IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 4, 2021, accepted May 14, 2021, date of publication May 26, 2021, date of current version June 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3083972

Sequence-to-Sequence Acoustic-to-Phonetic Conversion Using Spectrograms and Deep Learning

MUSTAFA A. QAMHAN[1], **YOUSEF AJAMI ALOTAIBI**[1], **(Senior Member, IEEE)**, **YASSER MOHAMMAD SEDDIQ**[2], **ALI HAMID MEFTAH**[1], **AND SID AHMED SELOUANI**[3], **(Senior Member, IEEE)**

[1]College of Computer and Information Sciences, King Saud University, Riyadh 12372, Saudi Arabia
[2]King Abdulaziz City for Science and Technology (KACST), Riyadh 12354, Saudi Arabia
[3]LARIHS Laboratory, Université de Moncton at Shippagan, Shippagan, NB E8S 1P6, Canada

Corresponding author: Mustafa A. Qamhan (mqamhan@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 (DPFs) abstractedly describe the place, manner of articulation, and voicing of the language phonemes. While DPFs are powerful features of speech signals that capture the unique articulatory characteristics of each phoneme, the task of DPF extraction is challenged by the need for efficient computational model. Unlike the ordinary acoustic features that can be directly determined form speech waveform using closed-form expressions, DPF elements are extracted from acoustic features using machine learning (ML) techniques. Therefore, for the objective of developing an acoustic-to-phonetic converter of high accuracy and low complexity, it is important to select the input acoustic features that are simple, yet carry adequate information. This paper examines the effectiveness of using spectrogram as the acoustic feature with DPFs modeled using two deep learning techniques: the deep belief network (DBN) and the convolutional recurrent neural network (CRNN). The proposed method is applied on Modern Standard Arabic (MSA). Multi-label modeling is considered in the proposed acoustic-to-phonetic converter. The learning techniques were evaluated by proper evaluation measures that accommodate the imbalanced nature of DPF elements. The results showed that the CRNN is more accurate in extracting the DPFs than the DBN.

INDEX TERMS Distinctive phonetic features, spectrograms, speech processing, convolutional recurrent neural network, deep belief networks, KAPD corpus, Arabic, MSA.

I. INTRODUCTION

Distinctive phonetic features (DPFs) are relevant and highly descriptive features of speech waveforms that have the remarkable ability to capture and represent the unique articulatory characteristics of each phoneme [1]. which are relevant and highly descriptive features of speech waveforms [1]. A DPF vector is simply organized as a sort of binary elements that outlines the phonemes in terms of their articulatory and vocal properties [1]. Each phonological component of a language is either present which is typically marked as "+," or absent which is typically marked as "−." By way of illustration, in phonology we encounter this typical phoneme /θ/ which is phonetically characterized as "unvoiced," "fricative," "interdental," or "consonant." Hence, a potential DPF vector of /θ/ can be pointed out as voiced−, consonant+, fricative+, interdental+. Since languages differ in terms of their DPF elements, DPFs are not neutral but actually a language-dependent.

A. BACKGROUND ON DPF ELEMENTS

Phonemes are uttered by realizing the relevant DPF elements in coordination between the speaker's brain and vocal system [2]. As an example of phoneme classification by DPFs, consider the two English phonemes /p/ and /b/ that have all DPF elements equal except the "voicing" element. That is, since /b/ is generated by vibrations of the vocal folds, the voicing element is "+." Conversely, no vibrations of

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

TABLE 1. DPF values of MSA phonemes [1].

	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	sˤ	−	+	−	+	−	−	+	+	+	+	+	−	+	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−
15	ض	db10	dˤ	−	+	−	+	−	−	+	+	+	+	−	−	+	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	+	−
16	ط	tb10	tˤ	−	+	−	+	−	−	+	−	+	+	−	−	+	−	−	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−
17	ظ	zb10	ðˤ	−	−	−	+	−	−	+	+	+	+	+	−	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−
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	ه	hs10	h	−	−	−	−	−	−	+	+	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
27	و	ws10	w	−	−	−	+	−	+	−	+	−	−	−	−	−	−	−	+	−	−	−	−	−	+	+	−	−	−	−	−	+	−
28	ي	ys10	j	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−	−	+	−
29	ـَ	as10	a	−	−	−	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	+
30	ـَا	as21	a:	−	−	−	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+
31	ـِ	is10	i	−	−	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	+
32	ـِي	is21	i:	−	−	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+
33	ـُ	us10	u	−	−	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	+	−	−	−	−	+	+
34	ـُو	us21	u:	−	−	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	+	+

vocal cords when uttering /p/, and, hence, the voicing element is "−" in the DPF elements vector [2].

