IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 4, 2020, accepted October 26, 2020, date of publication October 29, 2020, date of current version November 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034762

Optimizing Arabic Speech Distinctive Phonetic Features and Phoneme Recognition Using Genetic Algorithm

AHMED B. IBRAHIM[1], YASSER MOHAMMAD SEDDIQ[2], ALI HAMID MEFTAH[3], MANSOUR ALGHAMDI[4], SID-AHMED SELOUANI[5], (Senior Member, IEEE), MUSTAFA A. QAMHAN[3], YOUSEF A. ALOTAIBI[3], (Senior Member, IEEE), AND SALEH A. ALSHEBEILI[1,6]

[1]KACST-TIC in Radio Frequency and Photonics (RFTONICS), King Saud University, Riyadh 11421, Saudi Arabia
[2]King Abdulaziz City for Science and Technology (KACST), Riyadh 11442, Saudi Arabia
[3]College of Computer and Information Sciences, King Saud University, Riyadh 11421, Saudi Arabia
[4]Education and Training Evaluation Commission (ETEC), Riyadh 11537, Saudi Arabia
[5]LARIHS Laboratory, Université de Moncton, Campus de Shippagan, Shippagan, NB E8S 1P6, Canada
[6]Electrical Engineering Department, King Saud University, Riyadh 11421, Saudi Arabia

Corresponding author: Ahmed B. Ibrahim (ahahmed@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research, King Saud University, under Grant RG-1439-033.

ABSTRACT Distinctive phonetic features have an important role in Arabic speech phoneme recognition. In a given language, distinctive phonetic features are extrapolated from acoustic features using different methods. However, exploiting lengthy acoustic features vector in the sake of phoneme recognition has a huge cost in terms of computational complexity, which in turn, affects real time applications. The aim of this work is to consider methods to reduce the size of features vector employed for distinctive phonetic feature and phoneme recognition. The objective is to select the relevant input features that contribute to the speech recognition process. This, in turn, will lead to a reduced computational complexity of recognition algorithm, and an improved recognition accuracy. In the proposed approach, genetic algorithm is used to perform optimal features selection. Therefore, a baseline model based on feedforward neural networks is first built. This model is used to benchmark the results of proposed features selection method with a method that employs all elements of a features vector. Experimental results, utilizing the King Abdulaziz City for Science and Technology Arabic Phonetic Database, show that the average genetic algorithm based phoneme overall recognition accuracy is maintained slightly higher than that of recognition method employing the full-fledge features vector. The genetic algorithm based distinctive phonetic features recognition method has achieved a 50% reduction in the dimension of the input vector while obtaining a recognition accuracy of 90%. Moreover, the results of the proposed method is validated using Wilcoxon signed rank test.

INDEX TERMS Arabic speech distinctive phonetic feature, phoneme recognition, genetic algorithm.

I. INTRODUCTION

Performance of automatic speech recognition (ASR) systems is highly affected by input features that are extracted from the speech waveform. Features can be affected, in turn, by the speech variability that is caused by many factors such as speaker variability (e.g., speech tempo, speaker's age and gender, etc.), or by external sources such as background noise. Some features that are used in ASR systems are acoustic features such as spectrogram, mel-frequency cepstral coefficients (MFCCs), and short-time energy just to name a few. There are also other types of features that are highly representative, which are the distinctive phonetic features (DPFs). These features are introduced to a system as binary vectors where each bit of that vector describes the presence or absence (denoted as + or –, respectively) of some articulatory and acoustic properties that are associated with a particular phoneme utterance. DPFs are

The associate editor coordinating the review of this manuscript and approving it for publication was Fan-Hsun Tseng.

language dependent and each spoken language has its own finite set of DPFs, where a unique binary vector is assigned to each phoneme of the language [1].

Theoretically, the DPFs can describe all phonemes with uniquely distinctive binary patterns. DPF elements (bits) can be very useful also in categorizing phonemes based on similarity among them that can be directly traced by matching the DPF vectors of the different phonemes [2].

Here is an example, the phonemes /s/ and /z/ are very close to each other in the DPF space. Both phonemes share the same phonetic features (e.g., consonant, fricative, alveodental, etc.) except for the voicing feature, which refers to the physiological activity on the vocal folds, in which this pair of phonemes shows contrary values. That is, vocal folds must vibrate in order to vocalize /z/, otherwise a pure /s/ will be uttered [4]. The DPF elements in modern standard Arabic are listed in Table 1, depending on most references [1].

A. ARABIC LANGUAGE OVERVIEW

Modern standard Arabic (MSA) has 34 phonemes: three short vowels /a, i, u/, three long vowels /a:, i:, u:/, and 28 consonants that are grouped under a number of subcategories such as plosives, affricatives, nasals, trills, etc. There are two subcategories of Arabic phonemes that are not found in many languages, which are the pharyngeal and the emphatic phonemes [4]. The duration of phonemes in Arabic is phonemic. That is, phonemes (vowels and consonants) can be uttered in short or long periods, where both ways directly affect the word meaning [5].

Words in Arabic consist of syllables, where each syllable must have at least one vowel. Therefore, a word would have as many syllables as there are vowels in that word [6].

B. LITERATURE REVIEW

Extracting DPFs has been tackled in many published studies. In [7], DPFs are extracted using multilayer perceptron (MLP), and have been demonstrated to enhance the robustness of ASR. In [8], a canonicalization process was proposed composing of multiple DPF extractors in order to neutralize the effect of speaker's gender on ASR system robustness. Similarly, in [9] multiple DPF extractors were deployed to eliminate the effect of hidden factors and to reduce the effect of noise. In extension to that, in [10], the DPF extractors are utilized to neutralize hidden factors of speakers' variability in addition to gender and to eliminate the effect of noise. In [11], a DPF extractor was proposed to enhance the accuracy of speech segmentation, using recurrent neural network (RNN) followed by an MLP neural network. In [12], a phoneme recognition system was proposed consisting of two-stage DPF extraction: the first stage converts acoustic features to a 45-bit DPF vector, while the second stage makes the vectors orthogonal before being fed to a hidden Markov model (HMM) classifier. The work in [13] proposed the use of recurrent neural networks to detect phonological features in continuous speech.

Articulatory Features (AFs) are utilized in [14] to develop pronunciation models for ASR systems. In [15], the articulatory features are investigated with respect to monolingual, cross-lingual, and multilingual ASR. The work published in [16] is an attempt to develop a large vocabulary ASR system utilizing the distinctive phonetic features instead of the ordinary short-term spectra features. DPF-based phone-level segmentation is reported in [17], where the system is built using recurrent neural networks and a multi-layer neural networks. In [3], a representation method is proposed such that a speech waveform is represented by some abstract linguistic descriptors from which a set of discriminative features is derived and fed to ASR systems. The work in [18] attempted to improve the ASR performance by adopting a multi-stream technique of DPFs and spectral features. A noise-robust ASR system that applies logarithmic normal distributions of HMMs for the purpose of approximating DPF elements was proposed in [19]. The robustness of ASR under Low-SNR of car environments was investigated in [20], where DPFs along with spectral cues are utilized to enhance system robustness. Phoneme classification for Bengali Language using DPFs and deep neural network was reported in [21]. In [22], a deep neural network is used to predict historical phonetic features drawn upon synchronic phonetic patterns arising from coarticulation and statistical constraints in Proto-Indo-European language. In [23], extracted acoustic features of speech signal using hamming window and pre-emphasis filter, in addition to extracted decompositional features using daubechies-filtered 5th-depth Wavelet Packet Decomposition (WPT), are optimized using genetic algorithm to classify Turkish vowels.

The relevance of evolutionary-based algorithms, like a genetic algorithm, that belong to a family of search algorithms inspired by the process of evolution in nature, was demonstrated in a recent study showing that optimizing the topology of an Artificial Neural Network may lead to a high classification rate of spoken utterances of both native and non-native English speakers [24].

C. DISTINCTIVE PHONETIC FEATURES IN THE ARABIC LANGUAGE

The Arabic language has a number of unique characteristics, such as the presence of a relatively large number of pharyngeal and emphatic sounds, in addition to various types of allophones, many of which are the result of emphaticness and gemination. Arabic also has several lexical stress systems, likely unknown in other languages, but regrettably unstudied and in need of thorough investigation. In the context of the present investigation of DPFs in Arabic, only a limited number of previous studies dedicated to the subject are available. In [25], Arabic DPFs were extracted using modular connectionist architectures with rule-based systems (SARPH). In [26], Selouani *et al.* deployed neural networks of mixed architectures fed with continuous speech in order to recognize complex Arabic phonemes.

TABLE 1. DPF values of Arabic phonemes.

	Arabic writing	KACST Symbol	IPA symbol	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
				affricative	alveodental	alveopalatal	anterior	aspirated	bilabial	consonant	continuant	coronal	emphatic	fricative	glottal	high	interdental	labiodental	labiovelar	lateral	nasal	palatal	pharyngeal	plosive	rounded	semivowel	short	trill	unvoiced	uvular	velar	voiced	vowel
1	ء	hz10	ʔ	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	ب	bs10	b	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-
3	ت	ts10	t	-	+	-	+	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-
4	ث	vs10	θ	-	-	-	+	-	-	+	+	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
5	ج	jb10	ʤ	+	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
6	ح	hb10	ħ	-	-	-	-	-	-	+	+	-	-	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	-	-	-	-
7	خ	xs10	X	-	-	-	-	-	-	+	+	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-
8	د	ds10	d	-	+	-	+	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-
9	ذ	vb10	ð	-	-	-	+	-	-	+	+	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
10	ر	rs10	r	-	+	-	+	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-
11	ز	zs10	z	-	+	-	+	-	-	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
12	س	ss10	s	-	+	-	+	-	-	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
13	ش	js10	ʃ	-	-	+	-	-	-	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
14	ص	sb10	$\mathrm{s}^{ʕ}$	-	+	-	+	-	-	+	+	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
15	ض	db10	$\mathrm{d}^{ʕ}$	-	+	-	+	-	-	+	+	+	+	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-
16	ط	tb10	$\mathrm{t}^{ʕ}$	-	+	-	+	-	-	+	-	+	+	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-
17	ظ	zb10	$ð^{ʕ}$	-	-	-	+	-	-	+	+	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
18	ع	cs10	ʕ	-	-	-	-	-	-	+	+	-	-	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-
19	غ	gs10	ʁ	-	-	-	-	-	-	+	+	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-
20	ف	fs10	f	-	-	-	+	-	-	+	+	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
21	ق	qs10	q	-	-	-	-	-	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	+	+	-	-	-
22	ك	ks10	k	-	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	+	-	-
23	ل	ls10	l	-	+	-	+	-	-	+	+	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-
24	م	ms10	m	-	-	-	+	-	+	+	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-
25	ن	ns10	n	-	+	-	+	-	-	+	+	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-
26	هـ	hs10	h	-	-	-	-	-	-	+	+	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
27	و	ws10	w	-	-	-	+	-	+	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	+	+	-	-	-	-	-	+	-
28	ي	ys10	j	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	-	+	-
29	ـَ	as10	a	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	+
30	ـَا	as21	aː	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+
31	ـِ	is10	i	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	+
32	ـِي	is21	iː	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+
33	ـُ	us10	u	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	+	+
34	ـُو	us21	uː	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	+

In our previous works, the multidimensional phonological feature structure of Arabic was investigated by assessing the performance of statistical and connectionist approaches in performing the complex mappings between DPFs and associated acoustic cues [27]. In a review paper [28], a background on Arabic DPFs, highlighting the historical and geographical varieties, the problem of ambiguous definitions between classical and modern phonology, and the deviations in phonemes and DPF elements across dialects of Arabic were investigated and presented. HMM was used with an original normalization technique to perform Arabic phoneme classification using the DPF elements and utilized DPFs for the purpose of introducing a canonical process for phoneme level classification by means of substituting the speech waveform with its phonetic binary DPF vector [29]. In another work [30], the problem of DPF modeling and extraction of modern standard Arabic is tackled by using deep neural networks (DNNs) and compared with the classical MLP models. The representativeness of several acoustic cues for different DPF elements was measured additional to the proper evaluation measures satisfying the imbalanced nature of the DPF elements which was addressed. It is important to note that our previous work on DPF modeling using DNN had an objective acoustic-to-phonetic conversion, where Arabic DPFs were extracted from acoustic features using DNNs. However, input feature selection was not within the scope of that previous work. On the other hand, the present work has a different objective and scope, which is to come up with a unified reduced set of acoustic features that can be used to extract any DPF element using any machine learning technique.

