A State of Art of Machine Learning Algorithms Applied Over Language Identification and Speech Recognition Models

A. Razia Sulthana and Aakansha Mathur

Abstract Over the years, we have observed several applications of machine learning. One of the applications is in speech recognition. One of the sub-topics in speech recognition is language identification. A typical language identification system involves a feature extraction and classification stage. Machine learning classifiers have been extensively used in the classification stage of the LID system. The current paper examines the chronological involvement of machine learning in the LID system. The paper also details the chronological development and refinement of the LID system.

Keywords Machine learning · Speech recognition · Support vector machines · Classification · Language

1 Introduction

The objective of a language identification (LID) system is to determine the correct language in the audio speech signal. A LID system generally consists of two stages, feature extraction stage and classification stage. The feature extraction stage also consists of some pre-processing steps such as pre-emphasis, framing and windowing. The feature extraction stage involves different types of speech features. The speech features are broadly classified in high-level features and low-level features. Acoustic, phonotactic and prosodic features are low-level features. Lexical and syntactic features are high-level features. The classification stage involves using different machine learning classifiers like support vector machine (SVM), Gaussian mixture models (GMM) and K-nearest neighbors.

Most researchers have proposed LID systems with low-level feature extraction techniques. An additional stage, normalization, has been applied to refine the LID

A. Razia Sulthana · A. Mathur (✉)
BITS Pilani Dubai Campus, Dubai, United Arab Emirates

A. Razia Sulthana
e-mail: razia@dubai.bits-pilani.ac.in

R. J. Kannan et al. (eds.), *International Virtual Conference on Industry 4.0*,
Lecture Notes in Electrical Engineering 355,
https://doi.org/10.1007/978-981-16-1244-2_10

system results. Other refinements to the LID system include splitting the classification stage into the learning and recognition stage. The current paper looks at chronological development of research in language identification in recent years. The next six sections explain a few of the LID models and finally the conclusion.

2 Literature Review

2.1 Language Identification System—I

The research proposal in [1] develops an automatic language identification system that classifies languages into modern standard Arabic (MSA) and Kabyle, two commonly spoken languages in Algeria. The research uses two different databases. One of the databases is in MSA, and the other database is in Kabyle dialect. The databases do not have an equal number of samples. All the speakers were Algerian, so they were fluent in MSA and Kabyle. Moreover, the speakers spoke the sentences with dialect. The research creates a bilingual database by taking 10 sentences from MSA. The MSA 10 sentences were spoken by 13 different speakers and were repeated 10 times. This makes the total number of utterances to be 1300. The sampling frequency of each MSA sentence is 44,100 Hz, and each sentence is coded in 16 bits. The sentences selected were phonetically balanced. This was done so that the greatest number of phonemes can be obtained without varying the frequency of each sentence. For the Kabyle database, the research took dialogs spoken by four male and two female speakers. And, these dialogs were repeated five times. This makes the total number of Kabyle utterances to be 720. Moreover, the Kabyle utterances were sampled at a frequency of 16,000 Hz. The Kabyle utterances, like MSA utterances, were coded in 16 bits. The system proposed by the research involves two phases: the learning phase and recognition phase. Both the phases have the same steps. The only difference is that in the learning phase, the language of the speech sentence sample is known while in the recognition phase, the language of the sentence is unknown. The steps for both the phases are as follows: pre-processing, feature extraction, Arabic and Kabyle Language Model, decision and identify language. The research has extracted two types of low-level features: prosodic features and acoustic features. Melody and stress are two types of prosodic features extracted by the research. The research extracted MFCC acoustic features. A melody is made up of high- and low-pitch notes. Melody is a prosodic feature which represents the speech signal frequency in function of time. Melody is a speaker-dependent feature and is related to the speaker's emotions. The research used PRAAT software to extract the melody feature. Stress is also a prosodic feature which represents the speech signal's intensity. The research used support vector machines as classifiers for its language identification system. For the learning phase, the research took 16 Kabyle sentences from 24 Kabyle sentences and 6 MSA sentences from 10 MSA sentences. The research first used prosodic features to classify unknown audio signals into one of the two languages.

