient than the traditional approaches, where a minimum of three minutes of utterances from each speaker is recommended to achieve an acceptable level of accuracy, according to [22]. On the other hand, in our experience, high accuracy was achieved based on only around 25 s of training speech samples per speaker, which means around 87% fewer speech data was needed.

Concerning EMAL, it can be argued that using several CNNs may result in parameter explosion, but our experiments prove otherwise. In particular, each CNN comprising the proposed SI requires significantly fewer parameters as EMAL facilities the learning task by distributing it among several learners. In our experiments, an EMAL CNN dealt with only 22,913 trainable parameters meaning that the proposed SI needed overall 14 M trainable parameters to model all TIMIT speakers (22,913 parameters per CNN · 630 CNNs) in comparison to the benchmark CNN with more than 279 M trainable parameters, and 28 M parameters (22,913 · 1251) compared to VoxCeleb 1 benchmark CNN with over 106 M parameters. Consequently, the EMAL SI optimization resulted in around 95% reduction in the number of DNN trainable parameters over the TIMIT benchmark SI and 78% over the VoxCeleb 1 benchmark SI, and yet the proposed SI system improved the performance over both benchmark systems. This complexity optimization of SI resulting from integrating the single sound-based SI concept and EMAL can potentially enable devices within the lower-end processing power spectrum to perform accurate and offline speaker identification.

To put it differently, employing a deep learning model with hundreds of millions of parameters could be very challenging for any low-end processor, but the same processor can train and apply each CNN in the proposed SI sequentially, meaning that at any point in time, it deals with a considerably smaller CNN. Likewise, a lower amount of memory is required. In our experiments, we did not use any computer with GPUs; instead, we trained the proposed SI using three typical laptops by structuring them to train a range of EMAL CNNs in parallel.

5.3 Summary of contributions

The contributions of this study can be summarized as follows:

1. Speaker Identification using a single acoustic sound frame decreases the complexity of large SI and facilitates it since SI learners require to learn a fewer number of acoustic features in comparison to the traditional stacked-based approaches.
2. Single-frame SI requires shorter speech data.
3. EMAL framework decreases the SI structural complexity due to the reductions in trainable parameters.

4. The optimizations offered by the proposed system make SI more affordable since less expensive hardware may be required.
5. The proposed SI system offers state-of-the-art accuracies with considerably smaller CNNs, even over noisy speech data.

The proposed approach is open source and available from [31].

6 Conclusion

This paper proposed an optimized speaker identification system that integrates Enhanced Multi-Active Learners and the single sound segment approach. A text-independent speaker identification system that employed a network of CNNs to learn the speaker models and distribute the complexity of speaker identification was proposed and evaluated. The speaker models were formed according to the speakers' voice biometrics in a single sound segment presented by an acoustic frame of 60-dimensional MFCCs. Overall, we conducted experiments with 1881 CNNs, including 630 CNNs for each TIMIT speaker and 1252 CNNs for VoxCeleb 1 speakers, during which a standalone CNN was considered for each speaker model. Compared with similar CNN-based speaker identification systems trained and tested on the same speakers, the proposed system delivered comparable performance but significantly reduced the number of DNN trainable parameters. In particular, the proposed speaker identification system reduced the number of trainable parameters by up to 95% while delivering the top-1 accuracy of 98.61%.

Combining the reduction of the number of trainable parameters as the result of EMAL SI with the reduction of input dimension due to using a single sound segment, it can be concluded that the proposed SI system optimizes the complexity of challenging SI tasks. In other words, in an EMAL-based speaker identification system, each learner focuses solely on the acoustic features related to one speaker rather than encompassing all speaker features. Likewise, using a single sound approach, only one frame of acoustic features is fed to the learners in contrast to multiple stacked sequential frames, which means the input length is considerably smaller, as shown by formula 5. This may enable large-scale, offline speaker identification for devices without specific neural chips or GPUs, making SI more affordable. Finlay, we can highlight the novelty of the proposed approach as follows:

1. The applications of EMAL in SI.
2. Large-scale speaker identification using a single sound segment.
3. Implementation of a CNN-based SI system that benefits from (1) and (2).

The source code of the proposed approach is available from [31].

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Competing interests The author has no competing interests to declare that are relevant to the content of this article.