D. MOTIVATION AND OBJECTIVES

The aim of this work is to consider methods to reduce the size of features' vector employed for phoneme recognition by using a genetic algorithm-based approach. The objective is to select the relevant input features that yield to reduce the computational complexity of the recognition algorithm while improving the DPF recognition accuracy. Genetic algorithms (GAs) have been successfully integrated into various speech-processing applications such as speaker adaptation of acoustic models or speech enhancement [31]. GAs have also shown advantage in enhancing the performance of voice communication systems [32]. The main advantage of using GA to optimize the feature selection is their ability to extend the search space of best parameters by applying the principle of maintaining and manipulating a large population of solutions. Their methodology consists of implementing a 'survival of the fittest' strategy in their search for better solutions. Thus, the original idea of this article is to use the ability of GAs to select the relevant acoustic features from speech. GAs and neural networks are very common and effective in processing digital speech mainly in recognition and classification problems. Reducing speech acoustic features while keeping the nominal system accuracy is a very important goal that will help to reduce central processing unit (CPU) time and memory requirements. Hence, the main contribution of this work is to build a robust features selection model, whose input is a wide range of multiple acoustic features. This proposed model is in the form of a hybrid of genetic algorithm and neural network, to predict distinctive phonetic features of phonemes in modern standard Arabic.

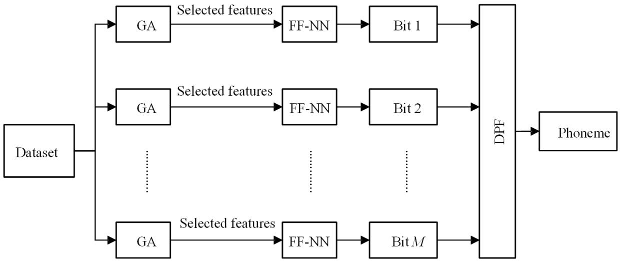


FIGURE 1. Phonemes classification using DPF elements and GA for adaptive features selection.

E. PAPER'S ORGANIZATION

After Section I, Introduction, the remaining of this article is organized as follows. Section 2 presents an introductory background about the genetic algorithm and gives an overview of the proposed GA-based DPF recognition method. Section 3 provides information about the dataset and features used in this study. Also, in this section, the extracted features are examined and preprocessed for the purpose of normalization and reducing the outliers. Section 4 presents experimental results pertaining to the development of a baseline model based on Feedforward Neural Networks (FF-NN). This model will be used to evaluate the phoneme recognition accuracy of the GA-based features selection method against the whole feature vector-based method. Details of the GA-based features selection method is given in this section. Section 5 presents and analyzes the performance of phoneme recognition based on DPF elements, while the discussion is presented in Section 6. Section 7 gives the concluding remarks.

II. OVERVIEW ON THE GENETIC ALGORITHM BASED METHOD

This sections presents the architecture of proposed GA-based DPF recognition method. Also, it gives brief introduction about the basics of GA.

A. SYSTEM MODEL

Figure 1 shows the proposed architecture for phonemes classification using DPF elements and GA for adaptive features selection. In this model, the dataset consists of N preprocessed features. The N-point features vector is applied to M GAs followed by M FF-NNs working in parallel, where M is the number of DPF elements. The output of each branch is one bit with value either '0' or '1', depending on the DPF element it represents.

The selected features by each GA and the parameters of each FF-NN are determined through a training process. In the testing phase, a phoneme is identified by measuring the Euclidean distance between the outputs of M FF-NNs, a vector of M bits, and the actual DPF vectors of all phonemes.

B. GENETIC ALGORITHM

GAs belong to a family of computational models, namely evolutionary algorithms, inspired by the process of natural evolution [33]. They have received increasing popularity due to their robustness and efficiency in solving complex problems in which many classical mathematical methods fail [34]. In particular, GAs work by executing five main steps [35].

- *Coding*. The parameters of a given problem are encoded often using a binary string.
- *Initiation of population*. A set of randomly generated strings (chromosomes) are generated as candidate solutions.
- *Evaluation of responses*. A goodness of fit is applied to each string (chromosome) to determine its chance to be selected for creating the next generation.
- *Reproduction*. It involves two steps: 1) selecting a set of strings from the previous population, 2) generating a new population through combining parts of selected strings (cross over operation).
- *Mutation*. It maintains genetic diversity from one generation of a population to the next. It alters one or more elements (genes) in a string (chromosome) from its initial state. In binary encoding, a gene of value '1' gets changed to '0' and vice versa.

The proposed genetic algorithm based feature selection method using feedforward neural network is a metaheuristic method for dimensionality reduction. It has a potential application in automatic speech recognition and its applications. For example, in [24], the authors used evolutionary algorithms based optimized deep neural network for recognition of diphthong vowel sounds in the English phonetic alphabet. In [36], the author tied genetic algorithm with Manhattan distance to classify plain and emphatic vowels in continuous Arabic speech. Also in [37], the genetic algorithm was exploited in the segmentation of Arabic speech, and in [38], it has been used with the K-nearest neighbour algorithm to build a voice command recognition system. Using the proposed GA-based optimization method allows to maintain the selected indices of features after finding them during the training process. The training process as to be applied once through the corpus. Therefore, the reduced size of features' vector contributes in reducing time cost when performing the real time applications. In addition, using genetic algorithm can help in removing redundancy in the dataset under investigation [39].

III. DATASET AND FEATURES EXTRACTION

This section is to address a fundamental step in this work, pertaining to the preparation of data for the subsequent experiments. It is of great importance to provide the system with suitable data that carries rich phonetic information. Also, feature extraction is an essential preprocessing step, where representative acoustic features are extracted and prepared for training the acoustic-to-phonetic conversion models.

TABLE 2. Information of the dataset used in this work.

Number of subjects	7 males
Dataset size	13,766 phonemes
Training subset	9,636 phonemes (70%)
Test subset	4,130 phonemes (30%)

A. KAPD DATASET

The dataset used in this work is extracted from the KACST Arabic Phonetic Database (KAPD) [40], [41] as summarized in Table 2. KAPD is a phonetically rich speech corpus recorded by seven native Saudi male subjects. Each Arabic phoneme appears in a carrier word in one of three different positions: initial, middle, and final position. Also, for each position of these, three different carrier words exist, where the target phoneme co-articulates with one of the three short vowels (i.e., /a/, /i/, and /u/) of Arabic in each word. For a consonant phoneme in a middle position, the carrier word contains the phoneme in one of two states: single and geminated. Each one of the aforementioned combinations is uttered by each one of the seven subjects in eight different experiment each one of them aims at capturing some physical characteristics of the speech signal, in addition to recording the uttered word. KAPD is composed of the following subsets: Subset A for aerodynamic data, Subset C for lip-labeled face images, Subset E for epiglottal imaging data, Subset G for electroglottographic measurement data, Subset N for nasal and oral air pressure measurements data, Subset P for electropalatal imaging data, Subset V for vocal folds imaging data, and Subset X for velopharyngeal imaging data. The experiments carried out in this work are based on random samples taken from KAPD dataset. That is, 13,766 phonemes are used and split into two subsets: training subset consisting of 9,636 phonemes (70%), and test subset consisting of 4,130 phonemes (30%).

B. FEATURES EXTRACTION

The input acoustic features are extracted from each phoneme waveform. That is, a number of 15, evenly spaced, 20-ms long frames are sampled from each waveform. Spacing between frames vary from one waveform to another since phonemes are not all equal in the time duration. Each frame is windowed by a 20-ms Hamming window. Other preprocessing of DC removal and pre-emphasis using $\alpha = 0.97$ are also applied. The following acoustic features are extracted from each frame:

1) 256-point Spectrogram calculated using short-time Fourier Transform. The number of spectrogram points in an input vector is: 256 points per frame × 15 frames = 3840 points.
2) 39-coefficient MFCC, where log energy, first derivative and second derivative are computed for each frame. This results in a total of 585 coefficients per phoneme, which has 15 frames.

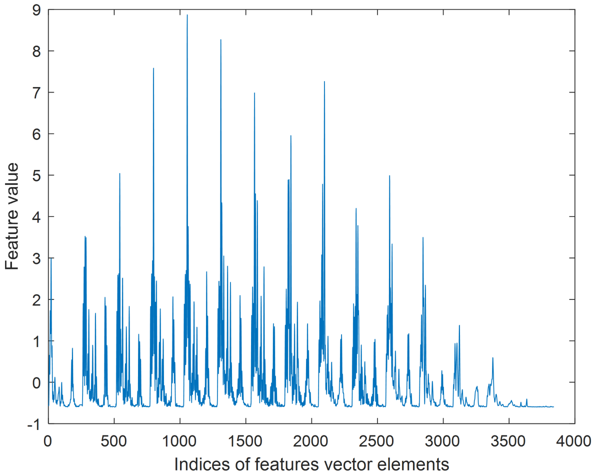


FIGURE 2. A sample of the normalized spectrogram features of phoneme 'as21'.

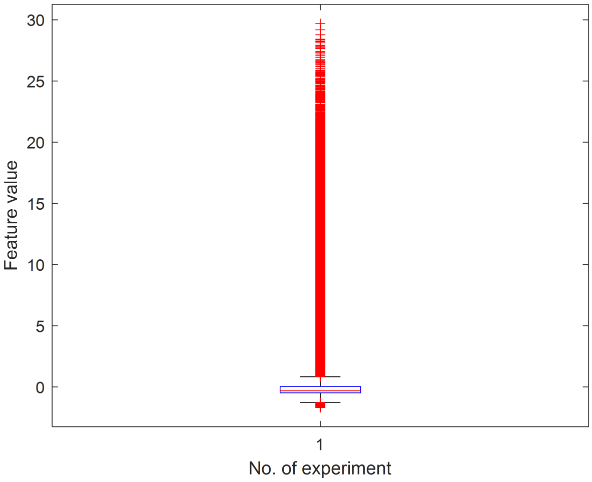


FIGURE 3. Boxplot of normalized spectrogram features.