2.2 Language Identification System—II

A learning model is proposed in [2] for a language identification system with optimized machine learning algorithms. The input signals are converted into frames of length 25 s. These frames had an overlap of 20 s. Following which vocal tract length normalization (VTLN) is applied on the input speech signal to obtain seven mel-frequency cepstral coefficients (MFCC). The research proceeded by performing cepstral mean and variance normalization along with representations relative spectra RASTA filtering. Next, the shifted delta cepstral (SDC) features are calculated.

The extreme learning machine (ELM) approach involves calculating the output weights of a neural network using least-squares solutions. The technique involves generating biases randomly for hidden layer weights. A disadvantage of ELM, as mentioned by the research, is that it does not have a specific approach for determining the input-hidden layer weights. Therefore, the ELM becomes subject to local minima. So, optimal weights must be identified. The identification of optimal weights can be done by using an optimization approach. This will ensure that ELM gives its best performance. Further, the ameliorated teaching–learning-based optimization (ATLBO) algorithm is applied. For the language identification system, the research worked with eight different languages. The research obtained audio files for the eight languages from broadcasting channels. The dataset consisted of 120 records for the 8 languages. And, the total number of features extracted was 600. There were 15 utterances for each language, and each utterance was of 30 s length. The dataset was split into training and testing dataset. The training set was 67% of the entire dataset. The testing set was 33% of the entire dataset. The research selected enhanced ameliorated teaching–learning-based optimization (EATLBO) and enhanced self-adjusting extreme learning machine (ESA-ELM) as optimization approaches for the language identification system. The research evaluated EATLBO by comparing it with ATLBO. Each experiment involved applying a different mathematical function for optimization.

2.3 Language Identification System—III

The objective of this proposal [3] was to build a learning model for language identification systems. The study used an extreme learning machine (ELM) as a learning model for the language identification system. The first step was to extract the features. A drawback of ELM is that it gets subjected to local minima due to the absence of a specific approach to determine input-hidden layer weights. The drawback can be overcome by using optimization approaches which will give optimal weights for the ELM. The research used genetic algorithms for optimization tasks. The purpose of the genetic algorithms is to discover the optimal hidden layer weights and biases. A genetic algorithm involves: initial population, evaluation, selection and genetic

operators. The genetic algorithm will enhance the performance of the ELM. Generally, only one criteria are used by genetic algorithms during the selection process for optimization. Due to the shortcomings of genetic algorithms, the research optimized the genetic algorithm and proposed a new optimized genetic algorithm (OGA). The optimized genetic algorithm (OGA) has three selection criteria: roulette wheel, K-tournament and random. These criteria are used for parent selection. The parents are selected for generation of two new offspring for a new population in the crossover operation. The OGA is now used to enhance the performance of the ELM. The OGA-ELM has the three different selection criteria: selection, crossover and mutation process which optimizes the input weight values and hidden nodes bias. The parameter settings for ELM are as follows: The number of hidden nodes is 700–900, the output neurons are class values assigned to each language, the activation function applied is sigmoid activation function, the input weights range between –1 and 1, and the bias values range 0–1. The parameter settings for OGA are as follows: The number of iterations is 100, the population size is 50, the crossover operation is arithmetic operation, the mutation operation is uniform mutation, the crossover population is 70% of the population, the mutation population is 30% of the population, the gamma value is 0.4, and the tournament size is 3. The research worked with eight different languages. The class values for each language are as follows: 1 is for Arabic, 2 is for English, 3 is for Malay, 4 is for French, 5 is for Spanish, 6 is for German, 7 is for Persian, and 8 is for Urdu. The research architecture performs the following steps. The first step is random initialization of input weights and bias, determining the population size, objective function and maximum iteration. The third step is tabulating the fitness value. The fourth step is initialization of the crossover and mutation populations. The fifth step involves two parts. The first part is if the crossover population is less than or equal to 70%, then parents are selected based on the selected criteria. After which, crossover is performed, and the two offspring generated are saved in the new crossover population. The new crossover population is sent to the first part of the fifth step. Suppose the crossover population is greater than 70%, then the algorithm goes to the second part of the fifth step. The second part itself has if–then condition. If the population is less than or equal to 30%, then select parents randomly and perform mutation. The child created is added to the mutation population. After which merge the mutation and crossover population. However, if the mutation population is greater than 30%, then simply merge the crossover and mutation population. After merging, the algorithm checks if the termination criteria are satisfied. If it is not satisfied, then the algorithm goes back to calculation of the fitness value step (third step). However, if the termination criteria are satisfied, then the global optimum parameters are discovered. This marks the end of the OGA-ELM algorithm.