The literature on phonology offers language-specific tables listing the finite given DPF vector values [3]. However, owing to a contextual variation known as the *coarticulation effect*, the DPF vector of a spoken phoneme can differ from the theoretical presumption. Although the shape of the vocal tract adapts to the uttered sequence of phonemes, there is a limitation in the change rate of the utterances; consequently, the DPF elements gradually change during the transition periods before and after the designated phoneme. When this smooth transition causes overlap of adjacent phonemes, features that influence each other may be gained or lost [4].

Because DPFs describe speech signals both contextually and phonetically, they enhance the performance and robustness of a system [4]. Their advantages of language-specific experiments may be maximized. This paper focuses on Modern Standard Arabic (MSA) DPF modeling. In this research, the DPF elements considered are described in Table 1 [1], In addition, the table provides a mapping between International Phonetic Alphabet (IPA) [5] and King Abdulaziz City for Science and Technology (KACST) [6] symbols which will be used in the rest of the paper.

B. OBJECTIVES

The common approach applied in the literature for DPF extraction is converting acoustic features to DPF elements. Therefore, a preprocessing stage is always assumed where raw speech waveform is converted to acoustic features. Consequently, this stage has direct impact on the performance and complexity of the DPF extractor (the converter). If this preprocessing stage is designed to deliver acoustic features that are simple to compute yet carries significant information about the waveform, then that would greatly facilitate the design of the DPF extractor.

There are few acoustic features that are commonly considered in the published literature. The use of mel-frequency cepstral coefficients (MFCC) and local features (or just one of them) as input vector to the DPF extractor has been addressed by many studies such as [4], [7], [8]. The work in [9], [10] examined a more diverse combination of features consisting of spectrogram, MFCC, zero-crossing rate, short-time energy and pitch. The work in [11] investigates the use of mel-spectrogram and its first and second derivatives.

The aforementioned studies demonstrate how acoustic features, when combined with each other, are effective in DPF extraction. However, these acoustic features are either complex in terms of dimensionality or in steps to compute. We believe that input vectors can be further simplified without jeopardizing system performance. Moreover, there is a noticeable research gap related to Arabic DPF extraction. Thus, in this work, we investigate the effectiveness of spectrogram as input to the DPF extractor. To achieve this goal, a multi-label classification approach is proposed in order to provide a comprehensive model of the DPF vector as a whole entity. This approach is more aligned with how DPFs are naturally generated and perceived. We adopted sequence-to-sequence methodology where we convert data from the acoustic space to the phonetic one. For the purpose of DPF modeling, two techniques are used: the deep belief networks (DBN) and the convolutional recurrent neural networks (CRNN). Both types of networks are recognized for their significant modeling power.

II. DEEP LEARNING TECHNIQUES AND DATASET

A. DEEP BELIEF NETWORKS

If the weights of deep neural networks (DNNs) are properly initialized, they can achieve high accuracy [12] beside being able to model the complex, highly nonlinear relationships of speech signal [12]–[14]. Effective weight initialization in DNNs can be achieved by the use of the Restricted Boltzmann Machine (RBM) [15], [16]. An RBM is a bipartite, fully-connected networks containing two layers (visible and hidden). DBN is used to pre-train RBMs to generate initialization values. The weights initialized by a DBN must then be fine-tuned before we get a DBN–DNN [12], [13], [15]. DBN–DNNs have proven their advantages in digital speech-processing applications [13], [14], [16]–[24].

The probability $p(\mathbf{v}, \mathbf{h})$ of a joint configuration between a visible and a hidden unit, and the probability $p(\mathbf{v})$ of configuring a visible unit, are determined by normalizing the energy $E(\mathbf{v}, \mathbf{h})$ by the partitioning function. The calculations are respectively given by [13], [16]

$$p(\mathbf{v}, \mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{\sum_{\mathbf{v}',\mathbf{h}'} e^{-E(\mathbf{v}',\mathbf{h}')}} \tag{1}$$

$$p(\mathbf{v}) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}}{\sum_{\mathbf{v}',\mathbf{h}'} e^{-E(\mathbf{v}',\mathbf{h}')}}. \tag{2}$$

B. CONVOLUTIONAL RECURRENT NEURAL NETWORKS

The CRNN model combines a convolutional neural network (CNN) with long short-term memory (LSTM), thus exploiting the spatial and temporal features of both networks.