References

1. Ahmed AI, Chiverton JP, Ndzi DL, Al-Faris MM (2022) Channel and channel subband selection for speaker diarization. Comput Speech Lang 75. https://doi.org/10.1016/j.csl.2022.101367
2. Ali H, Tran SN, Benetos E, d'Avila Garcez AS (2018) Speaker recognition with hybrid features from a deep belief network. Neural Comput Appl 29(6):13–19. https://doi.org/10.1007/s00521-016-2501-7
3. Almaadeed N, Aggoun A, Amira A (2015) Speaker identification using multimodal neural networks and wavelet analysis. IET Biom 4(1):18–28. https://doi.org/10.1049/iet-bmt.2014.0011
4. An NN, Thanh NQ, Liu Y (2019) Deep CNNs with self-attention for speaker identification. IEEE Access 7:85327–85337. https://doi.org/10.1109/ACCESS.2019.2917470
5. Biagetti G, Crippa P, Falaschetti L, Orcioni S, Turchetti C (2016) Robust speaker identification in a meeting with short audio segments. Smart Innov Syst Technol 465–477. https://doi.org/10.1007/978-3-319-39627-9_41
6. Champiri ZD, Salim SSB, Shahamiri SR (2015) The role of context for recommendations in digital libraries. Int J Social Sci Humanity 5(11). https://doi.org/10.7763/ijssh.2015.v5.585
7. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. Accessed: May 29, 2019. [Online]. Available: http://arxiv.org/abs/1405.3531
8. Chen C, Wang W, He Y, Han J (2019) A bilevel framework for joint optimization of session compensation and classification for speaker identification. Digit Signal Process: Rev J 89. https://doi.org/10.1016/j.dsp.2019.03.008
9. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. Accessed: Jan. 28, 2020. [Online]. Available: http://arxiv.org/abs/1412.3555
10. Dash TK, Mishra S, Panda G, Satapathy SC (2021) Detection of COVID-19 from speech signal using bio-inspired based cepstral features. Pattern Recognit 117:107999. https://doi.org/10.1016/j.patcog.2021.107999
11. Dong Q, Gong S, Zhu X (2018) Imbalanced deep learning by minority class incremental rectification. IEEE Trans Pattern Anal Mach Intell 1–1. https://doi.org/10.1109/TPAMI.2018.2832629
12. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C: Appl Rev 42(4):463–484. https://doi.org/10.1109/TSMCC.2011.2161285
13. Garain A, Ray B, Giampaolo F, Velasquez JD, Singh PK, Sarkar R (2022) GRaNN: feature selection with golden ratio-aided neural network for emotion, gender and speaker identification from voice signals. Neural Comput Appl 34(17):14463–14486. https://doi.org/10.1007/S00521-022-07261-X/TABLES/20
14. Garofolo J et al (1993) TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Linguistic Data Consortium, vol 10, no 5, p 1, [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1; https://doi.org/10.35111/17gk-bn40
15. Hansen JHL, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process Mag 32(6). https://doi.org/10.1109/MSP.2015.2462851
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
17. Islam MA, Jassim WA, Cheok NS, Zilany MSA (2016) A robust speaker identification system using the responses from a model of the auditory periphery. PLoS ONE 11(7). https://doi.org/10.1371/journal.pone.0158520
18. Karthikeyan V, S SP (2022) Modified layer deep convolution neural network for text-independent speaker recognition. J Exp Theor Artif Intell. https://doi.org/10.1080/0952813X.2022.2092560