3) Zero-crossing Rate (ZCR), where each frame yields one scalar value that brings to a total of 15 ZCR values per input vector.
4) Short-time Energy, where there is also one scalar value per frame summing up to 15 values per input vector.
5) Voicing Percentage, which is one value representing the percentage of frames (among the 15 frames) that carry valid (nonzero) pitch values. There is only one percentage value in each input vector.

Therefore, the total length of original features vector is 4,456 (= 3,940+585+15+15+1) points. In this study, only the first 15 points are considered from MFCC coefficients for each frame instead of considering all 39 coefficients. This is because the remaining 24 MFCC coefficients have been found significantly of small values. Selecting a large number of MFCC coefficients results in more complexity in the model [42]. Based on this modification, the number of MFCC coefficients for all 15 frames is now 15 × 15 = 225 points instead of 585 points, which brings the features vector used in our experiments to 4,096 points.

C. FEATURES NORMALIZATION

A features'vector consists of the aforementioned five different types of features, each of which has its own dynamic range. Therefore, it is essential that features'vectors are normalized before being applied to a machine learning algorithm. In our development, each type of features of a given features vector is normalized so that it has unity variance. Figure 2 shows a sample of the normalized spectrogram features of a phoneme represented by 15 records.

Note that the normalized spectrogram has spikes, corresponding to resonances in the vocal tack. The amplitudes of the spikes vary between the records of a phoneme, and also vary between those of other different phonemes. Figure 3 shows the boxplot of normalized spectrogram features of the dataset under consideration. A boxplot is a standardized way to display data distribution by using five statistical measures, which are the minimum, first quartile (Q1), median, third quartile (Q3), and maximum of dataset [43]. It tells about the skewness of data distribution and presence of outliers.

In this context, the minimum and maximum of a dataset are defined as Q1−1.5xIQR and Q3+1.5xIQR, respectively. Here, IQR is the difference between Q3 and Q1. Therefore, any sample of value less than the minimum or greater than the maximum is considered an outlier. For the normalized spectrogram features, the minimum value is -1.2661, the first quartile (Q1) is -0.47869, the third quartile (Q3) is 0.046241, and the maximum value is 0.83364. However, there is a large number of outliers due to the presence of spikes. In fact, the presence of these spikes causes a large dynamic range for spectrogram features. Therefore, it is important to limit spikes'amplitudes so that they all have relatively comparable values.

Let $s(n)$ be the n^{th} sample of the features shown in Figure 2. The new scaled feature sample, $s'(n)$, is then given by

$$s'(n) = s(n)/(1 + \beta|s(n)|) \tag{1}$$

The parameter 'β' is a scalar whose value is greater than or equal to 0. This parameter controls the amount of scaling. If $\beta = 0$, then no scaling is performed; that is, $s'(n) = s(n)$. However, for extremely large values of β, the value of $s'(n)$ becomes zero. In what follows, the value of β is set to 1. With this value, samples of large amplitudes will undergo high attenuation, while those of relatively low amplitudes will pass almost unchanged. Figure 4 shows the resulting normalized and scaled spectrogram features.

Figure 5 shows the normalized MFCC features of the same phoneme considered above. The boxplot of whole MFCC features of normalized KAPD dataset is shown in Figure 6. It is clear from the figure that these features have large dynamic range. Thus, the scaling operation is applied to these features in a similar manner to what is described in (1). Figure 7 shows the normalized and scaled MFCC features. Figure 8, on the other hand, shows the complete features vector, including the

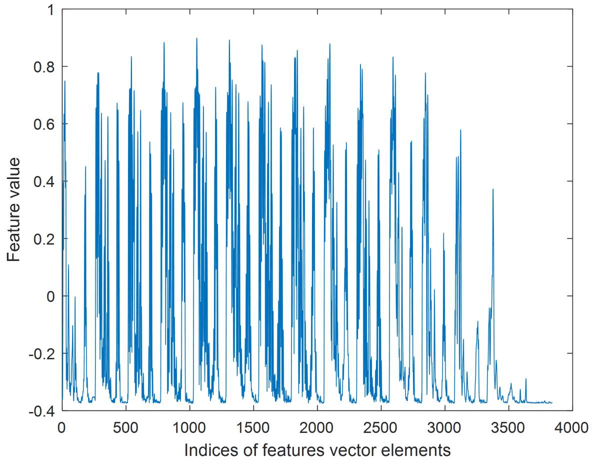


FIGURE 4. The resulting normalized and scaled spectrogram features of phoneme 'as21'.

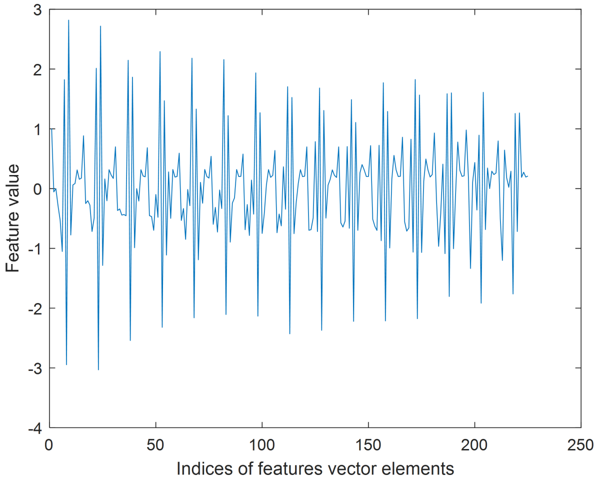


FIGURE 5. The normalized MFCC features of phoneme 'as21'.

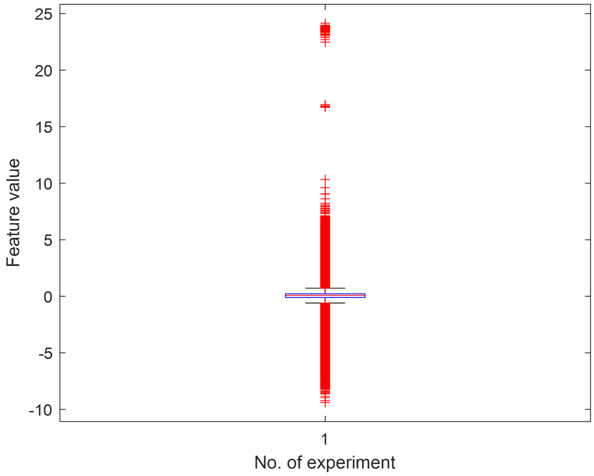


FIGURE 6. The boxplot of whole normalized MFCC features.

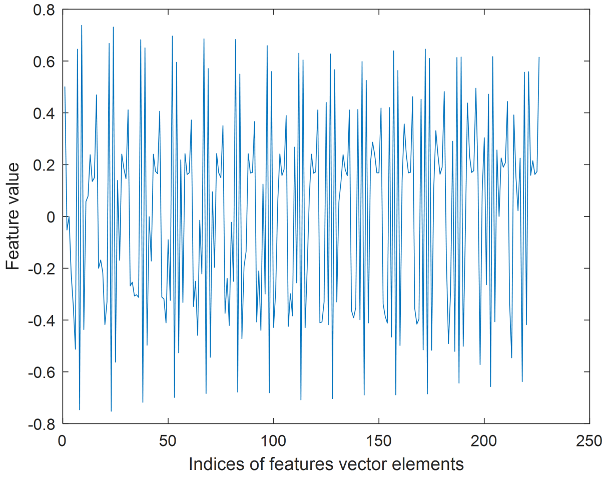


FIGURE 7. The normalized and scaled MFCC features.

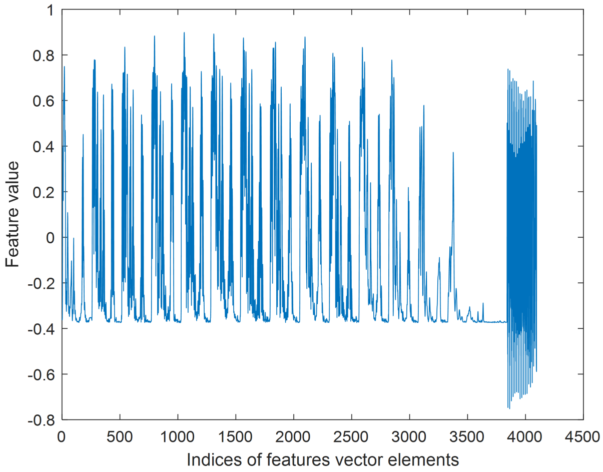


FIGURE 8. The complete features vector.

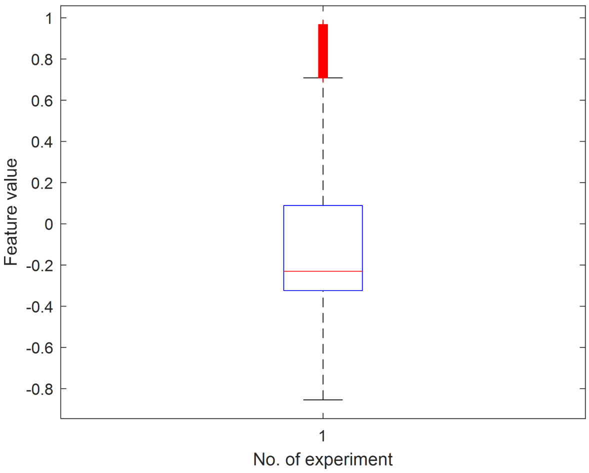


FIGURE 9. The boxplot of all features vectors after pre-processing.

pitch percentage and the normalized energy and zero-crossing features. The boxplot of all features vectors after being pre-processed is shown in Figure 9.

IV. FEATURE SELECTION APPROACH

This section considers the selection of appropriate configuration for the GA-based features section process. Therefore, a model suitable for phoneme classification is first introduced. This model is needed to act as a base line against which the performance of the GA-based features selection method is compared.

A. DEVELOPMENT OF A BASELINE MODEL

With the normalized and scaled dataset described in Subsection III-C, machine learning algorithms can be used to classify different phonemes. However, the performance of such algorithms is greatly affected by many factors, including specific parameters related to the input data; e.g., the length of features vector and the correlation among its entries, and the signal-to-noise ratio. For the normalized and scaled KAPD (NS-KAPD) dataset, the length of each features vector is 4096, as described in Subsection III-B With this high dimensional vector, its entries may not be all equally important, as it may contain redundant and/or irrelevant features and/or noise. Therefore, it is essential that the size of the input features vector be reduced to contain only the features which contribute to the classification process. By doing so, the original representation of data will not be affected, and may even provide better readability and interpretability. Furthermore, the computational complexity will be reduced, and the classification accuracy could be improved.