Any CNN architecture is a collection of ordered neural networks with various layers placed in a specific order. Each layer in one network makes a specific contribution. A typical model comprises several convolutional layers that convolve the input features A with a variety of kernels K (known as filters) to obtain a feature map S as follows:

$$S = A^{*}K, \text{ where S(i,j)} = \sum_{\text{n}} \sum_{m} A(i-m, i-n).K(m, n) \tag{3}$$

The kernel coefficients (like the neural network weights) are learned throughout the training phase.

The pooling layers in the CNN gather the concentrated activation features H obtained by adding the bias matrix B to S and applying a nonlinear activation function on S. In this way, the pooling layer reduces the spatial resolution of the maps as mentioned earlier.

The LSTM network, originally designed to overcome the vanishing gradient problem of a conventional RNN, adaptively learns the patterns in a time-variable sequence. We considered that combining LSTM with a CNN would reap the benefits of both connectionist architectures. This hybrid configuration is detailed in the following subsections.

C. USED CORPUS

In this paper we used the KACST Arabic Phonetic Database (KAPD) [6], which contains 35,981 phonemes that have imbalanced distribution. The cumulative period of the corpus is 1.2 hours. The KAPD is an isolated-word MSA speech corpus. Seven male's speakers have participated in recording audio material. The 26,499 tokens in the dataset are randomly split into a training subset (80%) and a test subset (20%). The KAPD data were manually segmented at the phoneme level by trained personnel under expert supervision.

III. EXPERIMENTS

A. SELECTED FEATURES (SPECTROGRAMS)

The spectrogram carries much useful information on speech signals, such as formants (F1, F2, F3, ...), high frequency components, pitch information and periodicity, and energy. It discriminates almost all phonetic features to various degrees [25], [26]. Each spectrogram is represented as a two-dimensional graph with time on the horizontal axis and frequency on the vertical axis. The strength of the color in the spectrogram represents the amplitude of the corresponding signal. Additionally, the color intensity at a particular point in the graph correlates with the signal frequency. The color ranges from light blue at low amplitudes to dark red at the highest amplitude.

The spectrograms in this work were obtained by short-time fast Fourier transform (FFT) of the speech signals, which generates a time–frequency representation. Here, the spectrograms were extracted by applying 64-point FFT on a set of frames sampled across a certain phoneme. Through this operation, we extracted 15 evenly spaced frames per phoneme. Each frame was preprocessed by 20-ms Hamming windowing, DC removal, and pre-emphasis ($\alpha = 0.97$). As the phonemes vary in length, the frame step size was set to cover the phoneme duration satisfying the required number of frames. For a small portion of short-length phonemes in the corpus, the phenome length was augmented with zero-padding prior to the framing process.

It is often preferred to train DNN on learning complex representations rather than imposing them. For that, in some applications, spectrogram produces better accuracies against the commonly used MFCC.

B. MODEL ARCHITECTURE

The implemented CRNN model consists of three convolutional layers (*Conv1, Conv2*, and *Conv3*) connected to one bidirectional LSTM layer, followed by one fully-connected (FC) layer and a softmax layer. The network input is a (64 × 15) feature matrix.

The input spectrogram segments are processed by the first convolutional layer (i.e., *Conv1*), which has 16 kernels of size 12 × 16 applied with a stride of one. This is followed by an exponential linear unit (ELU) activation function and a max pooling layer of size 2 × 2 with a stride of two. Here, the ELU replaces the typical sigmoid function to improve the efficiency of the training process. The second layer *Conv2* has 24 kernels of size 8 × 12, which are applied to the input with a stride of one. Similarly, *Conv3* has 32 kernels of size 5 × 7. Each of these convolutional layers is followed by an ELU unit. After applying these three layers, the bidirectional LSTM layers are applied with a batch size of 128. To avoid overfitting, the bidirectional-LSTM layer is followed by a dropout layer with a dropout ratio of 40%. Finally, one FC layer (connected to the previous layer) is applied with 34 different phonemes for phoneme classification, or 30 binaries ("+" or "−") for DPF classification. In this research, the 34 phonemes and binary DPF outcomes were constrained by the used KAPD corpus. The proposed DBN model contains two hidden layers composed of RBMs with 256 neurons (processing units) in each layer and a sigmoid activation function.

