Neural Computing and Applications (2022) 34:21319–21333
https://doi.org/10.1007/s00521-022-07613-7

ORIGINAL ARTICLE

Development of novel automated language classification model using pyramid pattern technique with speech signals

Erhan Akbal[1] · Prabal Datta Barua[2,3] · Turker Tuncer[1] · Sengul Dogan[1] · U. Rajendra Acharya[4,5,6]

Received: 5 November 2021 / Accepted: 4 July 2022 / Published online: 25 July 2022

Abstract

Language classification using speeches is a complex issue in machine learning and pattern recognition. Various text and image-based language classification methods have been presented. But there are limited speech-based language classification methods in the literature. Also, the previously presented models classified limited numbers of languages, and few are accents. This work presents an automated handcrafted language classification model. The novel pyramid pattern is presented to extract the features extraction. Also, statistical features and maximum pooling are used to generate the features. We have developed our speech-language classification model using *two* datasets: (i) created a new big speech dataset containing 14,500 speeches in 29 languages, and (ii) used the VoxForge dataset. The neighborhood component analysis method is used to select the most informative 1000 features from the generated features, and these features are classified using a quadratic support vector machine classifier (QSVM). Our developed method yielded $98.87 \pm 0.30\%$ and $97.12 \pm 1.27\%$ accuracies for our and VoxForge datasets, respectively. Also, geometric mean, average precision, and F1-score evaluation parameters are calculated, and they are presented in the results section. This paper presents an accurate language classification model developed using *two* big speech-language datasets. Our results indicate the success of the proposed pyramid pattern-based language classification method in classifying various speech languages accurately.

Keywords Pyramid pattern · Speech-based language classification · Speech language classification dataset · Machine learning

✉ U. Rajendra Acharya
aru@np.edu.sg

Erhan Akbal
erhanakbal@firat.edu.tr

Prabal Datta Barua
prabal.barua@usq.edu.au

Turker Tuncer
turkertuncer@firat.edu.tr

Sengul Dogan
sdogan@firat.edu.tr

[1] Department of Digital Forensics Engineering, College of Technology, Firat University, Elazig, Turkey

[2] School of Business (Information System), University of Southern Queensland, Toowoomba, QLD 4350, Australia

[3] Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia

[4] Department of Electronics and Computer Engineering, Ngee Ann Polytechnic, Singapore 599489, Singapore

[5] Department of Biomedical Engineering, School of Science and Technology, SUSS University, Singapore, Singapore

[6] Department of Biomedical Informatics and Medical Engineering, Asia University, Taichung, Taiwan

1 Introduction

Speech is the basic component of communication. There are variable societies in the world that have developed their own language for communication. Communication cannot be established if someone is speaking an unknown language. In such cases, a translation process should be performed to provide communication by using a translator/interpreter [1]. Humans are the best system for language identification [2]. However, there are many languages in the world. Hence it is very difficult to determine which language the speaker speaking. In addition, each language consists of different accents and dialects.

Language identification and classification identify the target language with high accuracy using the acoustic properties of speech signals [3]. The language identification process is different from the speech or speaker recognition process. The purpose of language identification using speech is to use the characteristics of sounds using the text-based features of the language [4]. Speech has acoustic, phonetic, and prosodic features of the language. In addition, the alphabet, words, morphology, syntax, and grammatical structure are factors affecting speech [5]. Therefore, languages show different acoustic properties regardless of speaker and computer-aided automatic language identification systems (ALIS) [6].

ALIS performs automatic language identification based on the features extracted from the speech signal of each language and used for different purposes [7]. In determining the languages of refugees caught by law enforcement, it is essential to determine the language of the person with communication problems that may occur during border crossings. Language identification with manual methods is challenging and taxing [8]. Identifying the target language can take days. In forensics, there is a need to identify the language of the speech content using the audio files. Many different audio files need to be examined for digital evidence. Analysis of evidence for the content in the unknown language depends on language identification [9]. In addition, language detection is needed in many applications, such as speech recognition systems and speech-to-speech translation [10–12]. According to the results obtained from the automatic language identification system, speech translation systems ensure that the speech is translated into the target language. This feature is used as a preliminary step in telephone call systems and automatic translation systems.

Acoustic properties are extracted from raw speech signals using feature extraction methods [13]. The most commonly used feature extraction methods are linear predictive coefficients (LPC) [14], linear predictive cepstral coefficients LPCC [15], Mel frequency cepstral coefficients (MFCC) [16, 17], perceptual linear prediction features (PLP) [18], Gaussian mixture models (GMM) [19], and hidden Markov model (HMM) [20]. The classifiers used for the automated classification of speech are neural network [21], k-nearest neighbors (KNN) [22], linear discriminant analysis (LDA) [23], deep neural network (DNN) [24], and recurrent neural network (RNN) [25]. The machine learning models are more popular due to their low computational cost and high-performance results. Therefore, studies on smart systems and machine learning have been presented in many different fields in the literature [26–28].

1.1 Motivation and our method

This research focuses on *two* primary problems in language classification. These are the construction of a huge database of speech signals and presenting highly accurate automated language identification/classification methods. Language classification is one of the hot-topic in the research. Many image processing and text classification methods have been presented to achieve high classification accuracy using images and texts. The presented many speeches based models aimed to detect the limited number of languages or accents of a language. This research focuses on the need for the testbed for speech-based language classification. Therefore, an extensive database is collected from Youtube, and the second aim of this work is to obtain high classification accuracy using more classes. The literature states that a highly accurate learning model should have an efficient feature generation/method. Therefore, a multilevel generation method is presented. This model uses pooling decomposition (maximum pooling) [29] for generating levels to extract high-level features. Also, statistical and textural features are generated by applying statistical moments and the proposed pyramid pattern. As seen from the literature review (see previous section/Sect. 1), conventional/shallow sound descriptors like LPCC, MFCC, PLP, GMM, and HMM have been used. These feature extractors have limited speech classification ability. Thus, a sensitive and robust feature extraction model should be presented. To realize this purpose, a new 3D shape-based feature extraction function has been proposed and this extractor is named pyramid pattern. The introduced pyramid pattern can detect speech differences and is a 3D graph-based descriptor. Our feature extraction motivation is to investigate the feature vector generation ability of a graph-based local feature extractor and evaluate the classification ability of this feature extractor for the language identification problem. Using the presented pyramid pattern, statistical pattern, and maximum pooling, both textural (pyramid pattern extracts textural features), and statistical (using statistical moments) features are extracted.

Furthermore, deep learning models have high classification performance since they used more levels and can learn features at the higher abstract level than hand-modeled methods but they are expensive models since their time complexity is generally exponential. By mimicking deep learning architecture, a highly accurate model with linear time complexity (the time complexity of the proposed feature extraction model is $O(nlogn)$) have been presented. The recommended pyramid pattern based language classification model inspired by the deep learning network such as AlexNet [30], and GoogLeNet [31]. These networks can be used to generate a large number of features. Then, neighborhood component analysis (NCA) [32] is applied to developed features to select clinically significant features and finally fed to the quadratic SVM classifier for automated classification.