In this subsection, the problem of selecting a small subset of entries of a features vector is addressed by applying GA. Features selection based on GA has been widely studied and a large number of methods have been developed in different applications; in [44], the authors used genetic algorithm to design decoder-tailored polar code, where in [45], the problem of finding optimal distance for a traveling salesman is solved using genetic algorithm. In [46], the effect of using different configurations of GA is investigated. Using the available NS-KAPD dataset, a model suitable for phoneme classification is considered here. This model will be used as a base line against which the performance of the GA-based features selection method is compared. The proposed model is a simple Feedforward Neural Network (FF-NN), consisting of an input layer, two hidden layers each of which has 100 neurons, and an output layer in the form of a binary vector of size 34. In the ideal case, one element of the output binary vector is ‘1’ and the remaining elements are zeros. The active element corresponds to one of the 34 different phonemes. Figure 10 shows the architecture of proposed FF-NN model, where **W** is the weights vector and **b** is the bias.

The FF-NN model is trained using 70% randomly selected features vectors of the NS-KAPD dataset. The remaining 30% features vectors are used for testing. Table 3 presents the performance in terms of four measures: the average classification accuracy, Area Under Curve (AUC), G-mean, and F-score. These numbers are our baseline to evaluate the performance of GA to select a subset out of the 4096 features of an input vector. In other words, the performance given in the table is the yield of the system when our full features vector is used without the involvement of GA selection scheme.

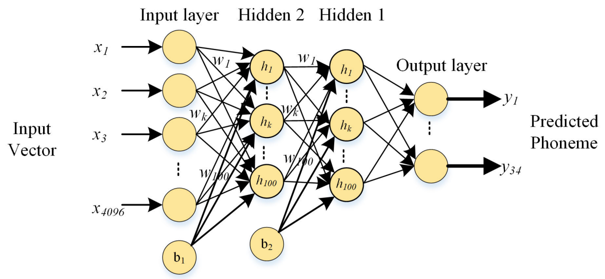


FIGURE 10. The architecture of proposed FF-NN model.

TABLE 3. The accuracy, AUC, G-mean, and F-score by using all 4096 baseline features.

Accuracy	AUC	G-mean	F-score
68%	0.84	0.81	0.677

B. GENETIC ALGORITHM BASED FEATURES SELECTION METHOD

In our development, each features vector is encoded by 61 bits divided in order as follows:

a) 15 bits encode the 15 spectrogram records. That is, each bit encodes one spectrum record. If the bit value is ‘1’, this means the corresponding record will be included in the new features vector, otherwise it will not be included.

b) 15 bits encode the 15 MFCC records. That is, each bit encodes one MFCC record. If the bit value is ‘1’, this means the corresponding record will be included in the new features vector, otherwise it will not be included.

c) 15 bits encode the 15 zero-crossing values.

d) 15 bits encode the 15 energy values.

e) 1 bit encodes the value of pitch percentage.

Figure 11 shows the schematic diagram of the encoding process. For each possible binary string of length 61 bits, the corresponding features are selected and used to train and test the proposed FF-NN, as described in Figure 12. Note that the GA needs to search for the binary string which gives the maximum possible classification accuracy. Each spectrum or MFCC record is encoded by one bit to reduce the search space of GA. Following this encoding scheme, the search space becomes 2^{61} candidate features vectors. If, however, entries of spectrogram and MFCC records are not encoded, then this leads to a search space of size 2^{4096}.

For the remaining genetic operations, there are many possibilities each of which may be effective for one type of application but worse for another. In fact, there is no one choice fitting all, and the majority of research efforts are focused on finding an optimum choice for a specific setting. In what follows, the GA performance is investigated using the commonly used configurations in literature, as follows. For parent selection operator, Roulette wheel selection, tournament selection [47], and their hybrid combination [48] are considered. For crossover operator, the single point crossover, double point crossover, and uniform crossover [49], [50]

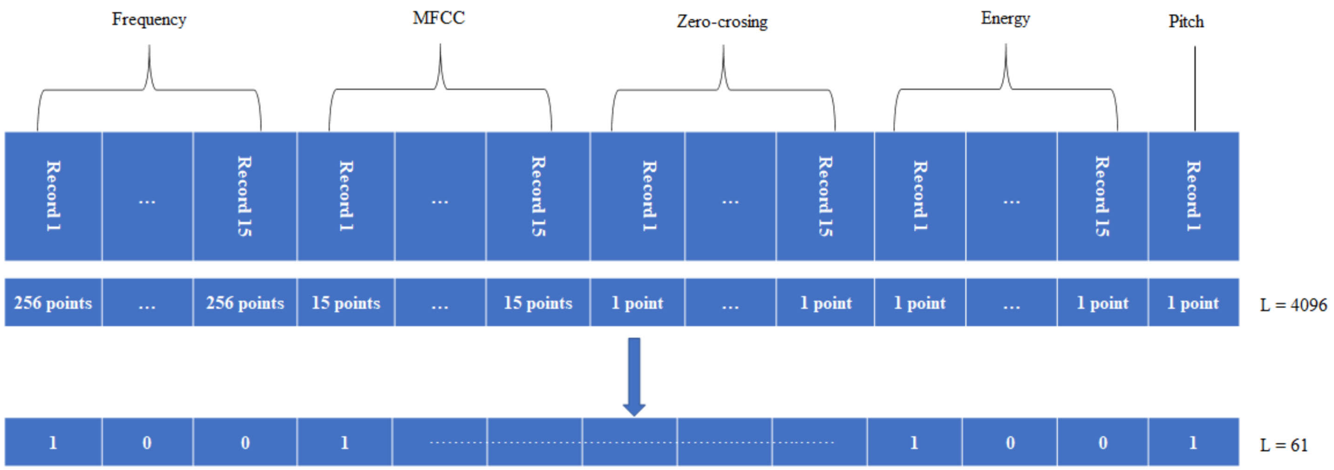


FIGURE 11. The schematic diagram of the encoding process.

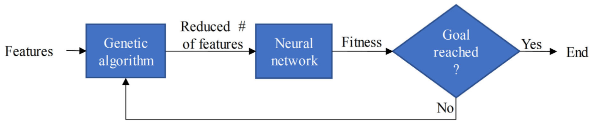


FIGURE 12. The schematic diagram of GA with neural network.

TABLE 4. The performance of different GA configurations.

Selection method	Crossover operator	Accuracy
Hybrid selection	Uniform crossover	68.2 %
Roulette wheel selection		67.77 %
Tournament selection		66.87 %
Hybrid selection	Single point	67.2 %
	Double point	67.6 %

are considered. The crossover operator is followed by bit flip mutation. The details of each operator are well explained in its relevant reference.

Table 4 shows the performance of the five configurations, in terms of the classification accuracy, when the GA is applied along with the FF-NN to the NS-KAPD dataset. It is evident from the table that the GA with the selection operator combing Roulette wheel and tournament schemes, and uniform point crossover is the best performing algorithm. Therefore, this GA configuration is selected for our analysis to follow.

Table 5 gives further details about the performance of the best performing GA in terms of AUC, G-mean, and F-measure. Compared with the performance of FF-NN alone, it is observed that the GA gives almost similar results but with a reduced size features vector. By comparing the performances using all four measures, it is noticed that there is almost a full match with the corresponding figures given in Table 2. By this, it can be concluded that the same performance is kept by using almost 50% of the features vector length generated by the GA.

TABLE 5. The accuracy, AUC, G-mean, and F-score by using the reduced feature vector by GA algorithm.

Accuracy	AUC	G-mean	F-score
68.2 %	0.842	0.81	0.674

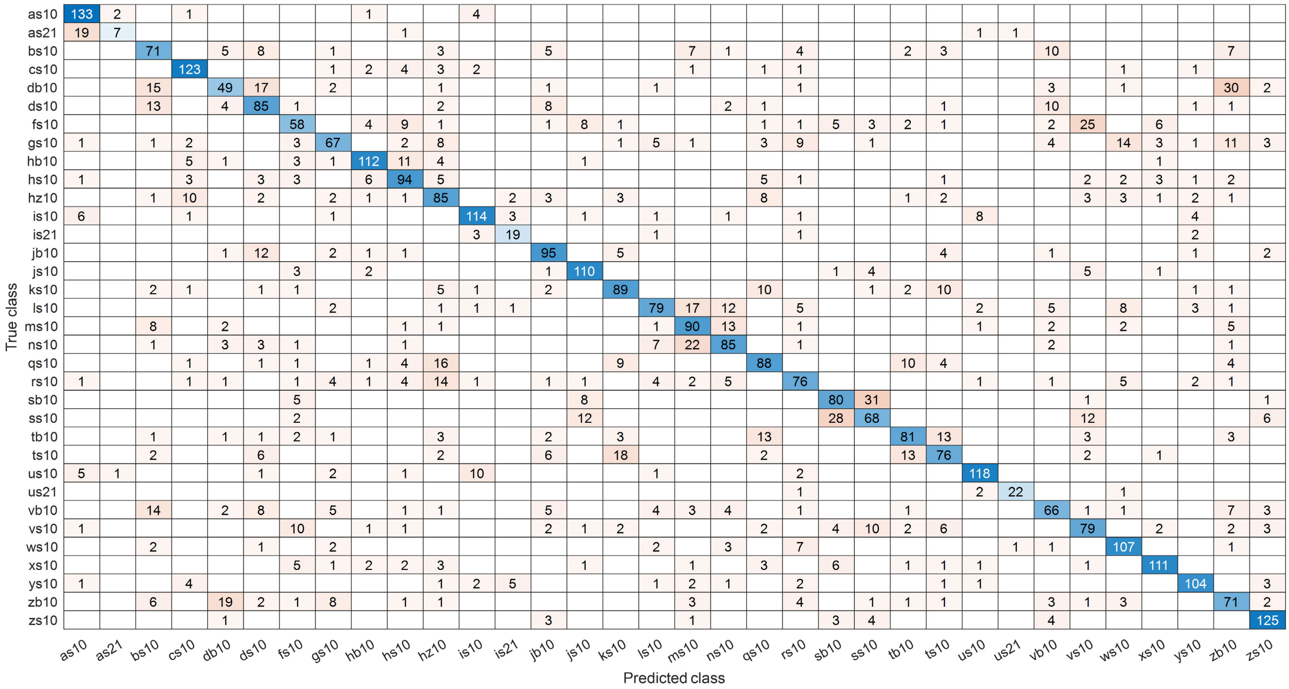
In particular, the GA shows that only a features vector of size 3231 is sufficient to achieve the performance of the full-fledge features vector. The confusion matrix is depicted in Figure 13. This matrix shows that the phonemes ‘sb10’, ‘db10’, and ‘fs10’ are greatly confused with the phonemes ‘ss10’, ‘zb10’, and ‘vs10’. This is intuitively not surprising because the features vectors of these phonemes may not be well separable.

Figures 14 (a) and (b) show results when the t-distribution stochastic neighbor embedding (t-SNE) algorithm [51] is applied to the corresponding features of phonemes ‘sb10’ and ‘ss10’, and the two mostly separable phonemes (‘hz10’ and ‘ss10’). The t-SNE algorithm is used to reduce the data dimensionality from 4096 to 2, while preserving both local and global structure of data, hence it facilitates its visual inspection.