19. Kiran Reddy M et al (Sep. 2021) The automatic detection of heart failure using speech signals. Comput Speech Lang 69:101205. https://doi.org/10.1016/j.csl.2021.101205
20. LeCun YA, Bengio Y, Hinton GE (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
21. Lee H, Largman Y, Pham P, Ng A (2009) Unsupervised feature learning for audio classification using convolutional deep belief networks. Adv Neural Inf Process Syst. https://doi.org/10.1145/1553374.1553453
22. Lu H, Bernheim Brush AJ, Priyantha B, Karlson AK, Liu J (2011) SpeakerSense: Energy efficient unobtrusive speaker identification on mobile phones. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp 188–205. https://doi.org/10.1007/978-3-642-21726-5_12
23. Lukic Y, Vogt C, Dürr O, Stadelmann T, Durr O, Stadelmann T (2016) Speaker Identification and Clustering Using Convolutional Neural Networks. In: Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on, IEEE, pp 13–16. https://doi.org/10.1109/MLSP.2016.7738816
24. Lyons J. Python Speech Feature extraction. MIT. [Online]. Available: https://pypi.org/project/python_speech_features/0.4/
25. Nagrani A, Chung JS, Zisserman A (2017) VoxCeleb: a large-scale speaker identification dataset. In: Interspeech 2017, ISCA: ISCA, pp 2616–2620. https://doi.org/10.21437/Interspeech.2017-950
26. Porwik P, Doroz R, Wrobel K (2018) An ensemble learning approach to lip-based biometric verification, with a dynamic selection of classifiers. Expert Syst Appl Aug. https://doi.org/10.1016/J.ESWA.2018.08.037
27. Reynolds D, Rose RC (1995) Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans Speech Audio Process 3(1):72–83. https://doi.org/10.1109/89.365379
28. Richardson F, Member S, Reynolds D, Dehak N (2015) Deep neural network approaches to Speaker and Language Recognition. In: IEEE Signal Processing Letters. IEEE, Queensland, pp 1671–1675. https://doi.org/10.1109/LSP.2015.2420092
29. Roger V, Farinas J, Pinquier J (2022) Deep neural networks for automatic speech processing: a survey from large corpora to limited data. EURASIP J Audio Speech Music Process 2022(1). https://doi.org/10.1186/s13636-022-00251-w
30. Saeed K, Nammous MK (2007) A speech-and-speaker identification system: feature extraction, description, and classification of speech-signal image. IEEE Trans Ind Electron 54(2):887–897. https://doi.org/10.1109/TIE.2007.891647
31. Shahamiri SR (2023) Enhanced-Multi-Learner Single-Sound-Segment Speaker Identification. GitHub, Accessed: May 19, 2023. [Online]. Available: https://github.com/rshahamiri/Enhanced-Multi-LearnerSingle-Sound-Segment-Speaker-Identification
32. Shahamiri SR (2021) Neural network-based multi-view enhanced multi-learner active learning: theory and experiments. J Exp Theor Artif Intell: 1–21. https://doi.org/10.1080/0952813X.2021.1948921
33. Shahamiri SR, Kadir WMNW, Ibrahim S (2010) A single-network ANN-based oracle to verify logical software modules. In: ICSTE 2010–2010 2nd International Conference on Software Technology and Engineering, Proceedings, 2010. https://doi.org/10.1109/ICSTE.2010.5608808
34. Shahamiri SR, Tahbtah F (2020) An investigation towards speaker identification using a single-sound-frame. Multimed Tools Appl 79: 31265–31281. https://doi.org/10.1007/s11042-020-09580-4
35. Shahamiri SR, Thabtah F, Abdelhamid N (2022) A new classification system for autism based on machine learning of artificial intelligence. Technol Health Care 30(3). https://doi.org/10.3233/THC-213032
36. Shahin I, Nassif AB, Hamsa S (2018) Novel cascaded gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments. Neural Comput Appl 1–13. https://doi.org/10.1007/s00521-018-3760-2
37. Shi Y, Huang Q, Hain T (2020) H-vectors: utterance-level speaker embedding using a hierarchical attention model. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., pp 7579–7583. https://doi.org/10.1109/ICASSP40776.2020.9054448
38. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Accessed: May 27, 2019. [Online]. Available: http://arxiv.org/abs/1409.1556
39. Sun S, Zhang Q (2011) Multiple-view multiple-learner semi-supervised learning. Neural Process Lett 34(3):229–240. https://doi.org/10.1007/s11063-011-9195-8

40. Sun L, Gu T, Xie K, Chen J (2019) Text-independent speaker identification based on deep Gaussian correlation supervector. Int J Speech Technol 22(2):449–457. https://doi.org/10.1007/s10772-019-09618-5
41. Tirumala SS, Shahamiri SR (2017) A deep autoencoder approach for speaker identification. In: 9th International Conference on Signal Processing Systems (ICSPS), ACM, Auckland, pp 175–179. https://doi.org/10.1145/3163080.3163097
42. Xu L, Lee KA, Li H, Yang Z (2018) Generalizing I-vector estimation for rapid speaker recognition. IEEE/ACM Trans Audio Speech Lang Process 26(4):749–759. https://doi.org/10.1109/TASLP.2018.2793670
43. Yadav S, Rai A (2018) Learning discriminative features for Speaker Identification and Verification. In: Interspeech 2018. ISCA, ISCA, pp 2237–2241. https://doi.org/10.21437/Interspeech.2018-1015.
44. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, Association for Computational Linguistics (ACL), 2016, pp 1480–1489. https://doi.org/10.18653/v1/n16-1174
45. Zhang Z, Wang L, Kai A, Yamada T, Li W, Iwahashi M (2015) Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification. EURASIP J Audio Speech Music Process 2015(1). https://doi.org/10.1186/s13636-015-0056-7
46. Zhao X, Wang Y, Wang D (2014) Robust speaker identification in noisy and reverberant conditions. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp 3997–4001. https://doi.org/10.1109/ICASSP.2014.6854352