2.4 Language Identification System—IV

The research proposed [4] a system that performs three tasks: speaker identification, gender identification and language identification. The system classifies speakers into eight categories, languages into three categories and gender into two categories. The authors created their own database. The database primarily consists of speech recordings in Arabic, English and Polish. The speech recordings were taken from online and TV broadcasts. There were eight speakers for each language. So, in total, the dataset consisted of 24 speech recordings. The speakers were male and female. The database consisted of talks of 2 h 20 min and 27 s. The speech recordings had a frequency of 44 kHz. Firstly, the speech recordings are pre-processed. The pre-processing step involved eliminating the silence in the recordings. Moreover, low-energy fragments are eliminated from the input file. After pre-processing, the input file time length was reduced to 2 h 7 min and 22 s. The research used radial neural network (RNN), probabilistic neural network (PNN) and the used long short-term memory recursive neural network LSTM neural network (NN) architecture for classification. The LSTM NN architecture consists of six layers. The input layer consists of input feature vectors. The length of the input features depends upon the features extracted. The research extracted Burg's estimation, TM-eigenvalues and MFCC features. The feature vector length of Burg's estimation is 129. The feature vector length of TM-eigenvalues is 125. The feature vector length of MFCC is 23. After the input layer, the LSTM NN consists of 24 hidden units. After the hidden layer, there is a dropout layer with dropout possibility of 50%. Then, there is a fully connected layer (fc) consisting of 24 neurons. Following the fc layer, there is a SoftMax layer. The purpose of the SoftMax layer is to apply SoftMax function to the input. Finally, there is an output layer which calculates the cross-entropy loss function. The number of neurons in LSTM and output layer has been reduced for language and gender identification tasks because of the small number of final categories. The dataset was distributed in the following way. Firstly, the dataset was split into Arabic, English, Polish, Male and Female subset. Each subset was further split into a training set and testing set. Furthermore, 30 s of each speaker was taken. So, the Arabic training set consisted of 30 speakers. This totaled to 4 min for Arabic training set. For the Arabic testing set, the remaining speech signal is taken. This process is repeated for English, Polish, Male and Female subsets. Hence, the training data was 9% of the complete database. While the testing data was 91% of the complete database, the training–testing ratio was 9:91. These features were inputted into probabilistic NN and radial NN for speaker, language and gender identification.

2.5 Language Identification System—V

A two-stage language identification system (TS-LID) [5] is proposed for Indian languages. The research used the two databases for experimentation. The first

database consists of telephone speech 11 languages. The telephone speech is at a frequency of 8 kHz. The speech includes 90 utterances from 90 different speakers per language. Out of the 11 languages, only 2 are tonal languages. The remaining languages are non-tonal languages. The second database consists of studio quality 12 Indian languages. There were five tonal languages and seven non-tonal languages. The research has analyzed the performance of MHEC and MFCC features individually and in combination. The first step in the classification stage involves a tonal/non-tonal pre-classification. Here, i-vector-based SVM will be used as classifiers. The second step involves language identification. The architecture proposed by the research is subjected to three conditions. The first condition is a conventional language identification system where each language has a model, and the model is trained. So, during the identification stage, it is determined which language model is the most likely for a given test sample. The second condition involves the languages to be first pre-classified into tonal and non-tonal. The pre-classification step is the first step. Depending upon the outcome of the first step, irrespective of its correctness, the speech signal is routed to either the tonal or non-tonal module of the second step. The third condition also involves a pre-classification step but only those samples which are correctly classified into tonal or non-tonal are forwarded to the next step, which is the language identification step.