From the figures, it is observed that features of phonemes ‘sb10’ and ‘ss10’ overlap, which makes phonemes’ discrimination difficult. This overlap led to 59 times confusions between these two phonemes. Therefore, it can be concluded that there is a big similarity between the two phonemes ‘sb10’ and ‘ss10’, but, on the other hand there is a big dissimilarly between ‘hz10’ and ‘ss10’ phonemes.

V. PHONEME RECOGNITION PERFORMANCE USING DISTINCTIVE PHONETIC FEATURES ELEMENTS

In this study, each phoneme is represented by 30 DPF elements that are listed in Table 1, which can be used for phoneme recognition. Note that the NS-KAPD dataset has feature vectors of dimension 4096. It is possible that the 4096 features may not all contribute to the recognition of a



	as10	as21	bs10	cs10	db10	ds10	fs10	gs10	hb10	hs10	hz10	is10	is21	jb10	js10	ks10	ls10	ms10	ns10	qs10	rs10	sb10	ss10	tb10	ts10	us10	us21	vb10	vs10	ws10	xs10	ys10	zb10	zs10
as10	133	2		1					1			4																						
as21	19	7								1																1	1							
bs10			71		5	8		1			3			5				7	1		4			2	3			10					7	
cs10				123				1	2	4	3	2						1		1	1									1		1		
db10			15		49	17		2			1			1			1				1							3		1			30	2
ds10			13		4	85	1				2			8					2	1					1			10				1	1	
fs10							58		4	9	1			1	8	1				1	1	5	3	2	1			2	25		6			
gs10	1		1	2			3	67		2	8					1	5	1		3	9		1					4		14	3	1	11	3
hb10				5	1		3	1	112	11	4				1																1			
hs10	1			3		3	3		6	94	5									5	1				1				2	2	3	1	2	
hz10			1	10		2		2	1	1	85		2	3		3				8				1	2				3	3	1	2	1	
is10	6			1				1				114	3		1		1		1		1					8						4		
is21												3	19				1				1											2		
jb10					1	12		2	1	1				95		5									4			1				1		2
js10							3		2					1	110							1	4						5		1			
ks10			2	1		1	1				5	1		2		89				10			1	2	10							1	1	
ls10								2			1	1	1				79	17	12		5					2		5		8		3	1	
ms10			8		2					1	1						1	90	13		1					1		2		2			5	
ns10			1		3	3	1			1							7	22	85		1							2					1	
qs10				1		1	1		1	4	16					9				88				10	4								4	
rs10	1			1	1		1	4	1	4	14	1		1	1		4	2	5		76					1		1		5		2	1	
sb10							5								8							80	31						1					1
ss10							2								12							28	68						12					6
tb10			1		1	1	2	1			3			2		3				13				81	13				3				3	
ts10			2			6					2			6		18				2				13	76				2		1			
us10	5	1				1		2		1		10					1				2					118								
us21																					1					2	22			1				
vb10			14		2	8		5		1	1			5			4	3	4		1			1				66	1	1			7	3
vs10	1						10		1	1				2	1	2				2		4	10	2	6				79		2		2	3
ws10			2			1		2									2		3		7						1	1		107			1	
xs10							5	1	2	2	3				1			1		3		6		1	1	1			1		111			
ys10	1			4							1	2	5				1	2	1		2				1	1						104		3
zb10			6		19	2	1	8		1	1							3			4		1	1	1			3	1	3			71	2
zs10					1									3				1				3	4					4						125

FIGURE 13. The confusion matrix of the 34 phonemes.

DPF element. In this section, GA is used to determine the features well represent a particular DPF element, and build for each element an FF-NN for its classification. Table 6 shows the performance of the developed 30 FF-NNs, along with the number of selected features for each DPF element. The table also displays the number of ‘1s’ in the final output code vector of GA.

By virtue of Table 6, it is of interest to note that the average accuracy across all over the 30 DPF elements is 90%, while the average AUC, GM, and F_Score are 0.85, 0.84, 0.78, respectively. This excellent performance has been achieved with a great reduction in the required number of features, which ranges from 2982 down to 1131 with an average 50% (=2047/4096) of total number of features.

Table 7 gives more details about the selected features for each DPF element, and provides the final 61-binary string output of GA. For example, the first DPF element has a 61-bit string vector with 31 entries of ‘1s’. This corresponds to the selection of spectrogram features computed from 8 frames, MFCC features computed from 10 frames, zero-crossing features computed from 6 frames, and energy features computed from 7 frames. The pitch percentage for this PDF element, however, is not selected.

Figure 15 depicts the number of times each entry of the 61-bit string vectors carries the value of ‘1’. It is evident from the figure that entries number 42 and 57 have the lowest frequency of having the value of ‘1’, while entry number 15 has the highest. These three entries represent the corresponding frames of zero-crossing percentage, energy, and the spectrogram features, respectively.

As described in Subsection II-A, Figure 1 shows the targeted architecture for phonemes classification using DPF elements and GA for adaptive features selection. In this figure, $N = 4096$ and $M = 30$. Therefore, the 4096-features vector is applied to 30 GAs followed by 30 FF-NNs working in parallel. The output of these FF-NNs constitutes a binary vector of length 30 of ‘0s’ or ‘1s’, depending on the DPF elements of phoneme under consideration.

Figure 16 shows the performance, in terms of the confusion matrix of proposed classification system, where the features that are nominated for training and testing are those that are selected by the GA. In this setting, the outputs of 30 FF-NNs constitute the predicted DPF vector corresponding to a particular phoneme. Therefore, a phoneme is identified by measuring the Euclidean distance between the output of the model and the actual DPF vectors of all phonemes. The phoneme whose DPF vector has the minimum distance is selected. The test set is composed of 100 samples of each phoneme except the phonemes (‘as21’, ‘is21’, ‘us21’) that have representation of 28, 26, and 26, respectively, in the KAPD dataset. Note that the two confusion matrices in Figure 13 and Figure 16 represent results of two methods of phoneme recognition using the output of GA-FFN model. The first method computes the confusion matrix right after the neural network, while in the second method the confusion matrix is computed from the predicted DPF elements.

Figure 17 shows the identification accuracy for each phoneme computed from Figure 16, where the two phonemes ‘bs10’ and ‘fs10’ are that of the worst performance as each phoneme gets confused with other phonemes. The two

TABLE 6. The performance of the developed 30 FF-NNs, along with the number of selected features for each DPF element.

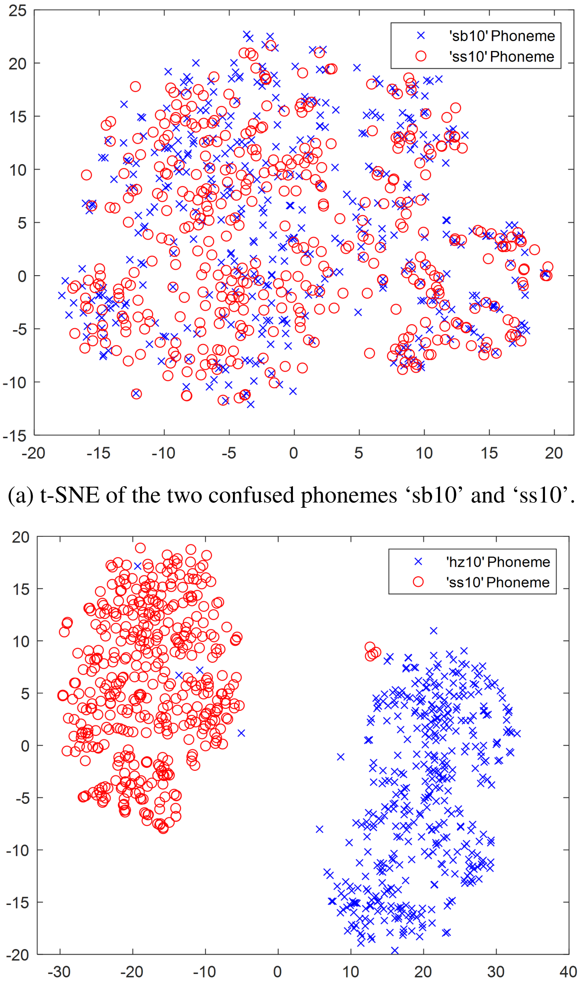
No.	DPF	Accuracy (%)	AUC	GM	F_Score	No. of '1s' in the final 61-bit output vector of GA	Length of Features Vector out of 4096 entries
1	affricative	98.5	0.853	0.84	0.75	31	2211
2	alveodental	84.9	0.814	0.81	0.749	28	2648
3	alveopalatal	97.4	0.873	0.866	0.785	21	1649
4	anterior	96.8	0.853	0.84	0.740	25	1894
5	aspirated	94.2	0.797	0.777	0.6648	25	2177
6	bilabial	96.6	0.947	0.947	0.9796	24	1411
7	consonant	93.3	0.895	0.893	0.958	29	2422
8	continuant	91.69	0.762	0.734	0.621	27	2689
9	coronal	92.15	0.918	0.918	0.904	34	2427
10	emphatic	96	0.804	0.784	0.663	25	1625
11	fricative	92.4	0.719	0.6738	0.535	28	1925
12	glottal	97.3	0.676	0.597	0.4554	30	1969
13	high	98.8	0.856	0.845	0.791	41	2703
14	interdental	97.7	0.808	0.788	0.651	27	2392
15	labiodental	97.8	0.911	0.907	0.824	32	2694
16	labiovelar	98.5	0.849	0.836	0.756	19	1151
17	lateral	98.3	0.933	0.931	0.872	27	1683
18	nasal	93.3	0.904	0.902	0.847	32	1915
19	palatal	97.6	0.904	0.900	0.835	22	1622
20	pharyngeal	97.7	0.872	0.864	0.8067	23	1679
21	plosive	98.1	0.770	0.737	0.631	27	1145
22	rounded	95.4	0.950	0.950	0.934	31	2508
23	semivowel	94.7	0.830	0.817	0.720	27	1938
24	short	97.9	0.779	0.75	0.626	24	1652
25	trill	96.2	0.961	0.961	0.968	27	2675
26	unvoiced	98.7	0.967	0.966	0.948	22	1863
27	uvular	89.2	0.882	0.880	0.913	37	2982
28	velar	85.5	0.855	0.855	0.849	30	2692
29	voiced	87.3	0.793	0.780	0.694	23	1920
30	vowel	98.9	0.965	0.965	0.946	27	1131

phonemes 'sb10' and 'ss10' are of lower performance as they are mutually confused. This later observation is consistent with our previous observation in Section IV.B. The t-SNE plot for the features of both phonemes, 'sb10' and 'ss10', reveals the presence of severe overlap between them, as depicted in Figure 14 (a).

TABLE 7. GA selected features with the corresponding 61-bit string for each DPF element.