1.2 Literature review

Many studies on language identification are concentrated on four areas: language identification, language classification, language diarization, and voice activity detection [33]. Language identification is aimed to determine whether there is a single language in a speech signal [34–36]. Language classification aims to determine whether the language in the sound signal belongs to a certain language class [2, 36–38] and can be classified with a high accuracy rate. The language diarization is aimed to diarization the languages in speech files containing multiple languages [39–41]. Voice activity detection aims to identify human speech sounds among different environmental sounds. Target speech can be identified by separating speech and environmental noise [42–45]. These systems can be used as a preliminary module for language identification, classification, and diarization studies using techniques such as Mel frequency cepstral coefficients (MFCC) [46], autocorrelation function analysis [47], Gaussian mixture model [42], multilayer perceptron [42] and linear predictive coding [48]. The most concentrated areas among these four areas are language identification and voice activity detection. Because identifying a single language or speech from an audio file is an easier task than language diarization and classification [33]. It is more difficult to obtain high accuracy in classification and diarization studies performed with multiple languages [49]. Our proposed study is more difficult as it is automated language classification. There is the limited number of studies in the literature on language classification. Reference [3] proposed a model for classifying four languages: Tamil, Malayalam, Hindi, and English. Feedforward back-propagation neural network is used. By using perceptual linear prediction, Mel frequency cepstral coefficients (MFCC), and relative spectral transform perceptual linear prediction hybrid feature extractors, a maximum accuracy rate of 94.6% is achieved. Reference [50] proposed a method to classify the northeast Indian languages. Experiments conducted using two databases consisting of 11 and 7 languages obtained a maximum accuracy rate of 80%.Reference [51] presented a model using a GMM classifier with MFCC and PLP hybrid feature extractors. It has been shown that a maximum accuracy of 88.75% is obtained using a database with three different Indian languages. Reference [52] presented a method using a graph-based feature extractor. Tests are conducted on three other sound-based speech databases. They have achieved the classification performance between 87.7 and 91.23%. Authors in [53] proposed a method for language identification using speech sounds with HMM, SVM, and neural networks. The tests performed on four languages obtained the highest classification accuracy of 70%. The best performance is obtained with the HMM as it uses temporal properties. Reference [54] presented a method for detecting three different indigenous Indonesian languages from their speech sounds. They obtained the accuracy of 77.42 and 75.94% using phonotactics methods. Reference [55] proposed a forensic speaker recognition system using MFCC feature extractor and deep learning methods. They obtained the detection accuracy of 95.1% in identifying the Urdu language.

1.3 Novelties

The novelties of our proposed method are given below:

- A new big database is used as a testbed for our proposed speech-based language classification model.
- A novel pyramid pattern textural feature generation function is proposed in this work.

1.4 Key contributions

Our key contributions are;

- The used databases for language classification corpora include a limited number of languages. To assess the performance of our learning model with a wide spectrum, a big speech database is collected from the Youtube of various languages of male and female speakers.
- An efficient and highly accurate language classification method is presented. This method uses handcrafted features, and hence there is no need to set millions of parameters like deep learning methods.

2 Dataset collection

Two different speech datasets were used in our study. The first data set ($\boldsymbol{DS_a}$) is collected from youtube, which consists of 29 different languages. The second dataset ($\boldsymbol{DS_b}$) is VoxForge [56], which consists of 16 different languages. The details of these two datasets' are given below.

DS_a: In our study, scenario speakers and texts were not used to provide the usability of the proposed method in real-time systems. A new dataset is created with only audios (speeches) by language learning and listening videos on YouTube [57]. These videos feature a formal speaking accent of all languages. There are various speakers of different genders in the used records. In addition, the recordings have different environmental noise and sound recorder features. Thus, it is aimed to obtain performance measurement of algorithm independent of environment, speaker, and text. Herein, a hand-modeled learning model has been tested on the acquired dataset. Deep learning models like convolutional neural network (CNN)s inspire this model. In this model, both textural (using a proposed 3D graph-based local feature extractor), and statistical features have been extracted in each level, and levels are created using maximum pooling like CNN. Using this strategy, both low and high-level features have been extracted, and the feature selector has selected the most appropriate features. In this aspect, we proposed a multileveled hand-modeled speech classification model and is a feature engineering model. To create and appropriate speech dataset, these steps were conducted. First, non-speaking voices in the recordings were cleaned manually. Windows operating system and NHC WavePad [58] program were used for manual cleaning. Then, the recordings were divided into pieces of 8–15 s duration. Ten different sound recordings were used for each language. Thus, 500 sample speech files were created for each language. The file formats of the sample speech files are m4a and wav, and the sampling frequency is 48 kHz. The database consists of 14,500 speech files with 29 languages and a total of 2856.75 min. The details of the speech signal database used database is demonstrated in Table 1. Furthermore, our collected dataset was publicly published, and this dataset can be downloaded from http://web.firat.edu.tr/turkertuncer/lang_data.rar URL.

DS_b: This dataset consists of the speech of the speakers on Voxforge and has been used in different studies [52, 59, 60]. Voxforge has speaker speech from 16 different languages. These languages are Albanian, Croatian, English, Spanish, French, German, Greek, Hebrew, Italian, Catalan, Netherlands, Persian, Portuguese, Russian, Turkish, and Ukrainian. The $\boldsymbol{DS_b}$ dataset consists of 16 languages, 300 samples for each language, and 4800 samples. Samples have wav file extension and 16 kHz frequency.

3 The proposed pyramid pattern

In this work/research, a novel 3D graph-based textural feature extractor has been proposed, and this pattern is a histogram-based textural feature extractor. Details of the proposed pyramid pattern is given in below.

1. Divide one-dimensional into 25-sized blocks.

$$\text{window}^i(j) = S(i+j-1), j = \{1,2,\ldots,25\}, i = \{1,2,\ldots,\text{Len}-24\} \quad (1)$$

   where window^i describes ith window/block with a length of 25.

2. Employ vector to matrix transformation to divided windows.

$$P^i(k,l) = S(j),\ k = \{1,2,\ldots,5\},\ l = \{1,2,\ldots,5\} \quad (2)$$

   where P^i is the ith matrix with a size of 5 × 5.

3. Deploy signum function to determine values by assigning the presented pyramid pattern (see Fig. 1) and extracting binary features.

   The red circles are utilized as nodes and the relationship of the edge of the value. The center value (P(3,3)) is considered as the top point of the pyramid. The mid 3 × 3 sized matrix of the used 5 × 5 matrix is used as the first floor of the pyramid. The rest values consist of the floor of the pyramid.

   Mathematical definitions of the bit extraction process of the proposed pyramid pattern are given in Eqs. (3)–(6).

$$\begin{vmatrix} b1(1) \\ b1(2) \\ b1(3) \\ b1(4) \\ b1(5) \\ b1(6) \\ b1(7) \\ b1(8) \end{vmatrix} = \text{signum}\left(\begin{vmatrix} P^i(1,1) & P^i(3,3) \\ P^i(1,3) & P^i(3,3) \\ P^i(1,5) & P^i(3,3) \\ P^i(3,1) & P^i(3,3) \\ P^i(3,5) & P^i(3,3) \\ P^i(5,1) & P^i(3,3) \\ P^i(5,3) & P^i(3,3) \\ P^i(5,5) & P^i(3,3) \end{vmatrix}\right) \quad (3)$$

$$\begin{vmatrix} b2(1) \\ b2(2) \\ b2(3) \\ b2(4) \\ b2(5) \\ b2(6) \\ b2(7) \\ b2(8) \end{vmatrix} = \text{signum}\left(\begin{vmatrix} P^i(2,2) & P^i(3,3) \\ P^i(2,3) & P^i(3,3) \\ P^i(2,4) & P^i(3,3) \\ P^i(3,2) & P^i(3,3) \\ P^i(3,4) & P^i(3,3) \\ P^i(4,2) & P^i(3,3) \\ P^i(4,3) & P^i(3,3) \\ P^i(4,4) & P^i(3,3) \end{vmatrix}\right) \quad (4)$$

Table 1 Summary of the collected DS_a speech sound dataset

No	Name of language	Number of samples	Total duration (minutes)
1	Arabic	500	96.07
2	Bulgarian	500	88.85
3	Cantonese	500	91.23
4	China	500	92.67
5	Danish	500	98.59
6	Dutch	500	99.66
7	English	500	97.45
8	Filipino	500	119.28
9	Finnish	500	107.96
10	French	500	104.39
11	German	500	63.57
12	Greek	500	104.44
13	Hebrew	500	105.83
14	Hindi	500	109.06
15	Hungarian	500	102.89
16	Indonesian	500	104.67
17	Italian	500	97.95
18	Japan	500	104.96
19	Korean	500	85.40
20	Polish	500	100.87
21	Portuguese	500	89.17
22	Romanian	500	104.28
23	Russian	500	87.85
24	Spanish	500	101.33
25	Swahili	500	101.58
26	Swedish	500	100.55
27	Thai	500	109.12
28	Turkish	500	89.56
29	Urdu	500	97.52
Total		14,500	2856.75

$$\begin{vmatrix} b3(1) \\ b3(2) \\ b3(3) \\ b3(4) \\ b3(5) \\ b3(6) \\ b3(7) \\ b3(8) \end{vmatrix} = \text{signum}\left(\begin{vmatrix} P^i(2,2) & P^i(1,1) \\ P^i(2,3) & P^i(1,3) \\ P^i(2,4) & P^i(1,5) \\ P^i(3,2) & P^i(3,1) \\ P^i(3,4) & P^i(3,5) \\ P^i(4,2) & P^i(5,1) \\ P^i(4,3) & P^i(5,3) \\ P^i(4,4) & P^i(5,5) \end{vmatrix}\right) \quad (5)$$

$$\text{signum}(a,b) = \begin{cases} 0, & a - b < 0 \\ 1, & a - b \geq 0 \end{cases} \quad (6)$$

As stated in Eqs. (10)–(13), three bits groups are generated by deploying signum function, and they are named $b1, b2$, and $b3$. Each group has eight bits, and the map signals are calculated by using these bits.