C. MODEL TRAINING

The proposed CRNN and DBN models were implemented in TensorFlow [27], applying Keras at the front end [28]. Spectrograms were generated for all audio files in the used dataset. Eighty percent of the data were dedicated to training; the remainder were reserved for the testing phase. The data were randomly split during each new training run. The models were trained using a NVIDIA GeForce RTX 2080 Ti graphics processing unit with 11 GB memory. The training process was run for a maximum of 200 epochs with a batch size of 64 samples. The proposed CRNN model was trained using the adaptive gradient descent algorithm Adam [29] as the optimizer with a learning rate of 0.001. For the DBN model, we pre-trained a stack of RBMs with a learning rate of 0.05 over 10 epochs, and fine-tuned the parameters with a learning rate of 0.1 over 200 iterations. The CRNN and DBM training times were approximately 100 minutes and 47 minutes, respectively.

D. ACOUSTIC-TO-PHONETIC CONVERSION

Every spoken language has its own set of phonemes and its own relevant and well-built-in DPF elements. For example, "emphatic" is a major DPF elements in Arabic languages and is included in all Arabic dialects, but is not found in English. This paper proposes two DPF extractors based on the DBN and CRNN models, which attempt to find the weak and strong correlations between the considered DPF elements and the acoustic features embedded in Arabic speech and language.

If a specific DPF elements has a strong and confirmed presence in a specific phoneme (or part of that phenome), that DPF elements should be extracted with a high accuracy rate. Conversely, if a given DPF elements is extracted with low accuracy by both extractors, it is probably irrelevant to the specific phoneme. This low accuracy can be attributed to a slight presence in the neighboring phonemes introduced by the co-articulation effect. The extractor learns the relevant DPF elements of MSA Arabic phonemes and maintains them in the proposed set of DPFs. This technique provides the relevant DPF elements of Arabic languages through acoustical experimental methods. In this sense, it differs from the purely linguistic approaches found in the literature.

E. PHONEME CLASSIFICATION

Thirty-four phenomes (see Table 1) were classified by two DNN classifiers (DBN and CRNN). Here we investigated the effect of batch size and number of iterations on the training performance. To optimize the batch size, we first observed the DBN performance for batch sizes of 32, 64, 128, and 256. In the given values, the highest accuracy performance (82.3%) was obtained for a batch size of 64. The batch-size

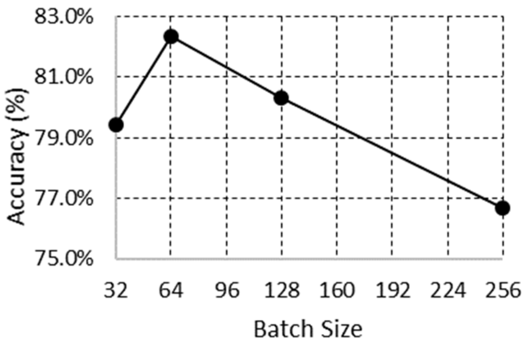


FIGURE 1. Tune up of the batch size.

tune-up results are illustrated in Figure 1. The optimal batch size depends on the corpus configuration, that is, on the number of audio files and the file-duration statistics.

TABLE 2. Model parameters of the DBN–DNN phoneme classifier.

Parameter	Value
Hidden layers structure	256, 256
Learning rate rbm	0.05
Learning rate	0.1
No. of epochs rbm	10
No. of iteration backpropagation	200
Batch size	64
Activation function	‘sigmoid'
Dropout	0.2

TABLE 3. Model parameters of the CRNN phoneme classifier.

Network	kernels
Conv1	kernel 16(12 × 16), stride 1
Pool1	kernel 2 × 2, stride 2
Conv2	kernel 24(8 × 12), stride 1
Pool2	kernel 2 × 2, stride 2
Conv3	kernel 32(5 × 7), stride 1
Pool3	kernel 2 × 2, stride 2
Activation function	'Relu'
Dropout	0.4

Next, the DBN–DNN and CRNN phoneme classifiers were developed under the specifications in Table 2 and Table 3, respectively. The parameters listed in these tables were selected and evaluated by trial-and-error. The shown parameters yielded the lowest system errors in all runs and classifiers.