No.	DPF	No. of '1s'	Frequency	MFCC	Zero Crossing	Energy	Pitch	The 61-Bit String Representation
1	affricative	31	8	10	6	7	0	1100010101100111110101111010010001100011010010010101110100100
2	alveodental	28	10	5	6	7	0	1100010111011111000100001001010101100001010010110001011001100
3	alveopalatal	21	6	7	1	7	0	1010000001100111101101000001100000000010000000110011110001000
4	anterior	25	7	6	5	7	0	1101000101100010110010000101100101011010000000100010011101100
5	aspirated	25	8	8	6	2	1	1110101100001010001101111110000111010001001000100010000000001
6	bilabial	24	5	8	5	5	1	0100001000001110111111000010100000100011001010100101000001011
7	consonant	29	9	7	5	7	1	0100101001111110001010110101010010001001000111101001010010011
8	continuant	27	10	8	5	3	1	1111100101001110010110001111100000001000010111000011000010 0001
9	coronal	34	9	7	8	9	1	1101011001010110001011011000110011111100000111111101100101001
10	emphatic	25	6	5	7	7	0	1010000100100111010010010100001010011110100000001000111100101 0
11	fricative	28	7	8	3	9	1	1011001000110011100010101111001000000100010000111010001011101 1
12	glottal	30	7	11	8	3	1	0101101110000011110111100111100001111010101010000000111000001
13	high	41	10	8	11	11	1	1101000011111111110000111100101011111110101101101100111101111
14	interdental	27	9	5	4	8	1	1100010011011111010100100000100101100000000101101100101101001
15	labiodental	32	10	8	8	5	1	1011110011011010100101001110111001001111001011010101100000001
16	labiovelar	19	4	8	3	4	0	1000000011000011001111011010000000001010000010000100010000101 0
17	lateral	27	6	9	4	7	1	1010000111000100101110011110011000000000110010111000110101001
18	nasal	32	7	7	7	11	0	1100010100011011010010011010010011001111001001110101101111100
19	palatal	22	6	5	6	5	0	0010010101001011110000001100001110100000100010001001000110010
20	pharyngeal	23	6	9	3	4	1	1000101001000110101101101101101100000000000010000110110000000 1
21	plosive	27	4	7	8	7	1	0000100100000111110100010010011000110111100101011110000000111
22	rounded	31	9	13	4	4	1	1011010110100111111010111111110101000011000000100000111000001
23	semivowel	27	7	9	6	5	0	1000011110000111110100010111101010001010110000000100001110010
24	short	24	6	7	5	6	0	1001010001100010100000111110101101001001000000100001110001010
25	trill	27	10	7	4	6	0	1111100100011111000110110001100001000110000010101011010000010
26	unvoiced	22	7	4	3	7	1	1010010001001110000000000110110100001000000101100101010100011
27	uvular	37	11	10	10	6	0	1111001011101111111011001100111110001111001110010001010001110
28	velar	30	10	8	4	7	1	1110010111100110100110111000110000101000100100011010111000011
29	voiced	23	7	8	3	5	0	1101010010010010011011010110011000010000100000000101110000001 0
30	vowel	27	4	6	8	8	1	1000001001000010110100100110001001001010011110000101111101011

Wilcoxon signed rank test [52] is used to judge the significance of the GA-FNN's results, as compared to the corresponding uttered phonemes. Wilcoxon signed-rank test is a non-parametric statistic test used for comparing two paired sets of observations whose difference comes from a distribution of zero median. In our experiments, the p-value of a two-sided Wilcoxon signed-rank test is 0.12. This result indicates that the test fails to reject the null hypothesis of zero median



(a) t-SNE of the two confused phonemes ‘sb10’ and ‘ss10’.

(b) t-SNE of the two separable phonemes ‘hz10’ and ‘ss10’.

FIGURE 14. The t-SNE algorithm of separable and non-separable phonemes.

in the difference at a significance level of 5%. That is, this means that the difference between the median of GA-FNN based outputs and that of the corresponding real phonemes’ sequence is zero; hence, the two paired of sets (model’s output and corresponding real pronounced sequence) are not statistically different.

These promising results demonstrate the effectiveness of the approach applied in this work in realizing an efficient compromise between the following three challenging requirements that are commonly encountered when developing speech processing systems: First, the ability to deal with variability in speech signal, which is a crucial requirement for system robustness. Such variability is captured in system input via a diversity of acoustic features that, in turn, would significantly increase input dimensionality and model complexity. Second, the urging need to reduce input dimensionality, which would greatly limit the involvement of multiple types of acoustic features in system input. Lastly, the fundamental requirement to increase system performance, which is directly affected by the aforementioned requirements.

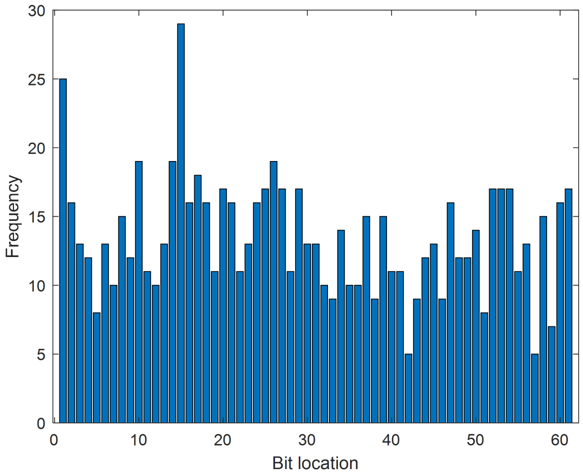
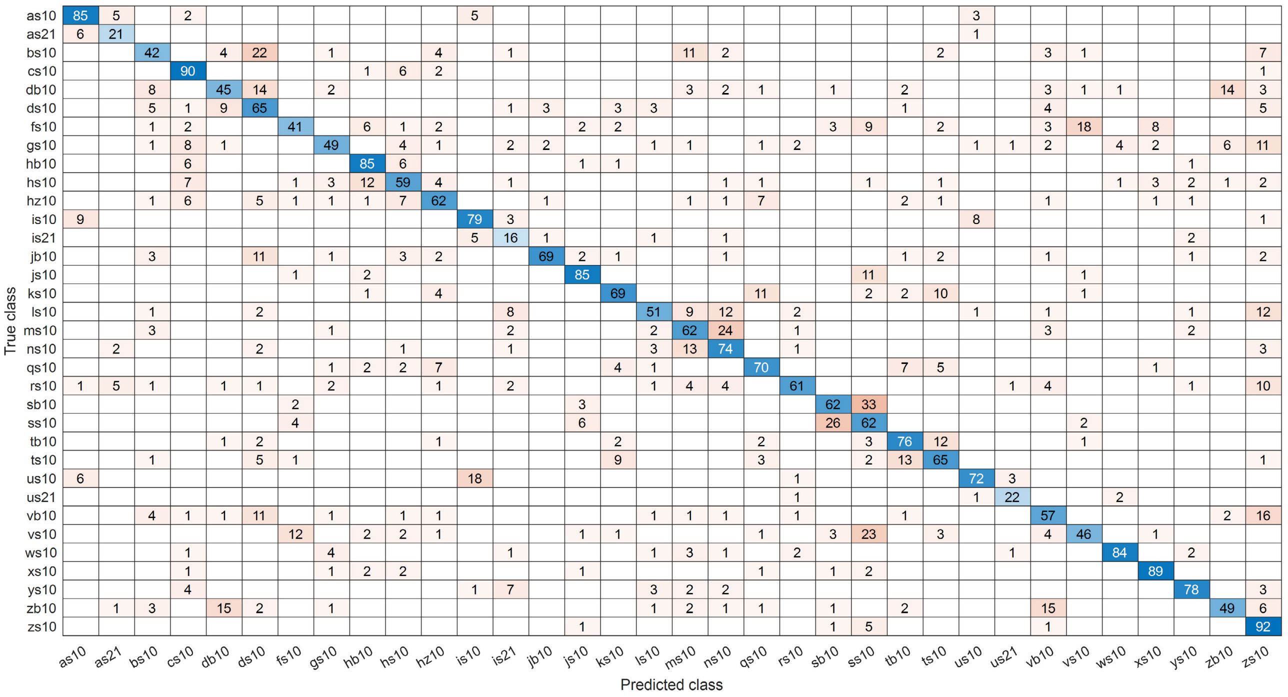


FIGURE 15. The frequency of occurrence of value ‘1’ for each entry of the 61-bit string.

Therefore, these results pave the way for more efficient DPF-based system design approaches in order to enhance system robustness by means of diversifying input acoustic cues, while maintaining lower input dimensions, and superior system performance. This effort is also validating the ability of GA to reduce features space of speech signal, in general, which is very useful in digital signal processing front-end in order to minimize CPU time and memory usage by removing duplicate redundant features. Certainly, that will have direct positive impact on real-time speech recognition systems that are implemented on low-resource computers.

VI. DISCUSSION

The GA has different parameters to configure and cost functions to estimate, which contribute to the total complexity of the entire algorithm. Therefore, each variant of GA has different time complexity based on algorithm implementation; for example, time complexity is shown to be polynomial of degree two in [53], where in [54], it is proportional to number of samples in the training set multiplied by squared number of total features, whereas in [55], it is proportional to number of features under investigation. By analyzing the whole process of the proposed GA-FFN model, as described in Section IV-B, it would not be difficult to determine the time complexity as $O(NPG)$, where N is the length of features vector, P is the population size, and G is the number of generations. This result is consistent with the finding reported in [56]. In the proposed model, the GA is only used in the training phase to select the best and optimal set on input features. The optimal configuration of input vectors composed of selected features is used in the testing phase. That is, in the testing phase GA is no longer needed, and the time complexity will be solely due to the FFN network, which is $O(N)$ [57]. On the other hand, for a training phase having a constraint of short time processing, GA full-parallel implementation on a dedicated hardware (e.g. Field Programmable Gate Arrays (FPGAs)) can be considered; see [58] and the references therein.



	as10	as21	bs10	cs10	db10	ds10	fs10	gs10	hb10	hs10	hz10	is10	is21	jb10	js10	ks10	ls10	ms10	ns10	qs10	rs10	sb10	ss10	tb10	ts10	us10	us21	vb10	vs10	ws10	xs10	ys10	zb10	zs10
as10	85	5		2								5														3								
as21	6	21																								1								
bs10			42		4	22		1			4		1					11	2						2			3	1					7
cs10				90					1	6	2																							1
db10			8		45	14		2										3	2	1		1		2				3	1	1			14	3
ds10			5	1	9	65							1	3		3	3							1				4						5
fs10			1	2			41		6	1	2				2	2						3	9		2			3	18		8			
gs10			1	8	1			49		4	1		2	2			1	1		1	2					1	1	2		4	2		6	11
hb10				6					85	6					1	1																1		
hs10				7			1	3	12	59	4		1						1	1			1		1					1	3	2	1	2
hz10			1	6		5	1	1	1	7	62			1				1	1	7				2	1			1			1	1		
is10	9											79	3													8								1
is21												5	16	1			1		1													2		
jb10			3			11		1		3	2			69	2	1			1					1	2			1				1		2
js10							1		2						85								11						1					
ks10									1		4					69				11			2	2	10				1					
ls10			1			2							8				51	9	12		2					1		1				1		12
ms10			3					1					2				2	62	24		1							3				2		
ns10		2				2				1			1				3	13	74		1													3
qs10								1	2	2	7					4	1			70				7	5						1			
rs10	1	5	1		1	1		2			1		2				1	4	4		61						1	4				1		10
sb10							2								3							62	33											
ss10							4								6							26	62						2					
tb10					1	2					1					2				2			3	76	12				1					
ts10			1			5	1									9				3			2	13	65									1
us10	6											18									1					72	3							
us21																					1					1	22			2				
vb10			4	1	1	11		1		1	1						1	1	1		1			1				57					2	16
vs10							12		2	2	1				1	1				1		3	23		3			4	46		1			
ws10				1				4					1				1	3	1		2						1			84		2		
xs10				1				1	2	2					1					1		1	2								89			
ys10				4								1	7				3	2	2													78		3
zb10		1	3		15	2		1									1	2	1	1		1		2				15					49	6
zs10															1							1	5					1						92

FIGURE 16. The confusion matrix of proposed classification system.