4. Use binary to decimal conversion for calculating map signal.

$$\text{map1}(i) = \sum_{j=1}^{8} b1(j) * 2^{j-1} \quad (7)$$

$$\text{map2}(i) = \sum_{j=1}^{8} b2(j) * 2^{j-1} \quad (8)$$

$$\text{map3}(i) = \sum_{j=1}^{8} b3(j) * 2^{j-1} \quad (9)$$

where map1, map2, and map3 are generated by three map signals.

5. Calculate the histograms of the map signals. As can be seen from Eqs. (7)–(9), these map signals are coded with eight bits. Therefore, the generated histograms have 256 elements/values.

In the first step of the histogram generation, the initial values of the histogram are assigned as zero.

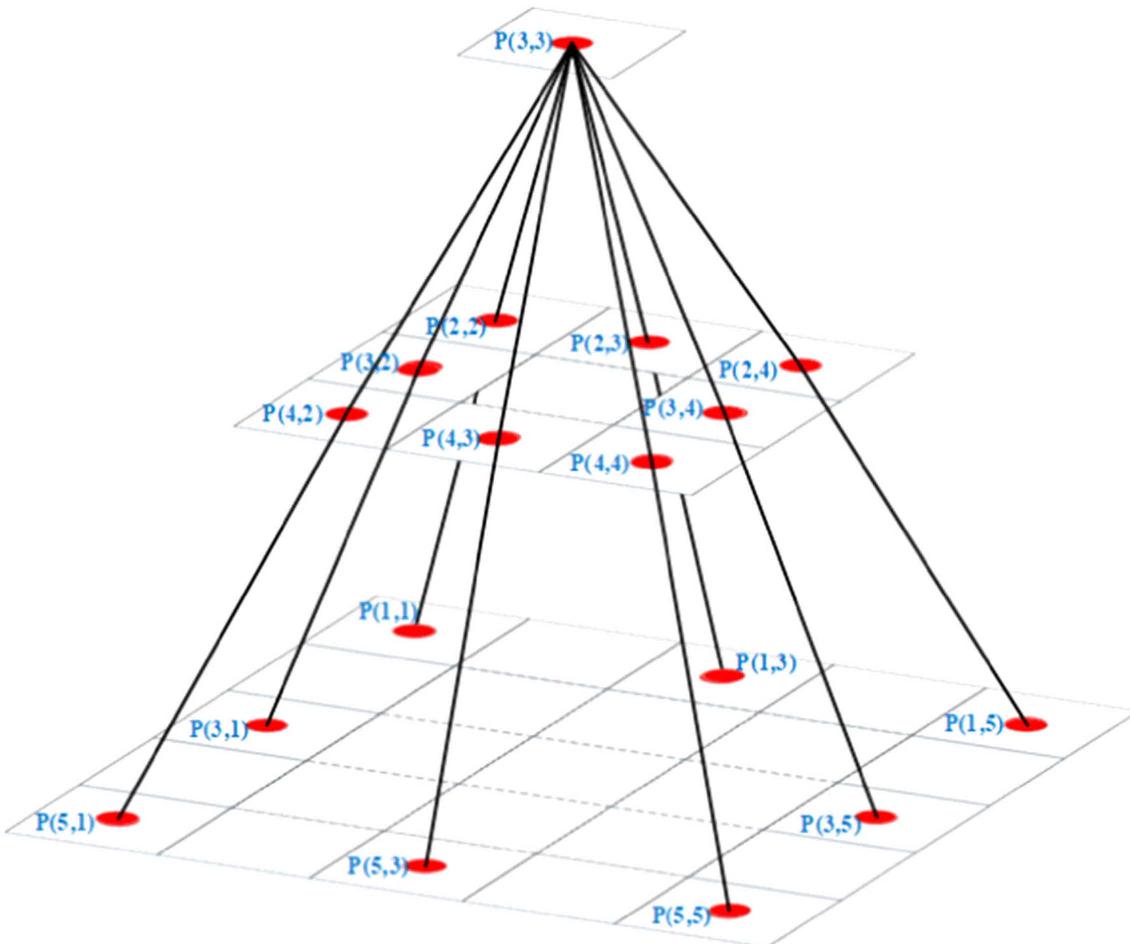

Fig. 1 Graphical demonstration of recommended pyramid pattern. The red circles are used as nodes, and the edges relationship of the value. The center value ($P(3,3)$) is considered as the top point of the pyramid. The mid 3 × 3 sized matrix used 5 × 5 matrix as the first floor of the pyramid. The rest values are the floor of the pyramid

$$\mathrm{hist}^1(t) = 0,\ t = \{1, 2, \ldots, 2^8\} \tag{10}$$

$$\mathrm{hist}^2(t) = 0 \tag{11}$$

$$\mathrm{hist}^2(t) = 0 \tag{12}$$

where $\mathrm{hist}^1, \mathrm{hist}^2$ and $hist^3$ are the histograms of the map1, map2 and map3 signals consecutively. The histogram extraction process is demonstrated in Eqs. (13)–(15) mathematically.

$$\mathrm{hist}^1(\mathrm{map1}(i) + 1) = \mathrm{hist}^1(\mathrm{map1}(i) + 1) + 1 \tag{13}$$

$$\mathrm{hist}^2(\mathrm{map3}(i) + 1) = \mathrm{hist}^2(\mathrm{map2}(i) + 1) + 1 \tag{14}$$

$$\mathrm{hist}^3(\mathrm{map3}(i) + 1) = \mathrm{hist}^3(\mathrm{map3}(i) + 1) + 1 \tag{15}$$

6. Merge the calculated/extracted histograms and obtain the final feature (*feat*) vector with a size of 768.

$$\mathrm{feat}((k-1) * 256 + j) = \mathrm{hist}^k(j),\ j = \{1, 2, \ldots, 256\},\ k = \{1, 2, 3\} \tag{16}$$

The steps 1–6 comprises of the $PP(.)$ feature generation function. $PP(.)$ function has been used to define the proposed multileveled feature generation model.

4 The presented pyramid pattern-based language classification model

The primary objective of the pyramid pattern-based identification model is to yield high accuracy database using this big database involving 29 languages. In this work, traditional statistical and textural feature generators are used. The statistical feature generation methods used linear and nonlinear statistical moments, and 18 statistical features are generated. A new pyramid pattern (it is a microstructure) is used to extract textural features. The recommended pyramid pattern uses 5×5 size matrix to create a pattern like a pyramid and generates 768 features. Also, the used statistical generator is deployed to generate 18 features. Therefore, $768 + 18 + 18 = 804$ features are generated by employing the feature generation functions. However, these functions are generated by low-level features from the speech. A decomposition model must be used like deep learning methods to generate high-level features. Thus, maximum pooling is utilized as a decomposition method, and *ten* leveled generation network is created. And this network generates $804 * 10 = 8040$ features. The NCA is applied to the 8040 features to obtain the 1000 most informative features. For automated classification, these features are fed to Quadratic SVM [62]. The graphical layout of the suggested pyramid pattern-based language classification model is demonstrated/depicted in Fig. 2.

The presented model has ten leveled feature generation network. The levels are created by maximum pooling decomposition. The used main feature generation functions are the suggested pyramid pattern and statistical feature generator. NCA chooses the generated features from the created ten levels. These features are classified using QSVM.