2.6 Language Identification System—VI

The research [6] proposed a language identification system that recognizes four languages: Tamil, Malayalam, Hindi and English. In fact, a number of recommendation systems [7–9] are supportive to be modified into speech recognition systems; in future, 50 utterances of each language are present in the dataset. Moreover, the dataset consists of equal numbers of male and female speakers. The total number of male speakers for each language is five. The total number of female speakers for each speaker is five. Studio quality acoustic systems are used to record the speeches by these speakers. The research used Sound Forge software for isolation of words and file organization. The audio recordings are saved in Mono Wav File Format. The sampling frequency of the speech recordings is 16 kHz. Once the words are isolated from each speech, a total of 200 input speech signals is obtained. This is primarily the pre-processing step. The speech signals now undergo pre-emphasis. The next step is the feature extraction step. The research extracts several features and experiments upon the combination of the features as well. The features extracted by research are MFCC, perceptual linear prediction features (PLP) relative perceptual linear prediction features (RASTA-PLP) and shifted delta cepstrum (SDC). 40 filters filter bank is used in MFCC. Of these 40 filters, 13 are linear filters and 27 are logarithmic filters. For each isolated word, a thirteen-coefficient matrix is obtained. The next step is the classification step. The research uses feed-forward back-propagation neural networks (FFBPNN) as a classifier. The features extracted from the earlier step are stored as feature matrices. The learning algorithms used are trainlm and trainscg. The objective

Table 1 Comparative analysis of speech features and machine learning method used by different papers

References	Speech features	Machine learning
[1]	MFCC, melody and stress	SVM
[2]	MFCC, cepstral mean, variance normalization along with RASTA filtering, SDC	Extreme learning machine
[3]	MFCC, SDC, cepstral mean and variance normalization with RASTA filtering	Extreme learning machine
[4]	MFCC	Radial neural network (RNN), probabilistic neural network (PNN) and LSTM
[5]	MFCC, MHEC	SVM
[6]	MFCC, PLP, SDC	Feed-forward backpropagation neural networks

of using two different learning algorithms is to identify which learning algorithm is the best to train the neural network. Moreover, nonlinear sigmoid activation function was used and softmax activation function. To evaluate the performance of FFBPNN, the research obtains error and accuracy rate for training, testing and validation sets. The research uses 100 hidden neurons in FFBPNN when the "trainlm" algorithm is used. The research uses 30 hidden neurons in FFBPNN when the "trainscg" algorithm is used.

3 Comparative Analysis

From Table 1, we can observe that MFCC features are extracted by all the papers. This is because MFCC features give better performance for the various LID systems built by different researchers.

4 Results Analysis

This section primarily discusses the results obtained by different researchers. Moreover, the limitation of the research is also addressed.

The melody features when classified using SVM by [1], the accuracy for MSA and Kabyle was 90.41 and 82.08, respectively. The average accuracy for melody features was 86.25. When stress features were used for language identification, the accuracy for MSA and Kabyle was 98.75 and 91.67, respectively. The average accuracy for melody features was 95.2. The research then combined the melody and stress

features. When the combined features were used for language identification, the accuracy for MSA and Kabyle was 99.17 and 91.67, respectively. The average accuracy for combined features was 95.42. 238 MSA samples were correctly classified into MSA language when combined features were used. 220 Kabyle samples were correctly classified into Kabyle language when combined features were used. 2 MSA samples were incorrectly classified into Kabyle language when combined features were used. 20 Kabyle samples were incorrectly classified into MSA language when combined features were used. The research then proceeded by using only MFCC features for language identification. The research obtained an accuracy of 97.91 for MSA language and 93.75 for Kabyle language. The average accuracy was 95.83. 235 MSA samples were correctly classified into MSA language when MFCC features were used. 225 Kabyle samples were correctly classified into Kabyle language when MFCC features were used. 5 MSA samples were incorrectly classified into Kabyle language when MFCC features were used. 15 Kabyle samples were incorrectly classified into MSA language when MFCC features were used. The research then combined melody, stress and MFCC features. The research obtained an accuracy of 98.75 for MSA language and 96.25 for Kabyle language. The average accuracy was 97.5. The research reinstated the fact that hybridization of acoustic and prosodic features increased the accuracy of language identification. The research used only two types of prosodic characteristics, melody and stress. Moreover, the research did not attempt to optimize the learning phase.