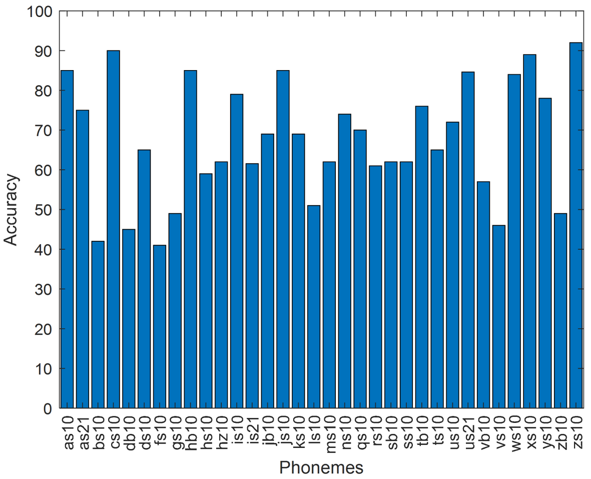


FIGURE 17. The identification accuracy for each phoneme.

This solution provides high-performance and higher speed when compared to sequential solutions.

Although the proposed GA-FNN method achieved good result in reducing the dimensionality of feature space, it has some limitations. Because this model uses GA for features selection, the running time during the training phase using Intel Core i9-9900k processor with 64 RAM is quite large (few hours). In order to cope with the problem of computational time, GA full-parallel implementation on a dedicated hardware (e.g. Field Programmable Gate Arrays (FPGAs)) can be considered [58]. Another limitation is related to the selection of the appropriate operators such as crossover and mutation to prevent algorithm divergence. Therefore, there is no guarantee of optimality of the obtained solution. On the other hand, besides its role as a classifier (phoneme recognizer) the FNN was used to estimate the GA objective function during the evaluation process of the huge amount of individuals produced through generations. This dual role assigned to FNNs as fitness estimators as well as classification and recognition engines must be assessed by comparing it to an approach using two different systems, dedicated separately to the estimation of the objective function on the one hand, and to the classification of phonemes on the other hand.

Compared to comparative methods, GA is selected to reduce the complexity of speech acoustic features in order to remove data redundancy in signal front-end processing. To the best of our knowledge, this is the first time to be considered in the literature. On the other hand, FNN was considered as a vehicle in performing GA task and to be used as a baseline in evaluating the huge amount of produced generations and chromosomes to help avoid exhaustive options. FNNs methods are well-known engines in performing classification and recognition with straightforward design methods and tune-ups. In linguistics, there are comparative methods that are based on systematic process of reconstructing the segmental and suprasegmental inventory of an ancestral language from cognate reflexes by performing a feature-by-feature comparison in the genetically related ancestor languages [59]. Indeed, comparative methods are supposed to deal with higher levels of language units such as phonemes in NLP disciplines, but the scope of the current work is to deal with acoustic feature engineering of speech signals.

VII. CONCLUSION

This work has considered the problem of reducing the size of features' vector employed for DPF and phoneme recognition. Specifically, the GA has been used to perform the features selection process. The experimental results obtained using the GA-based selection method show that a 79% reduction in the size of features' vector with performance, at least, as good as that obtained using the full-fledge features vector can be achieved. In particular, a features' vector of an average size of 3231 elements, selected by GA aided with FF-NN for phoneme recognition, has an accuracy of 68.2%, as compared to 68% obtained using the full-fledge features vector whose length is 4096 elements. For DPF recognition, the reduction in features' vector size is 50% in average with recognition accuracy of 90%. Therefore, the proposed method contributes to the reduction of computational complexity of the problem at hand with no degradation in the system's performance. Further, it opens a new direction for research, where other evolutionary algorithms can be tested for achieving further reduction in the size of features' vector.

REFERENCES

[1] Y. Alotaibi and A. Meftah, "Review of distinctive phonetic features and the Arabic share in related modern research," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 21, no. 5, p. 1426–1439, 2013.

[2] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York, NY, USA: Harper & Row, 1968.

[3] E. Eide, "Distinctive features for use in an automatic speech recognition system," in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, Aalborg, Denmark, 2001, pp. 1613–1616.

[4] M. Alkhouli, *Alaswaat Alaghawaiyah*. Amman, Jordan: Daar Alfalah, 1990.

[5] M. Alghamdi, *Arabic Phonetics*. Riyadh, Saudi Arabia: Al-Toubah Bookshop, 2001.

[6] Y. A. El-Imam, "An unrestricted vocabulary Arabic speech synthesis system," *IEEE Trans. Acoust.*, vol. 37, no. 12, p. 1829–1845, Dec. 1989.

[7] T. Fukuda, W. Yamamoto, and T. Nitta, "Distinctive phonetic feature extraction for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, Apr. 2003, p. II–25.

[8] T. Fukuda and T. Nitta, "Canonicalization of feature parameters for automatic speech recognition," in *Proc. 8th Int. Conf. Spoken Lang. Process. (ICC Jeju)*, Jeju-do, South Korea, 2004, pp. 2537–2540.

[9] T. Fukuda, M. Ghulam, and T. Nitta, "Designing multiple distinctive phonetic feature extractors for canonicalization by using clustering technique," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, 2005, pp. 3141–3144.

[10] M. N. Huda, M. Ghulam, T. Fukuda, K. Katsurada, and T. Nitta, "Canonicalization of feature parameters for robust speech recognition based on distinctive phonetic feature (DPF) vectors," *IEICE Trans. Inf. Syst.*, vol. 91, no. 3, p. 488–498, 2008.

[11] M. N. Huda, G. Muhammad, J. Horikawa, and T. Nitta, "Distinctive phonetic feature (DPF) based phone segmentation using hybrid neural networks," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, Antwerp, Belgium, 2007, pp. 94–97.

[12] M. N. Huda, K. Katsurada, and T. Nitta, "Phoneme recognition based on hybrid neural networks with inhibition/enhancement of distinctive phonetic feature (DPF) trajectories," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, Brisbane, QLD, Australia, 2008, pp. 1529–1532.

[13] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Comput. Speech Lang.*, vol. 14, no. 4, pp. 333–353, 2000.

[14] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezmaman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss, and K. Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, p. IV-621.

[15] S. Stuker, T. Schultz, F. Metze, and A. Waibell, "Multilingual articulatory features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 2003, p. 1.

[16] B. Launay, O. Siohan, A. Surendran, and C.-H. Lee, "Towards knowledge-based features for HMM based large vocabulary automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, May 2002, p. I-817.

[17] M. N. Huda, M. Ghulam, and T. Nitta, "DPF based phonetic segmentation using recurrent neural networks," in *Proc. Autumn Meeting Astronomical Soc. Jpn.*, 2006, pp. 3–4.

[18] H. Tolba, S.-A. Selouani, and D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, May 2002, p. I-837.

[19] T. Fukuda and T. Nitta, "Noise-robust ASR by using distinctive phonetic features approximated with logarithmic normal distribution of HMM," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, Geneva, Switzerland, 2003, pp. 2185–2188.

[20] S.-A. Selouani, H. Tolba, and D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for robust automatic speech recognition in low-SNR car environments," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol. Companion (HLT-NAACL)*, 2003, pp. 91–93.

[21] T. Bhowmik and S. K. Das Mandal, "Manner of articulation based Bengali phoneme classification," *Int. J. Speech Technol.*, vol. 21, no. 2, p. 233–250, 2018.

[22] F. Hartmann, "Predicting historical phonetic features using deep neural networks: A case study of the phonetic system of Proto-Indo-European," in *Proc. 1st Int. Workshop Comput. Approaches Historical Lang. Change*, Florence, Italy, 2019, pp. 98–108.

[23] Y. Korkmaz, A. Boyacğ, and T. Tuncer, "Turkish vowel classification based on acoustical and decompositional features optimized by genetic algorithm," *Appl. Acoust.*, vol. 154, pp. 28–35, Nov. 2019.

[24] J. J. Bird, E. Wanner, A. Ekárt, and D. R. Faria, "Optimisation of phonetic aware speech recognition through multi-objective evolutionary algorithms," *Expert Syst. Appl.*, vol. 153, Sep. 2020, Art. no. 113402.

[25] S.-A. Selouani and J. Caelen, "Spotting arabic phonetic features using modular connectionist architectures and a rule-based system.," in *Proc. NC*, 1998, p. 456–462.

[26] S.-A. Selouani and J. Caelen, "Arabic phonetic features recognition using modular connectionist architectures," in *Proc. IEEE 4th Workshop Interact. Voice Technol. Telecommun. Appl. (IVTTA)*, 1998, pp. 155–160.

[27] Y. Alotaibi, Y. Seddiq, A. Meftah, S.-A. Selouani, and M. S. Yakoub, "A new look at the automatic mapping between arabic distinctive phonetic features and acoustic cues," in *Proc. 40th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2017, pp. 368–371, doi: 10.1109/TSP.2017.8076007.

[28] Y. Seddiq, Y. Alotaibi, A. Meftah, S.-A. Selouani, and M. Alghamdi, "Revisiting distinctive phonetic features from applied computing perspective: Unifying views and analyzing modern arabic speech varieties," *Int. J. Speech Technol.*, vol. 21, no. 4, pp. 907–913, Dec. 2018, doi: 10.1007/s10772-018-9548-z.

[29] Y. A. Alotaibi, S.-A. Selouani, M. S. Yakoub, Y. M. Seddiq, and A. Meftah, "A canonicalization of distinctive phonetic features to improve arabic speech recognition," *Acta Acust. United Acust.*, vol. 105, no. 6, pp. 1269–1277, 2019.

[30] Y. Seddiq, Y. A. Alotaibi, S.-A. Selouani, and A. H. Meftah, "Distinctive phonetic features modeling and extraction using deep neural networks," *IEEE Access*, vol. 7, pp. 81382–81396, 2019.

[31] S.-A. Selouani, *Speech Processing and Soft Computing*. Cham, Switzerland: Springer, 2011.