The pseudo-code of the presented pyramid pattern-based method is given in Algorithm 1.

The method comprises *three* main phases: feature generation, feature selection, and classification. These phases are demonstrated in Algorithm 1. Lines 01–08 denote feature generation, line 09 shows NCA-based feature selection, and the classification phase is shown in line 10.

4.1 Feature generation method

It is the first phase of the presented model, which uses *ten* levels (features are extracted from ten signals, and these are a raw speech signal and nine decomposed signals). This phase extracts two feature generation functions: statistical and textural features. Steps of the suggested generation method are given below;

Step 1: Generate *nine* decomposed signals by applying maximum pooling decomposition.

We have used the maximum pooling method to decompose signals using two-sized non-overlapping blocks.

$$D^1(j) = \max(S(i), S(i+1)),\ j = \left\{1, 2, \ldots, \frac{\text{Len}}{2}\right\},\quad i = \{1, 3, \ldots \text{Len} - 1\} \tag{17}$$

$$D^k(j) = \max\left(D^{k-1}(i),\ D^{k-1}(i+1)\right),\ k = \{2, 3, \ldots, 9\} \tag{18}$$

$$\max(a, b) = \begin{cases} a, & a \geq b \\ b, & a < b \end{cases} \tag{19}$$

Algorithm 1. Pyramid pattern and maximum pooling-based language classification model.

Input: The collected speech dataset

Output: Validation predictions (vp).

00: Load the collected speech dataset

01: **for** k=1:14500 // The used dataset has 14500 speech signals

02: Read each speech (S).

03: **for** i=1:10

04: $X(k, (i-1) * 804 + 1{:}804 * i) = conc\left(PP(S), St(S), St\left(PP(S)\right)\right)$; // Generate features using the presented pyramid pattern ($PP(.)$), statistical feature generator ($St(.)$). This line also demonstrates the feature concatenation process.

05: $D = maxp(S)$; // Decompose the speech signal using maximum pooling ($maxp(.)$)

06: $S = D$; // Update speech signal

07: **end for i**

08: **end for k**

09: Select 1000 the most informative features by deploying NCA.

10: Classify the selected 1000 features and obtain vp.

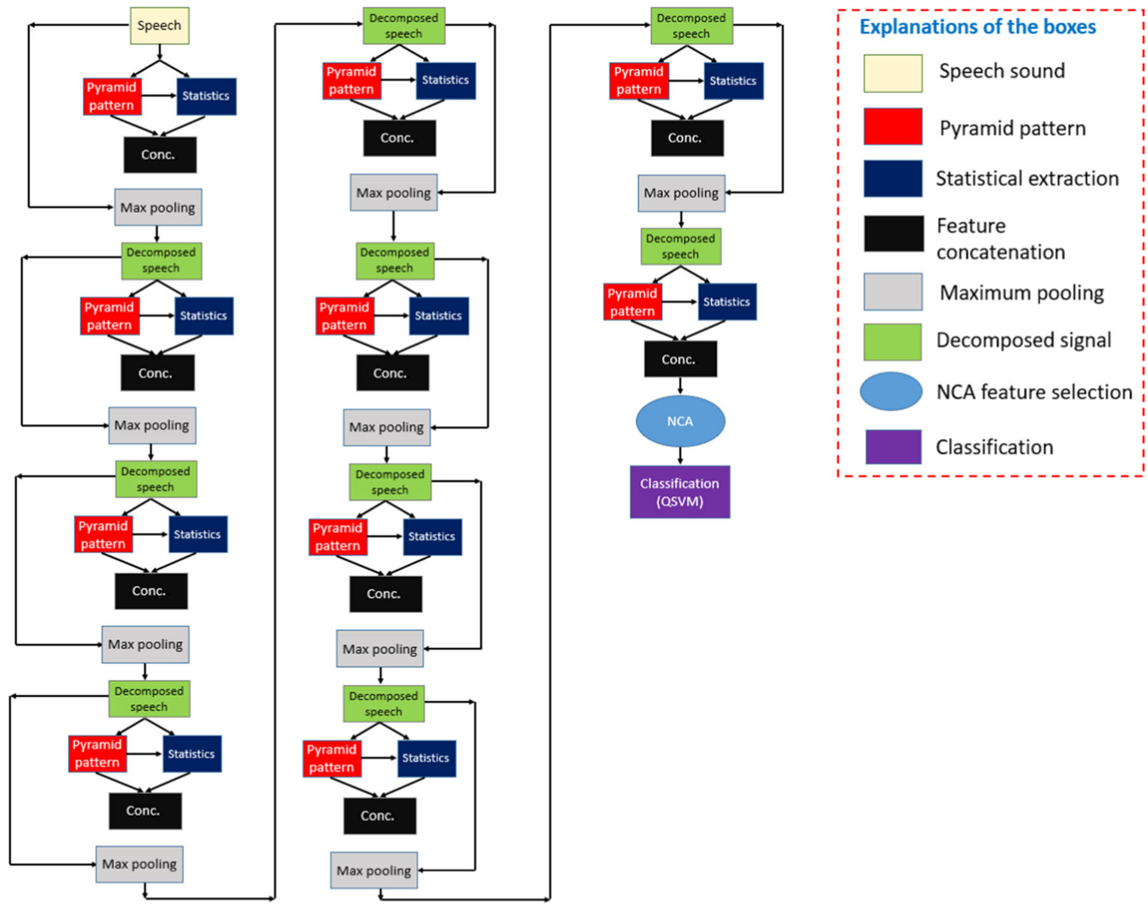

Fig. 2 Block diagram of pyramid pattern-based speech identification model. The presented model has ten leveled feature generation network

where D^k is kth level decomposed signal, S represents input signal, $\max(.,.)$ expresses maximum value calculation function, Len denotes the length of the signal, a, and b define input parameters of the $\max(.,.)$ function.

The Eqs. (1)–(3) defines the used maximum pooling method used (it is shown in Line 05 of Algorithm 1) mathematically.

Step 2: Generate statistical features ($feat^{St}$) of the speech signal and decomposed signal by using a statistical feature generator (St(.)).

$$\text{feat}^{\text{St}}(j) = \text{St}(S), \quad j = \{1, 2, \ldots, 18\} \tag{20}$$

$$\text{feat}^{\text{St}}(j + k * 18) = \text{St}\left(D^k\right), \quad k = \{1, 2, \ldots, 9\} \tag{21}$$

The moments which are consisted of the St(.) are listed in Table 2 [61].

Step 3: Generate 768 textural features deploying the presented pyramid pattern. In this step, the pyramid pattern has been detailed and explained.

$$\text{feat}^{t}(j) = \text{PP}(S), \quad j = \{1, 2, \ldots, 768\} \tag{22}$$

$$\text{feat}^{t}(j + k * 768) = \text{PP}\left(D^k\right), \quad k = \{1, 2, \ldots, 9\} \tag{23}$$

where feat^{t} describes textural features, and PP(.) is a pyramid pattern feature generation function (see Sect. 3 for details).

Step 4: Extract the statistical features of the generated textural features. These features are named statistical features of the textural features (feat^{ts}).

$$\text{feat}^{\text{ts}}(j) = \text{St}(\text{PP}(S)), \quad j = \{1, 2, \ldots, 18\} \tag{24}$$

$$\text{feat}^{\text{ts}}(j + k * 18) = \text{St}\left(\text{PP}\left(D^k\right)\right), \quad k = \{1, 2, \ldots, 9\} \tag{25}$$

Step 5: Merge the generated features to calculate the concatenated features (X) with a size of 8040.

$$X = \text{conc}\left(\text{feat}^{\text{St}}, \text{feat}^{t}, \text{feat}^{\text{ts}}\right) \tag{26}$$

where $conc(.)$ is the concatenation function.