IV. RESULTS AND DISCUSSION

When evaluating the performances of the DBN and the CRNN DPF extractors, the accuracy scores of the DPF element classifications must be carefully interpreted to avoid erroneous conclusions. In MSA phonology the distribution of + and − classes is unbalanced as well as in KAPD. The accuracy paradox is the paradoxical finding, which asserts that a high-accuracy classifier is not inherently better than a lower-accuracy one, is often invoked by classifying imbalanced results [30]. The paradox arises because the minority class is overwhelmed by the majority class. In consequence, the minority class will not affect the overall accuracy even when all of its members are not usefully classified.

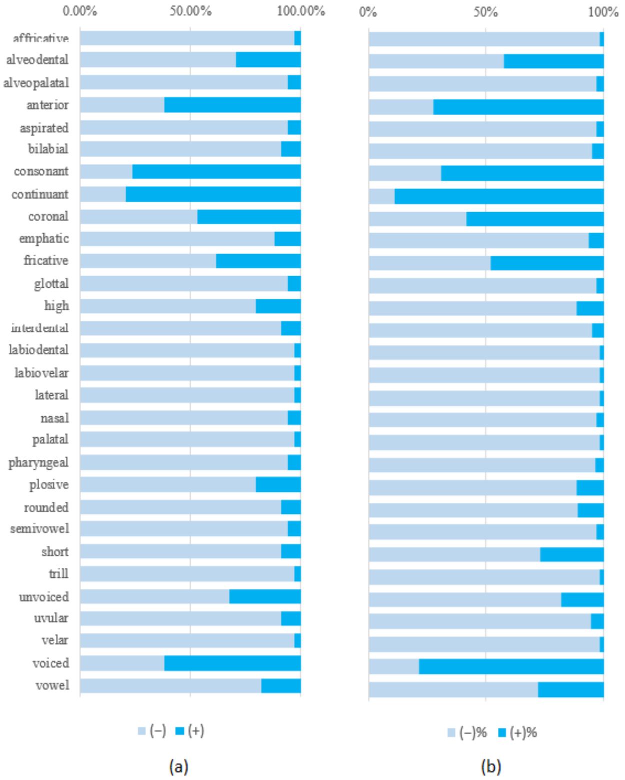


FIGURE 2. Distributions of + and − classes across the DPF elements in (a) the MSA phonology and (b) KAPD corpus.

Panels (a) and (b) of Fig. 2 show the +/− distributions of the DPF elements as derived from Table 1 and from KAPD corpus respectively, which are both typically imbalanced. The “−”class dominates all elements except the “continuant” element, which is dominated by the “+” class. Only four elements are almost balanced: the “anterior,” the “consonant,” the “coronal,” and the “voiced” features.

A. EVALUATION METRICS

Several measures can accurately evaluate the performances of imbalanced-data classifiers. Assuming that the accuracies of the majority and minority classes are the true negative rate (TNR) and true positive rate (TPR), respectively,