Albadr et al. [2] observed that EATLBO performs better than ALTBO. The research also evaluated ESA-ELM on the basis of multiple parameters of the learning model. The highest accuracy for ESA-ELM was 96.25%. The research inferred that ESA-ELM is better suited for language identification systems. The research also used the elitist genetic algorithm to enhance extreme learning machines. This approach was called the elitist genetic algorithm extreme learning machine (EGA-ELM). The highest accuracy for EGA-ELM was obtained when 750 hidden neurons were used. The lowest accuracy for EGA-ELM was obtained when 900 hidden neurons were used. However, the accuracy of ESA-ELM was higher than the accuracy of EGA-ELM for all iterations. Thus, the research reinstated the fact that ESA-ELM is better suited for language identification systems. The research did not experiment on time and cost optimization of front-end feature extraction. The research did not explore the use of metaheuristic algorithms in ESA-ELM for optimization of weights.

The OGA-ELM achieved 100% accuracy for K-tournament, 99.50% accuracy for roulette wheel and 99.38% for random. The research does not consider real-time aspects such as noise that could affect the performance of the model proposed by [3].

The accuracy or recognition rate for the nine cases (3 features $\times$ 3 identifications) was tabulated by [4]. The research deduced that probabilistic NN works better with longer speech signals. The LSTM NN performed well for all three identification cases. The research took a simple dataset, and the performance of LSTM NN was not observed on complex datasets.

Bhanja et al. [5] noted that the pre-classification step gave better performance for OGI-multilingual database than NIT Silchar language database (NITS-LD). That is,

the language identification accuracy was higher for OGI-MLTS than that of NITS-LD when the pre-classification step was applied. The research did not study the use of modified neural networks to help reduce processing time.

Deshwal et al. [6] deduced that highest FFBPNN accuracy and lowest testing error are achieved (0.10). The research also varied the number of epochs to observe the performance of the neural network. The epochs range from 30 to 60. It was observed that as the number of epochs increased, the classification accuracy also increased. It was also inferred that the "trainlm" algorithm gives better performance than "trainscg" algorithm. A limitation and future expansion of the research are using the proposed language identification system for continuous speech signals and using more training functions.

5 Conclusion

The paper examined the contribution of machine learning in LID systems from 2016 to 2020. It can be observed that the contribution of machine learning has increased in LID systems over the recent years. In this paper, we have seen how different machine learning classifiers such as SVM, neural networks and decision trees have been used in the classification stage of the LID system. Moreover, we have also seen how using a machine learning classifier in a pre-classification stage can improve the accuracy of the model. Some researchers have optimized the training parameters and observed the performance of the LID system. Further research can involve using continuous speech signals in the LID system.

References

1. Lounnas K, Demri L, Falek L, Teffahi H (2018, October). Automatic language identification for Berber and Arabic languages using prosodic features. In: 2018 International conference on electrical sciences and technologies in Maghreb (CISTEM). IEEE, pp 1–4
2. Albadr MAA, Tiun S, AL-Dhief FT, Sammour MA (2018) Spoken language identification based on the enhanced self-adjusting extreme learning machine approach. PLoS ONE 13(4)
3. Albadr MAA, Tiun S, Ayob M, AL-Dhief FT (2019) Spoken language identification based on optimised genetic algorithm–extreme learning machine approach. Int J Speech Technol 22(3):711–727
4. Nammous MK, Saeed K (2019) Natural language processing: speaker, language, and gender identification with LSTM. In: Advanced computing and systems for security. Springer, Singapore, pp 143–156
5. Bhanja CC, Laskar MA, Laskar RH, Bandyopadhyay S (2019) Deep neural network based two-stage Indian language identification system using glottal closure instants as anchor points. J King Saud Univ-Comput Inf Sci
6. Deshwal D, Sangwan P, Kumar D (2020) A language identification system using hybrid features and back-propagation neural network. Appl Acoust 164:107289

7. Sulthana AR, Gupta M, Subramanian S, Mirza S (2020) Improvising the performance of image-based recommendation system using convolution neural networks and deep learning. Soft Comput 1–14
8. Sulthana AR, Ramasamy S (2019) Ontology and context based recommendation system using neuro-fuzzy classification. Comput Electr Eng 74:498–510
9. Sulthana R, Ramasamy S (2017) Context based classification of Reviews using association rule mining, fuzzy logics and ontology. Bull Electr Eng Inf 6(3):250–255