4.2 Feature selection with NCA

Feature selection is one of the critical steps in machine learning. It helps to improve the performance and reduce the execution time of the classifier. In this work, NCA [32] is used to perform the feature selection. It measures

Table 2 Mathematical definitions of the used statistical moments used in this work

Num	Equation	Num	Equation
1	$\frac{1}{Ln}\sum_{j=1}^{\text{Len}} S(j)$	10	$\max(S) - \text{median}(S)$
2	$\sqrt{\frac{\sum_{i=1}^{\text{Len}}(S(i) - \frac{1}{\text{Len}}\sum_{j=1}^{\text{Len}} S(j))}{\text{Len}-1}}$	11	$\frac{1}{\text{Len}}\sum_{j=1}^{\text{Len}} \lvert S(j)\rvert$
3	$\max(S)$	12	$-\sum_{j=1}^{\text{Len}} \log(\text{prob}(S(j)))^2$
4	$\min(S)$	13	$\max(\lvert S\rvert) - \min(\lvert S\rvert)$
5	$\text{median}(S)$	14	$\min(\lvert S\rvert)$
6	$\frac{1}{\text{Len}}\left(\sum_{i=1}^{\text{Len}}(S(i) - \frac{1}{\text{Len}}\sum_{j=1}^{\text{Len}} S(j))\right)^2$	15	$\sqrt{\frac{\sum_{i=1}^{\text{Len}}(\lvert S(i)\rvert - \frac{1}{\text{Len}}\sum_{j=1}^{\text{Len}} \lvert S(j)\rvert)}{\text{Len}-1}}$
7	$\frac{1}{\text{Len}}\sum_{j=1}^{\text{Len}} S(j)^2$	16	$-\sum_{j=1}^{\text{Len}} \text{prob}(S(j)) * \log(\text{prob}(S(j)))$
8	$\frac{1}{\text{Len}}\sum_{i=1}^{\text{Len}} \lvert S(i) - \frac{1}{Ln}\sum_{j=1}^{\text{Len}} S(j)\rvert$	17	$\sum_{j=1}^{\text{Len}} S(j)^2$
9	$\max(S) - \min(S)$	18	$-\sum_{j=1}^{\text{Len}} \text{prob}(S(j))^2 * \log(\text{prob}(S(j)))^2$

where prob(.) defines probability

distances to calculate the weights of the features. It computes the weights step by step and uses regularization parameters. It generates non-negative weights using stochastic gradient descent or ADAM optimizers [62, 63]. Therefore, it is a back-propagation method. The generated weights are sorted in descending order, and the most valuable features are selected by using the indices of the sorted features of the weights.

The NCA selector has selected 1000 features from generated 8040 features in this work. The steps of the feature selection are;

Step 6: Employ NCA to generate 8040 features and calculated indices (*idx*).

$$idx = \text{NCA}(X, \text{target}) \quad (27)$$

Step 7: Select 1000 of the most informative/valuable features using the calculated *idx*.

$$\text{last}(d, j) = X(d, idx(j)),\ d = \{1, 2, \ldots, nOB\},\ j = \{1, 2, \ldots, 1000\} \quad (28)$$

where *last* represents the selected 1000 features and *nOB* is the number of instances/observations.

4.3 Classification

This is the last phase of the proposed work. In this phase, quadratic SVM is employed as a classifier. Various kernels are used for SVM. The second-degree polynomial order kernel is used for the SVM classifier. The MATLAB classification learner tool is used to implement this classifier, and it is named Quadratic SVM in this tool. The set parameters of the used quadratic SVM are [64];

Kernel: Polynomial.
Polynomial order: Two.
Kernel Scale: Auto.
C value (Box constraint): One.
Coding: One-vs-all.
Training and testing: Hold-out validation, 90:10.

The last step of the presented pyramid pattern-based language classification model is given below.

Step 8: Classify the selected 1000 features using the quadratic SVM classifier.

5 Results

This section provides the performance matrices of the proposed pyramid pattern-based language classification method using new big database. We have presented a novel handcrafted features-based classification model. In this work, we have employed the hold-out validation method to develop the model, with 90% of the database used for training and 10% for testing the developed model. The presented model is implemented using a desktop computer with a simple system configuration (Intel i9-9900 K microprocessor, 64 GB main memory, and Windows 10.1 operating system). MATLAB 2020 is utilized as a programming environment [65]. Accuracy, geometric mean, F1-score, and average precision values are calculated to evaluate the presented pyramid pattern-based model. The calculated results are listed in Table 3, and the developed Quadratic SVM classifier is executed 100 times to obtain robust results.

As shown in Table 3, the presented pyramid based model yielded 98.87 ± 0.30%(average ± standard deviation) classification accuracy, 98.88 ± 0.29% average precision, 98.87 ± 0.29% F1-score, and 98.83 ± 0.29% geometric mean values. The highest accuracy of 99.52%

Table 3 Various performance matrices obtained for the developed pyramid language classification model

Database	Evaluation metric	Statistics	Result (%)
DS_a	Accuracy	Standard deviation	0.30
		Minimum	97.87
		Average	98.87
		Maximum	99.52
	Average precision	Standard deviation	0.29
		Minimum	97.87
		Average	98.88
		Maximum	99.53
	F1-score	Standard deviation	0.29
		Minimum	97.87
		Average	98.87
		Maximum	99.52
	Geometric mean	Standard deviation	0.31
		Minimum	97.81
		Average	98.83
		Maximum	99.51
DS_b	Accuracy	Standard deviation	1.27
		Minimum	92.73
		Average	97.12
		Maximum	100.0
	Average precision	Standard deviation	1.15
		Minimum	93.10
		Average	97.39
		Maximum	100.0
	F1-score	Standard deviation	1.21
		Minimum	92.89
		Average	97.25
		Maximum	100.0
	Geometric mean	Standard deviation	1.37
		Minimum	91.83
		Average	96.93
		Maximum	100.0

(misclassified only seven) is obtained for our collected speech dataset (**DS_a**).

For the second speech dataset(VoxForge): our model yielded 97.12 ± 1.27% (average ± standard deviation) accuracy, 97.39 ± 1.15% average precision, 97.25 ± 1.21% F1-score, and 96.93 ± 1.37% geometric mean values.

6 Discussion

In this work, a new speech dataset for language classification is created, and also a novel pyramid pattern-based automated language classification method is proposed. The collected dataset is a comprehensive dataset with 1450 speeches in 29 languages. A new multi-leveled pyramid pattern-based feature generation model is presented to generate the most discriminative features. The main objective of the presented pyramid pattern is to show the feature generation ability of 3D shapes, and hence we have used pyramid pattern in this work. In the feature generation phase, a new 3D shape-based pattern (pyramid pattern) is presented to extract discriminative textural features. The presented pyramid pattern can extract the hidden patterns of speech effectively and hence achieve high classification performance. However, this pattern can extract low-level features. A multi-leveled feature generation method is presented by deploying a maximum pooling decomposer to generate high-level features. Furthermore, statistical features are generated in this model. Clinically significant features are selected by the NCA feature selection method. Therefore, our model yielded very high classification results for both datasets. Table 3 presents the various performance matrices obtained for the developed pyramid language classification model using *two* datasets. 1000 features are selected using NCA-like convolution neural network (CNN)-based feature generation models [66]. To select the best performing classifier, the selected features are fed to decision tree (DT) [67], linear SVM (LSVM) [68], Cubic SVM (CSVM) [69], quadratic SVM (QSVM) [64] and k nearest neighbor (kNN) [70] classifiers. It can be noted from the results that QSVM gave the best results among them. The graph of the highest accuracy (%) obtained using various classifiers with our proposed method is shown in Fig. 3.

Moreover, the chosen features are fed to neural network classifiers,: narrow neural network (NNN), medium neural network (MNN), wide neural network (WNN), bilayered neural network (BNN), and trilayered neural network (TNN). These classifiers belong to the MATLAB classification learner tool. Finally, the attributes of these classifiers are tabulated in Table 4.

The calculated maximum classification accuracies of the proposed models using neural network classifiers and SVM (it is the best classifier) are depicted in Fig. 4.

Figures 2 and 3 show that the most appropriate classifier is Quadric SVM for the proposed model. To get comparative results, the first (collected) dataset has been used.

In the literature, limited works have been presented on automated language classification. The prior models used low dimensional datasets, and the previously used datasets included a limited number of languages. Table 5 summarizes the accuracy (%) obtained using state-of-the-art automated speech-based language classification methods.