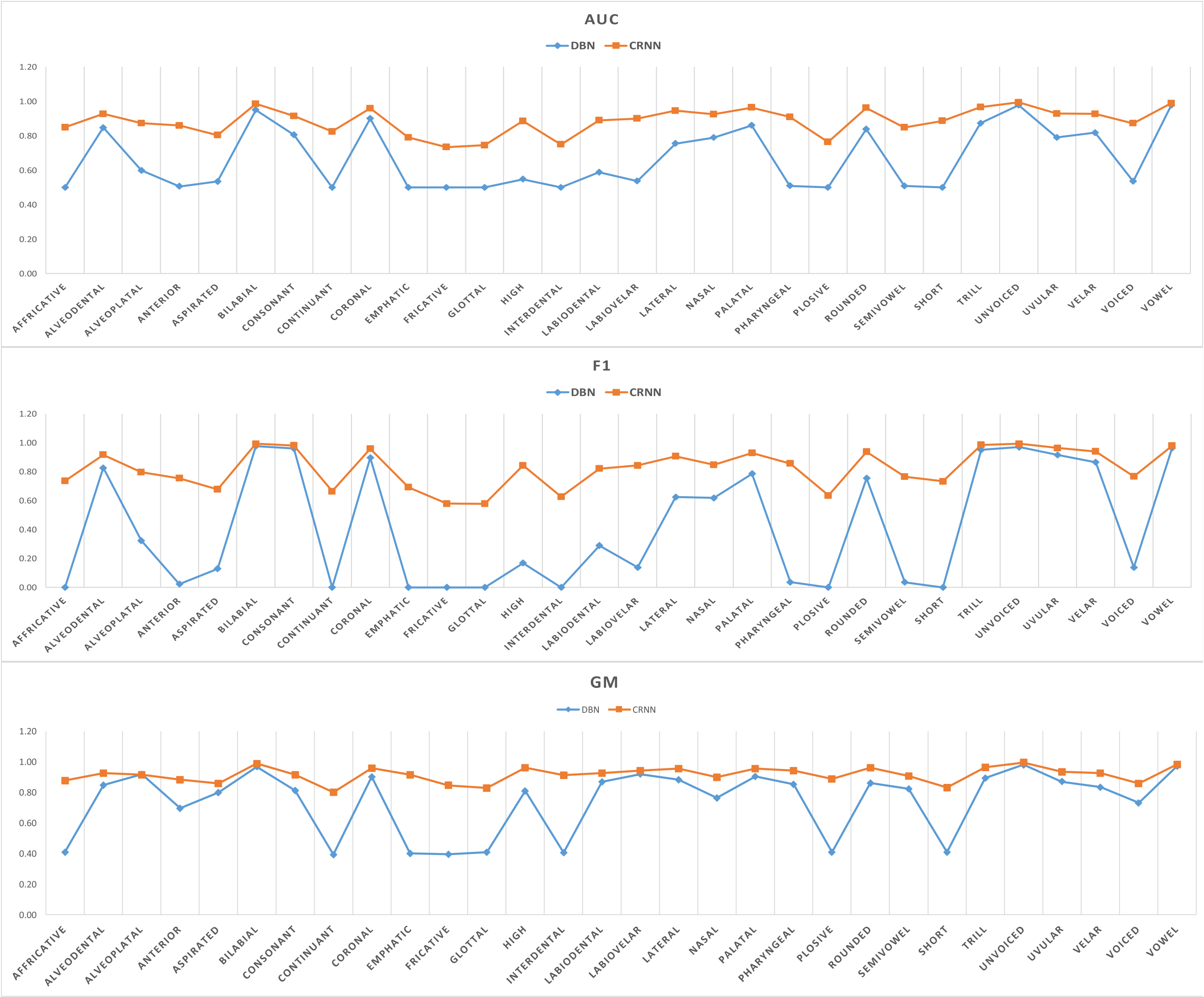


FIGURE 3. ***AUC*, F1, and *GM* performances of the extractors in the DBN (blue) and CRNN (orange) classifiers.**

we evaluate the performances of our proposed phoneme classifiers by three widely used measures, the area under the curve (AUC), the geometric mean (*GM*), and the F-measure [31].

The ROC is a graphical plot that depicts the tradeoff between *FPR* versus the *TPR* in the $x - y$ plane. The ROC checks that a classifier does not enhance the *TPR* by jeopardizing the *TNR*. Both the *TPR* and *TNR* must be high. Therefore, the area under the curve (*AUC*), which is proportional to the classifier performance, must also be high. for a binary classifier. The AUC can be calculated as follows [31]

$$AUC = \frac{1}{2}(1 + TPR - FPR) \tag{4}$$

The *GM*s of the classifier results balance the *TPR* and *TNR*. Any decrease in either value will reduce GM. The *GM* is computed as follows [31]

$$GM = \sqrt{TPR \cdot TNR}. \tag{5}$$

The classifiers were finally evaluated by the F-measure, which defines the harmonic mean of the precision (also called positive predictive value PPV) and recall (also known as True Positive Rate TPR). The F-measure can be calculated as follows [31]:

$$F = \frac{2 \cdot PPV \cdot TPR}{PPV + TPR}. \tag{6}$$

B. PERFORMANCE EVALUATION

The *AUC*, F1, and *GM* performances of the extractors in the DBN and CRNN classifiers are shown in Figure 4. These performance metrics are expected to resolve the data-imbalance problem, as explained previously. In Figure 4, all three metrics were more-or-less stable in the CRNN extractor, but the DBN performance was variable. Especially, the F1 metric of the DBN extractor repeatedly fell to zero. As an example of DPF elements extraction by the CRNN, we consider the