[32] W. Nabi, N. Aloui, and A. Cherif, "An improved speech enhancement algorithm for dual-channel mobile phones using wavelet and genetic algorithm," *Comput. Electr. Eng.*, vol. 62, pp. 692–705, Aug. 2017.

[33] D. Goldenberg, "Genetic algorithms in search, optimization, and machine learning," *Choice Rev. Online*, vol. 27, no. 2, p. 27-0936, 1989.

[34] J. J. Roberts, A. M. Cassula, J. L. Silveira, P. O. Prado, and J. C. F. Junior, "GAtoolbox: A MATLAB-based genetic algorithm toolbox for function optimization," in *Proc. 12th Latin-Amer. Congr. Electr. Gener. Transmiss.-Clagtee*, Mar del Plata, Argentina, 2017, pp. 1–12.

[35] R. Leardi, R. Boggia, and M. Terrile, "Genetic algorithms as a strategy for feature selection," *J. Chemom.*, vol. 6, no. 5, p. 267–281, 1992.

[36] M. Aissiou, "A genetic model for acoustic and phonetic decoding of standard Arabic vowels in continuous speech," *Int. J. Speech Technol.*, vol. 23, no. 3, pp. 425–434, 2020.

[37] A. H. Abo Absa, M. Deriche, M. Elshafei-Ahmed, Y. M. Elhadj, and B.-H. Juang, "A hybrid unsupervised segmentation algorithm for arabic speech using feature fusion and a genetic algorithm (July 2018)," *IEEE Access*, vol. 6, pp. 43157–43169, 2018.

[38] M. Wroniszewska and J. Dziedzic, "Voice command recognition using hybrid genetic algorithm," *Task Quart.*, vol. 14, no. 4, pp. 377–396, 2010.

[39] K. S. Mallikarjuna and N. Sushma, "A progressive technique for duplicate detection evaluating multiple data using genetic algorithm with real world objects," *Int. J. Sci. Eng. Adv. Technol.*, vol. 6, no. 5, pp. 283–288, 2018.

[40] M. M. Alghmadi, "KACST arabic phonetics database," in *Proc. Congr. Phonetics Sci.*, vol. 15, 2003, pp. 3109–3112.

[41] Y. Seddiq, A. Meftah, M. Alghamdi, and Y. Alotaibi, "Reintroducing KAPD as a dataset for machine learning and data mining applications," in *Proc. Eur. Modeling Symp. (EMS)*, Nov. 2016, pp. 70–74, doi: 10.1109/EMS.2016.022.

[42] R. Loughran, J. Walker, M. O'Neill, and M. O'Farrell, "The use of mel-frequency cepstral coefficients in musical instrument identification," in *Proc. Int. Comput. Music Conf.*, 2008.

[43] D. J. Norris, "Exploration of ML data models: Part 1," in *Machchine Learning With Raspberry Pi*. Cham, Switzerland: Springer, 2020, pp. 49–133.

[44] A. Elkelesh, M. Ebada, S. Cammerer, and S. ten Brink, "Decoder-tailored polar code design using the genetic algorithm," *IEEE Trans. Commun.*, vol. 67, no. 7, p. 4521–4534, Apr. 2019.

[45] P. M. Hariyadi, P. T. Nguyen, I. Iswanto, and D. Sudrajat, "Traveling salesman problem solution using genetic algorithm," *J. Crit. Rev.*, vol. 7, no. 1, p. 56–61, 2020.

[46] S. Mirjalili, "Genetic algorithm," in *Evolutionary Algorithms and Neural Networks*. Cham, Switzerland: Springer, 2019, p. 43–55.

[47] H. Zhang, D. Zhong, Y. Zhong, Z. Zhang, and S. Zhan, "Comparative analysis of selection schemes used in artificial bee colony algorithm," *Int. J. Comput. Sci. Math.*, vol. 8, no. 3, p. 218, 2017.

[48] R. Abd Rahman, R. Ramli, Z. Jamari, and K. R. Ku-Mahamud, "Evolutionary algorithm with roulette-tournament selection for solving aquaculture diet formulation," *Math. Problems Eng.*, vol. 2016, pp. 1–10, Jan. 2016.

[49] S. M. Lim, U. Malaysia of Computer Science, Engineering, A. B. M. Sultan, M. N. Sulaiman, A. Mustapha, and K. Y. Leong, "Crossover and mutation operators of genetic algorithms," *Int. J. Mach. Learn. Comput.*, vol. 7, no. 1, pp. 9–12, Feb. 2017.

[50] T. Kellegöz, B. Toklu, and J. Wilson, "Comparing efficiencies of genetic crossover operators for one machine total weighted tardiness problem," *Appl. Math. Comput.*, vol. 199, no. 2, p. 590–598, 2008.

[51] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[52] R. F. Woolson, "Wilcoxon signed-rank test," in *Wiley Encyclopedia of Clinical Trials*. Hoboken, NJ, USA: Wiley, 2007, pp. 1–3.

[53] A. K. Das, S. Sengupta, and S. Bhattacharyya, "A group incremental feature selection for classification using rough set theory based genetic algorithm," *Appl. Soft Comput.*, vol. 65, pp. 400–411, Apr. 2018.

[54] M. M. Kabir, M. Shahjahan, and K. Murase, "A new local search based hybrid genetic algorithm for feature selection," *Neurocomputing*, vol. 74, no. 17, pp. 2914–2928, Oct. 2011.

[55] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognit.*, vol. 33, no. 1, pp. 25–41, Jan. 2000.

[56] P. Liu, M. D. El Basha, Y. Li, Y. Xiao, P. C. Sanelli, and R. Fang, "Deep evolutionary networks with expedited genetic algorithms for medical image denoising," *Med. Image Anal.*, vol. 54, pp. 306–315, May 2019.

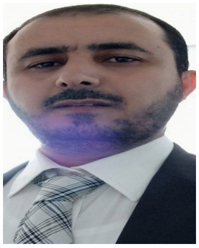
[57] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1424–1437, Nov. 2004.

[58] M. F. Torquato and M. A. C. Fernandes, "High-performance parallel implementation of genetic algorithm on FPGA," *Circuits, Syst., Signal Process.*, vol. 38, no. 9, pp. 4014–4039, Sep. 2019.

[59] M. Weiss, "The comparative method," in *The Routledge Handbook of Historical Linguistics*. London, U.K.: Routledge, 2015, p. 127.

AHMED B. IBRAHIM received the B.Sc. degree in computer science from Assuit University, Egypt, in 2007, and the Master of Biometrics degree in optics, image, vision, and multimedia from the University of Paris-Est Créteil, France, in 2016. He is currently a Researcher with the RF and Photonics for the e-Society (RFTONICS), King Saud University. His research interests include machine learning, data analysis, and signal/image processing algorithms related to biometrics and biomedical signals.

YASSER MOHAMMAD SEDDIQ received the B.S. degree in computer engineering from the King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, in 2004, and the M.S. and Ph.D. degrees in computer engineering from King Saud University (KSU), Riyadh, Saudi Arabia, in 2010 and 2017, respectively. He is currently an Assistant Research Professor with the King Abdulaziz City for Science and Technology (KACST), Riyadh. His research interests include digital signal processing, speech processing, image processing, computer arithmetic, and digital systems design.



ALI HAMID MEFTAH received the B.Sc. and M.Sc. degrees in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 2009 and 2015, respectively, where he is currently pursuing the Ph.D. degree. Since 2010, he has been a Researcher with King Saud University. His research interests include digital speech processing specifically speech recognition and Arabic language and speech processing.

MANSOUR ALGHAMDI received the Ph.D. degree in speech analysis, synthesis, and perception from the University of Reading, in 1990. He took several positions, including the General Director of scientific awareness and publishing with KACST. He is currently a Consultant at the Education and Training Evaluation Commission. He has more than 80 published articles and books, and five patents. He is also a PI and a Team Member of more than 20 scientific research projects that produced software systems, algorithms, and databases. He has supervised and examined several Ph.D. students. He has lectured and reviewed articles and projects in his field. His published work has more than 1000 citations with 20 H-index on Google Scholars. He has been working in public sectors for 47 years.

SID-AHMED SELOUANI (Senior Member, IEEE) received the Electrical Engineering degree and the master's degree in electronics from the University of Technology of Algiers (USTHB), in 1987 and 1991, respectively, and the Doctorate of Science degree in electrical engineering from Joseph Fourier University (USTHB), Grenoble, in 2000. He is currently a Full Professor with the Université de Moncton, Shippagan campus (UMCS), NB, Canada. He is also the Founder of the Human-System Interaction Research Laboratory (LARIHS), Université de Moncton, and the DILAN Center for the Development of the Language Industry, UMCS. He is also a Visiting Professor with INRS-Énergie-Matériaux et Télécommunications, Montreal. His main research interests include artificial intelligence, human–machine interaction, the Internet of Things, deep learning and bio-inspired algorithms, multimodal dialogue systems, and ubiquitous and distributed systems. He has more than 210 refereed publications and has received numerous grants from major Canadian research councils.

MUSTAFA A. QAMHAN received the B.Sc. degree in information technology from the Faculty of Engineering and Information Technology, Taiz University, Yemen, in 2008, and the master's degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 2015, where he is currently pursuing the Ph.D. degree. He was a Computer Engineer with Public Telecommunication Company (PTC), Yemen. He is currently a Research Assistant with King Saud University. His main research interests include digital signal processing, speech processing, and artificial intelligence.

YOUSEF A. ALOTAIBI (Senior Member, IEEE) received the B.Sc. degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 1988, and the M.Sc. and Ph.D. degrees in computer engineering from the Florida Institute of Technology, FL, USA, in 1994 and 1997, respectively. From 1988 to 1992 and 1998 to 1999, he was a Research Engineer with Al-ELM Research and Development Corporation, Riyadh. From 1999 to 2008, he was an Assistant Professor with the College of Computer and Information Sciences, King Saud University, where he was an Associate Professor, from 2008 to 2012. Since 2012, he has been a Professor with the College of Computer and Information Sciences, King Saud University. He has supervised many B.Sc., M.Sc., and Ph.D. thesis in the department. He has published tens of conference and journal papers. His research interests include digital speech processing specifically speech recognition and Arabic language and speech processing.

SALEH A. ALSHEBEILI was the Chairman of the Electrical Engineering Department, King Saud University, from 2001 to 2005. He has over 27 years of teaching and research experience in the area of communications and signal processing. He was a member of the Board of Directors with the King Abdullah Institute for Research and Consulting Studies, from 2007 to 2009, and a member of the Board of Directors with the Prince Sultan Advanced Technologies Research Institute, from 2008 to 2017, where he was the Managing Director, from 2008 to 2011, and the Director of the Saudi-Telecom Research Chair, from 2008 to 2012. He has been the Director of the Technology Innovation Center, RF and Photonics in the e-Society, funded by the King Abdulaziz City for Science and Technology (KACST), since 2011. He is currently a Professor with the Electrical Engineering Department, King Saud University. He has also an active involvement in the review process of a number of research journals, KACST general directorate grants programs, and national and international symposiums and conferences. He has been on the Editorial Board of the *Journal of Engineering Sciences*, King Saud University, from 2009 to 2012.