In Table 5, the bold values show the results of our method. It can be noted from Table 5 that other authors have used small datasets with the limited number of

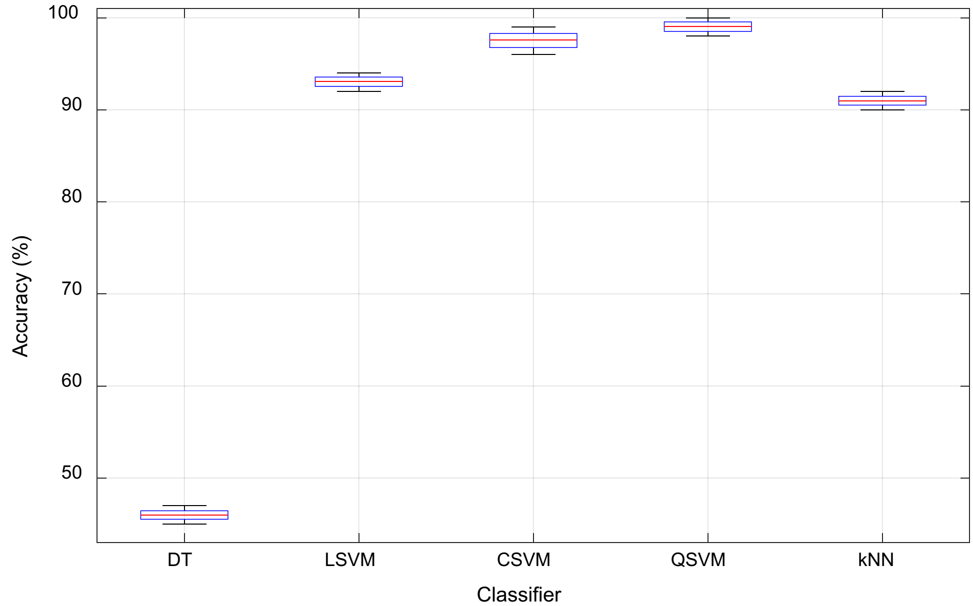

Fig. 3 Highest accuracy (%) obtained using various classifiers for our proposed method

Table 4 Details of the neural network classifiers used in this work

Classifier	First layer	Second layer	Third layer	Activation	Iteration limit
NNN	10	–	–	ReLu	1000
MNN	25	–	–	ReLu	1000
WNN	100	–	–	ReLu	1000
BNN	10	10	–	ReLu	1000
TNN	10	10	10	ReLu	1000

languages. Our dataset is more extensive than others (sizewise), with 29 languages. Also, the presented pyramid pattern-based model yielded 99.52% classification accuracy using our dataset. This result is higher than deep learning-based language classification models [55, 77]. Also, the presented pyramid pattern-based language classification model is tested on the VoxForge dataset with 16 languages. It obtained 97.12 $\pm$ 1.27% accuracy. The best result obtained for this dataset is 94.6% in classifying four languages [3]. Our presented model obtained better classification performance than others.

The benefits of this work are;

- A new micro structure (image descriptor) called pyramid pattern is used as a feature generator for speeches. In addition, the pyramid pattern is a very effective feature generation function for language classification.
- We have obtained the highest classification accuracy of 99.52% accuracy database in classifying 29 languages with tenfold cross-validation strategy.
- To the best of our knowledge, this is the first work to classify 29 languages and obtain high classification accuracy with a big database.

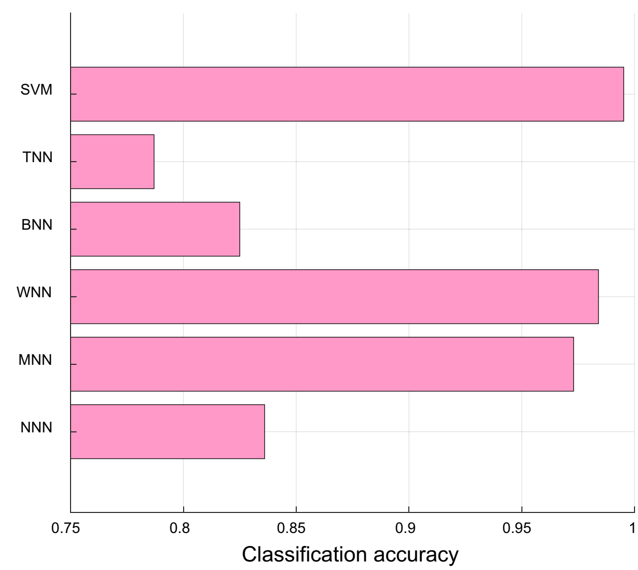

Fig. 4 Classification accuracies obtained using neural network and SVM classifiers. The NNN, MNN, WNN, BNN, and TNN classifiers achieved 83.62, 97.34, 98.40, 82.08, and 77.81% classification accuracies, respectively. The used SVM classifier reached 99.52% accuracy. This figure depicts that the best NN (WNN) attained 1.12% (= 99.52–98.40) lower accuracy than SVM

Table 5 Summary of accuracies (%) obtained using state-of-the-art automated speech-based language classification methods

Study	Features extraction method	Classifier	Number of languages	Dataset	Number of samples/ utterances	Accuracy
[3]	MFCC, Perceptual linear prediction, Relative perceptual linear prediction features	Feedforward back-propagation neural network	4	Voxforge [71, 72]	200 utterances	94.60%
[50]	MFCC	Artificial neural network, SVM, GMM	7	OGI-MLTS [73]	200 samples	73.40–80.00%
[51]	Perceptual linear prediction, Bark frequency cepstral coefficient, MFCC	GMM	3	CMDNYC [74]	70 utterances	88.75%
[52]	Short-Time Energy, MFCC	SVM, Random forest	1. 15 2. 28 3. 6	1. Collected data 2. Collected data 3. Voxforge [71, 72]	1. Unspecified 2. Unspecified 3. 90.000 samples	Ds1: 89.63% with RF classifier Ds2: 87.70% with RF classifier Ds3: 91.23% with SVM classifier
[53]	Hidden Markov models	SVM, Neural Network	4	Shtooka [75], Voxforge [71, 72] and Youtube	23.000 utterances	70.00%
[55]	MFCC, GMM	CNN	5	Collected data	500 samples	95.10%
[76]	Polymer pattern, Tent maximum absolute pooling	kNN	1. 45 2. 16	1. LI45 [76] 2. VoxForge [71, 72]	1. 4500 samples 2. 1650 samples	1. 97.87 2. 99.70
[77]	MFCC, GMM	Deep probabilistic neural network	Unspecified	CSTR VCTK [78]	20.000 utterances	87.78
Our method	**Pyramid pattern and maximum pooling based feature generation network**	**Quadratic SVM**	**29**	**Our collected dataset** **29 languages**	**14,500 samples**	**98.87% ± 0.30% (Average ± standard deviation)** **99.52% (Maximum)**
Our method	**Pyramid pattern and maximum pooling based feature generation network**	**Quadratic SVM**	**16**	**VoxForge dataset** **16 languages**	**1650 samples**	**97.12% ± 1.27% (Average ± standard deviation)** **100.0% (Maximum)**

- The developed model yielded the highest performance using both databases. This confirms the superiority of the proposed method.

The main limitation of this work is that the proposed pyramid pattern feature extraction is computationally intensive and takes time.

We intend to make a new language classification project for information security in the future. In this work, bigger databases (including more accents and languages) can be collected for language classification. New feature extraction functions can be presented like pyramid pattern using 3D shapes and special graphs. Also, new generation deep learning methods can be presented by using shape-based feature generators. Also, an accent detection application can be developed for refugees. The snapshot of the future application of this work is demonstrated in Fig. 5.

The intended project can be used for both language identification of the refugees and creating a language identification tool for digital forensics examiners.