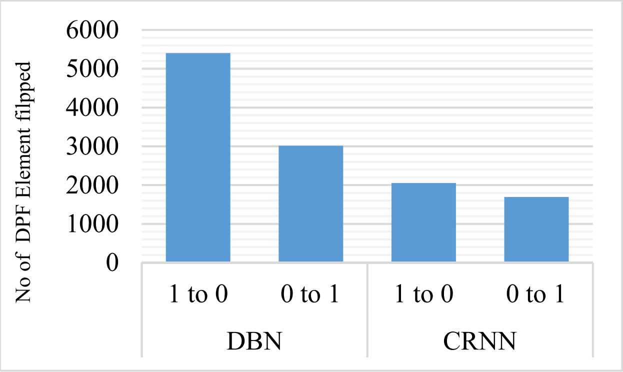


FIGURE 4. Numbers of mistakenly converted DPF elements in each classification system.

balanced element "consonantal" and the imbalanced element "high" in Table 1.

For the "consonantal" DPF elements, the normal *AUC*, F1, and *GM* measures were 0.97, 0.91, 0.98, and 0.81, respectively. Meanwhile, the normal *AUC*, F1, and *GM* measures of the "high" DPF elements were 0.1, 0.89, 0.84, and 0.71, respectively. Note that the normal accuracy measures were relatively consistent for the balanced DPF, but varied for the imbalanced DPF.

Ideally, one of the unique vectors mentioned in the search table should fit the DPF vector generated by the extractor, as depicted in Table 1. To evaluate the outcomes of the extractors, we thus seek a match in the lookup table. In fact, it is expected that the DPF vectors of spoken phonemes deviate slightly from their references in the search table based on general theoretical phonology. The deviation is at least partly caused by the co-articulation effect. Such deviations are unavoidable even under the idealized conditions of perfect DPF models. Other deviations can be introduced by modeling imperfections, which must be minimized. For this purpose, we must elucidate the contribution of the DPF extractor to the vector deviation from the ideal.

When evaluating a practical DPF extractor, we cannot expect 100% similarity to a reference vector, because the co-articulation effect is inevitable. Table 4 lists the similarities of the CRNN- and DBN-extracted files to the reference files, expressed as Hamming distances. When the Hamming distance equals 0, the extracted vector is 100% similar to a valid reference vector. Among the vectors extracted by DBN and CRNN, 51.33% and 78.68% were 100% similar to a reference vector, respectively. At a similarity-tolerance threshold of 90%, implying an error tolerance of 3 bits out of 30. The percentages of acceptable vectors returned by DBN and CRNN were 79.62% and 92.53%, respectively as can be seen in Table 4.

That is, a match is considered when any vector with a Hamming distance not exceeding 3 bits from a reference vector. In case of the output vector does not match any vector in the lookup table it will be considered as an invalid-output and is scored as an extraction error. Meanwhile, a confusion error occurs when the existing vector of a wrong phoneme matches some erroneous output vector.

TABLE 4. Test and reference phonemes classified by hamming distance.

Hamming Distance	Matching files out of 5,299			
	CRNN		**DBN**	
	No. of files	**%**	**No. of files**	**%**
0	4169	78.68	2720	51.33
1	241	4.55	449	8.47
2	266	5.02	554	10.45
3	227	4.28	496	9.36
4	152	2.87	454	8.57
5	110	2.08	357	6.74
6	66	1.25	157	2.96
7	41	0.77	85	1.60
8	14	0.26	20	0.38
9	11	0.21	4	0.08
10	2	0.04	3	0.06

1) PERFORMANCE OF THE DPF EXTRACTORS

The performances of the DBN and CRNN extractors are shown in Table 5. As shown in this table and other tables and figures in this paper, KACST symbolization is considered instead of IPA ones but mapping to IPA can be found in Table 1. Both extractors were evaluated on a test set of 5,299 phonemes. Each phoneme to be inspected by each extractor was described by 30 DPF elements, yielding a total of 158,970 (5299 × 30) DPF elements in the testing subset. The systems with the DBN and CRNN extractors missed 8,409 and 3,516 DPF elements out of 158,970, respectively. Most of the missed DPF elements in both extractors belonged to the "velar," "alveodental," "uvular," and "voiced" categories.