7 Conclusion

Most of the language classification models have used low-dimensional datasets and have not obtained high classification rates. Hence, a new big speech dataset is created to

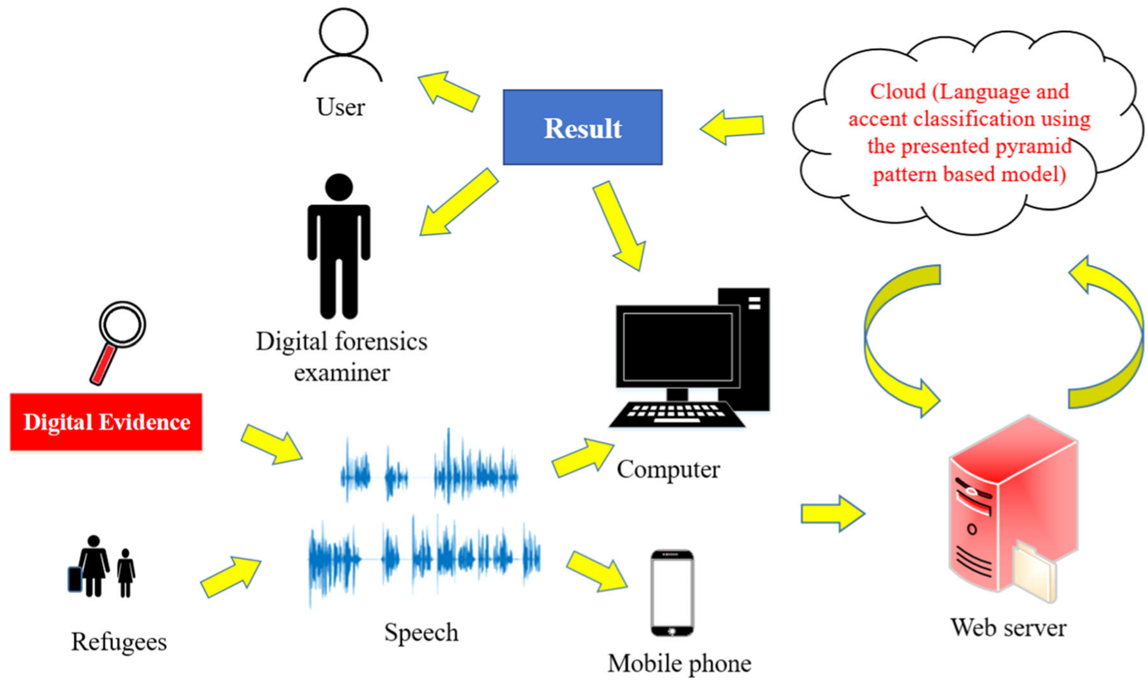

Fig. 5 Snapshot of the intended project. The intended project can be used for both language identification of the refugees and creating a language identification tool for digital forensics examiners

evaluate our novel pyramid pattern-based language classification architecture. The pyramid pattern-based model is an alternative for recurrent neural networks (RNN). The presented pyramid pattern-based model has a nonparametric feature generation and selection process. We have obtained the highest classification accuracy of 99.52%, average precision rate of 99.53%, F1-score of 99.52%, and the geometric mean of 99.51% using a pyramid pattern-based language classification model with our newly created dataset. Also, it yielded the classification accuracy of 97.12 ± 1.27%, average precision of 97.39 ± 1.15%, F1-score of 97.25 ± 1.21%, and geometric mean value of 96.93 ± 1.37% using VoxForge dataset. These results confirm the robustness and accuracy of the developed model. In the future, we intend to use this system to test more languages and accents and employ it for real-life applications.

In this research, we proposed a local feature extraction function. This model is named pyramid pattern, and the pyramid pattern can easily detect differences in speech signals. Therefore, our future project is to detect accents using the proposed 3D shape-based textural feature extractor and develop versions of our proposed pyramid pattern. Moreover, we will use the commonly known 3D shape to extract textural features, and we will propose a self-organized architecture to attain high classification performance on the accent datasets.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Demuro E, Gurney L (2021) Languages/languaging as world-making: the ontological bases of language. Lang Sci 83:101307
2. Das RK, Prasanna SM (2018) Speaker verification from short utterance perspective: a review. IETE Tech Rev 35(6):599–617
3. Deshwal D, Sangwan P, Kumar D (2020) A language identification system using hybrid features and back-propagation neural network. Appl Acoust 164:107289
4. Krčadinac O, Šošević U, Starčević D (2021) Evaluating the performance of speaker recognition solutions in E-Commerce applications. Sensors 21(18):6231
5. Ambikairajah E, Li H, Wang L, Yin B, Sethu V (2011) Language identification: a tutorial. IEEE Circuits Syst Mag 11(2):82–108
6. Muthusamy YK, Barnard E, Cole RA (1994) Reviewing automatic language identification. IEEE Signal Process Mag 11(4):33–41
7. Besacier L, Barnard E, Karpov A, Schultz T (2014) Automatic speech recognition for under resourced languages: a survey. Speech Commun 56:85–100
8. Singh G, Sharma S, Kumar V, Kaur M, Baz M, Masud M (2021) Spoken language identification using deep learning. Comput Intell Neurosci 2021:5123671. https://doi.org/10.1155/2021/5123671
9. Stutzman K (2007) The effects of digital audio files and online discussions on student proficiency in a foreign language. Iowa State University, Iowa
10. Wahlster W (2013) Verbmobil: foundations of speech-to-speech translation. Springer, Cham

11. Waibel A, Jain AN, McNair AE, Saito H, Hauptmann AG, Tebelskis J (1991) JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies. In: Acoustics, speech, and signal processing, IEEE international conference on, 1991. IEEE Computer Society, pp 793–796
12. Nakamura S, Markov K, Nakaiwa H, Kikui G-i, Kawai H, Jitsuhiro T, Zhang J-S, Yamamoto H, Sumita E, Yamamoto S (2006) The ATR multilingual speech-to-speech translation system. IEEE Trans Audio Speech Lang Process 14(2):365–376
13. Basu J, Majumder S (2020) Identification of seven low-resource North-Eastern languages: an experimental study. In: Intelligence Enabled Research. Springer, Cham, pp 71–81
14. Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P (2010) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(4):788–798
15. Fer R, Matějka P, Grézl F, Plchot O, Veselý K, Černocký JH (2017) Multilingually trained bottleneck features in spoken language recognition. Comput Speech Lang 46:252–267
16. Liu G, Sadjadi SO, Hasan T, Suh J-W, Zhang C, Mehrabani M, Boril H, Sangwan A, Hansen JH (2011) UTD-CRSS systems for NIST language recognition evaluation 2011. In: NIST 2011 Language recognition evaluation workshop, Atlanta, USA, pp 6–7
17. Singer E, Torres-Carrasquillo P, Reynolds DA, McCree A, Richardson F, Dehak N, Sturim D (2012) The MITLL NIST LRE 2011 language recognition system. In: Odyssey 2012-the speaker and language recognition workshop, 2012
18. Zhang Q, Liu G, Hansen JH (2014) Robust language recognition based on diverse features. In: ODYSSEY: The speaker and language and language recognition workshop, pp 152–157
19. Dustor A, Szwarc P (2010) Spoken language identification based on GMM models. In: ICSES 2010 international conference on signals and electronic circuits, 2010. IEEE, pp 105–108
20. Bharali SS, Kalita SK (2015) A comparative study of different features for isolated spoken word recognition using HMM with reference to Assamese language. Int J Speech Technol 18(4):673–684
21. Gelly G, Gauvain J-L, Le VB, Messaoudi A (2016) A divide-and-conquer approach for language identification based on recurrent neural networks. In: INTERSPEECH, 2016. pp 3231–3235
22. Bhatia M, Singh N, Singh A (2015) Speaker accent recognition by MFCC Using KNearest neighbour algorithm: a different approach. Int J Adv Res Comput Commun Eng 4(1):153–155
23. Abbas AW, Ahmad N, Ali H (2012) Pashto Spoken Digits database for the automatic speech recognition research. In: 18th International Conference on Automation and Computing (ICAC), 2012. IEEE, pp 1–5
24. Hautamäki V, Siniscalchi SM, Behravan H, Salerno VM, Kukanov I (2015) Boosting universal speech attributes classification with deep neural network for foreign accent characterization. In: Sixteenth annual conference of the international speech communication association, 2015
25. Rao K, Sak H (2017) Multi-accent speech recognition with hierarchical grapheme based models. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2017. IEEE, New York, pp 4815–4819
26. Barua PD, Dogan S, Tuncer T, Baygin M, Acharya UR (2021) Novel automated PD detection system using aspirin pattern with EEG signals. Comput Biol Med 137:104841
27. Aydemir E, Tuncer T, Dogan S, Gururajan R, Acharya UR (2021) Automated major depressive disorder detection using melamine pattern with EEG signals. Appl Intell 51(9):6449–6466
28. Tuncer T, Dogan S, Baygin M, Acharya UR (2022) Tetromino pattern based accurate EEG emotion classification model. Artif Intell Med 123:102210
29. Zubair S, Yan F, Wang W (2013) Dictionary learning based sparse coefficients for audio classification with max and average pooling. Digital Signal Process 23(3):960–970
30. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, 2012. pp 1097–1105
31. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. pp 1–9
32. Raghu S, Sriraam N (2018) Classification of focal and non-focal EEG signals using neighborhood component analysis and machine learning algorithms. Expert Syst Appl 113:18–32
33. Deshwal D, Sangwan P, Kumar D (2019) Feature extraction methods in language identification: a survey. Wireless Pers Commun 107(4):2071–2103
34. Li H, Ma B, Lee KA (2013) Spoken language recognition: from fundamentals to practice. Proc IEEE 101(5):1136–1159
35. Jothilakshmi S, Ramalingam V, Palanivel S (2012) A hierarchical language identification system for Indian languages. Digital Signal Processing 22(3):544–553
36. Li K-P (1997) Automatic language identification/verification system. Google Patents
37. Dey S, Rajan R, Padmanabhan R, Murthy HA (2011) Feature diversity for emotion, language and speaker verification. In: 2011 National Conference on Communications (NCC), 2011. IEEE, New York, pp 1–5
38. Morales L, Li FF (2018) A new verification of the speech transmission index for the English language. Speech Commun 105:1–11
39. Wong K-YE (2004) Automatic spoken language identification utilizing acoustic and phonetic speech information. Queensland University of Technology
40. Grachev AM, Ignatov DI, Savchenko AV (2019) Compression of recurrent neural networks for efficient language modeling. Appl Soft Comput 79:354–362
41. Lyu D-C, Chng E-S, Li H (2013) Language diarization for conversational code-switch speech with pronunciation dictionary adaptation. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing, 2013. IEEE, pp 147–150
42. Makowski R, Hossa R (2020) Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise. Appl Acoust 166:107344
43. Tan Z-H, Dehak N (2020) rVAD: an unsupervised segment-based robust voice activity detection method. Comput Speech Lang 59:1–21
44. Zhu M, Wu X, Lu Z, Wang T, Zhu X (2019) Long-term speech information based threshold for voice activity detection in massive microphone network. Digital Signal Process 94:156–164
45. Shin JW, Chang J-H, Kim NS (2010) Voice activity detection based on statistical models and machine learning approaches. Comput Speech Lang 24(3):515–530
46. Abraham J, Khan AN, Shahina A (2021) A deep learning approach for robust speaker identification using chroma energy normalized statistics and mel frequency cepstral coefficients. Int J Speech Technol, pp 1–9
47. Kingsbury B, Saon G, Mangu L, Padmanabhan M, Sarikaya R (2002) Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. IEEE, New York, pp I-53–I-56
48. Nemer E, Goubran R, Mahmoud S (2001) Robust voice activity detection using higher-order statistics in the LPC residual domain. IEEE Trans Speech Audio Process 9(3):217–231

49. Park TJ, Kanda N, Dimitriadis D, Han KJ, Watanabe S, Narayanan S (2022) A review of speaker diarization: recent advances with deep learning. Comput Speech Lang 72:101317
50. Bhanja CC, Laskar MA, Laskar RH (2019) A pre-classification-based language identification for Northeast Indian languages using prosody and spectral features. Circuits Systems Signal Process 38(5):2266–2296
51. Kumar P, Biswas A, Mishra AN, Chandra M (2010) Spoken language identification using hybrid feature extraction methods. arXiv preprint arXiv:10035623
52. Yasmin G, Das AK, Nayak J, Pelusi D, Ding W (2020) Graph based feature selection investigating boundary region of rough set for language identification. Expert Syst Appl, p 113575
53. Gazeau V, Varol C (2018) Automatic spoken language recognition with neural networks. Int J Inf Technol Comput Sci(IJITCS) 10(8):11–17
54. Safitri NE, Zahra A, Adriani M (2016) Spoken language identification with phonotactics methods on minangkabau, sundanese, and javanese languages. Proc Comp Sci 81:182–187
55. Saleem S, Subhan F, Naseer N, Bais A, Imtiaz A (2020) Forensic speaker recognition: A new method based on extracting accent and language information from short utterances. Forensic Sci Int Digital Invest 34:300982
56. VoxForge (2020) Open source speech corpus. http://www.voxforge.org/
57. YouTube (2020) www.youtube.com
58. NHC (2020) https://www.nch.com.au/wavepad/index.html
59. Savchenko AV, Savchenko LV (2015) Towards the creation of reliable voice control system based on a fuzzy approach. Pattern Recogn Lett 65:145–151
60. Reddy VR, Maity S, Rao KS (2013) Identification of Indian languages using multi-level spectral and prosodic features. Int J Speech Technol 16(4):489–511
61. Kuncan F, Kaya Y, Kuncan M (2019) Sensör işaretlerinden cinsiyet tanıma için yerel ikili örüntüler tabanlı yeni yaklaşımlar. J Faculty Eng Archit Gazi Univ 34(4)
62. Zhang Z Improved adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), 2018. IEEE, New York, pp 1–2
63. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
64. Nakano T, Nukala BT, Tsay J, Zupancic S, Rodriguez A, Lie DY, Lopez J, Nguyen TQ (2017) Gaits classification of normal vs. patients by wireless gait sensor and Support Vector Machine (SVM) classifier. Int J Softw Innovation (IJSI) 5(1):17–29
65. Aljerf L (2016) Reduction of gas emission resulting from thermal ceramic manufacturing processes through development of industrial conditions. Sci J King Faisal Univ 17(1):1–10
66. Tuncer T, Ertam F, Dogan S, Aydemir E, Pławiak P (2020) Ensemble residual network-based gender and activity recognition method with signals. J Supercomput 76(3):2119–2138
67. Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21(3):660–674
68. Cao X, Wu C, Yan P, Li X Linear SVM classification using boosting HOG features for vehicle detection in low-altitude airborne videos. In: 2011 18th IEEE international conference on image processing, 2011. IEEE, New York, pp 2421–2424
69. Jain U, Nathani K, Ruban N, Raj ANJ, Zhuang Z, Mahesh VG Cubic SVM classifier based feature extraction and emotion detection from speech signals. In: 2018 international conference on sensor networks and signal processing (SNSP), 2018. IEEE, New York, pp 386–391
70. Maillo J, Ramírez S, Triguero I, Herrera F (2017) kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data. Knowl-Based Syst 117:3–15
71. VoxForge (2020) VoxForge, Free Speech Recognition, www.voxforge.org
72. Lounnas K, Abbas M, Teffahi H, Lichouri MA (2019) language identification system based on voxforge speech corpus. International conference on advanced machine learning technologies and applications. Springer, Cham, pp 529–534
73. Muthusamy YK, Cole RA, Oshika BT The OGI multi-language telephone speech corpus. In: Second International Conference on Spoken Language Processing, 1992
74. Design CM (2020) https://www.cmdnyc.com/
75. Shtooka (2020) http://shtooka.net/
76. Tuncer T, Dogan S, Akbal E, Cicekli A, Acharya UR (2021) Development of accurate automated language identification model using polymer pattern and tent maximum absolute pooling techniques. Neural Comput Appl 34(6):4875–4888. https://doi.org/10.1007/s00521-021-06678-0
77. Bansal P, Singh V, Beg M (2019) A multi-featured hybrid model for speaker recognition on multi-person speech. J Electrical Eng Technol 14(5):2117–2125
78. Yamagishi J, Veaux C, MacDonald K (2019) CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.