Conversely, the four most accurately extracted DPF elements were "short," "labiovelar," "high," and "fricative" in the DBM system, and "labiovelar," "unvoiced," "high," and "short" in the CRNN system. The CRNN generally outperformed the DBN. Note that the four worst-extracted DPF elements, and three of the best-extracted DPF elements ("short," "labiovelar," and "high"), were common to both extractors. Table 7 details the performances of the DBN and CRNN extractors on each DPF elements. From this table, we can understand the accuracies of the extractors for both DPF elements and their related phonemes. Listed are the numbers of correctly predicted DPF elements among the 5,299 elements (31 DPF elements per phoneme) in the test files. The numbers of DPF elements that were mistakenly toggled from 1 to 0, or mistakenly toggled from 0 to 1, are also reported.

Figure 4 compares the numbers of DPF elements wrongly flipped from 1 to 0 (and vice versa) in the two classification systems. Both classifiers were more likely to flip the result from 1 to 0 than from 0 to 1.

TABLE 5. Phoneme recognition accuracies (%) (in ascending order).

Phonemes	CRNN	Phonemes	DBN
bs10	45.01	db10	35.48
db10	47.72	us10	44.55
tb10	49.91	bs10	44.65
zb10	57.83	ss10	46.55
ds10	58.77	is21	50.77
ss10	59.27	sb10	53.14
vs10	62.72	zb10	53.33
fs10	64.96	vb10	56.74
gs10	65.33	ms10	58.93
sb10	66.91	ds10	60.12
vb10	67.32	fs10	61.26
hz10	67.44	ls10	61.54
ts10	67.62	rs10	62.76
ks10	68.09	ts10	62.79
as21	68.65	as21	62.94
ms10	69.13	vs10	63.72
hs10	70.22	hz10	64.47
qs10	72.07	jb10	64.71
ns10	72.83	gs10	65.98
rs10	73.55	qs10	66.25
is21	73.61	tb10	67.57
jb10	75.56	ns10	68.57
ls10	79.98	ks10	72.24
ys10	87.23	hs10	72.38
xs10	87.72	xs10	80.10
ws10	89.07	js10	81.36
cs10	89.11	hb10	82.02
js10	90.91	cs10	84.22
hb10	91.46	ws10	85.06
us21	92.36	ys10	86.38
us10	92.56	as10	93.22
is10	94.10	is10	93.36
as10	94.50	us21	95.14
zs10	98.44	zs10	98.56
Overall	**84.02**	**Overall**	**81.75**

TABLE 6. Number of missed DPF elements (in ascending order).

DPF name	CRNN	DPF name	DBN
labiovelar	24	short	67
unvoiced	27	labiovelar	75
high	30	high	79
short	31	affricative	87
affricative	40	glottal	87
lateral	40	plosive	87
plosive	48	unvoiced	88
bilabial	55	vowel	101
glottal	55	interdental	104
interdental	55	lateral	117
vowel	55	labiodental	153
pharyngeal	56	pharyngeal	161
labiodental	62	alveoplatal	164
alveoplatal	63	anterior	171
anterior	71	bilabial	182
palatal	78	emphatic	187
emphatic	89	palatal	214
trill	106	aspirated	244
rounded	115	fricative	262
semivowel	121	semivowel	274
aspirated	149	continuant	314
consonant	166	consonant	387
fricative	169	trill	416
nasal	178	nasal	444
continuant	193	rounded	464
coronal	204	coronal	528
voiced	264	voiced	576
uvular	267	uvular	693
alveodental	346	alveodental	796
velar	359	velar	887
Total	**3516**		**8409**

2) PERFORMANCE OF THE PHONEME CLASSIFIERS

Table 5 lists the overall accuracies of all phonemes in both extractors. The overall accuracies of the DBN and CRNN phoneme classifiers were 81.75% and 84.02%, respectively. The classification accuracy of the zs10 phoneme exceeded 98% in both classifiers. This phoneme was biased in the KAPD corpus, as it exists in every training and testing audio file (as explained in the original corpus datasheet). The short vowel phonemes also ranked among the best recognized phonemes in both extractors. Meanwhile, the Arabic emphatic phonemes (db10, tb10, zb10, sb10) were poorly recognized by both classifiers. Inspecting the confusion matrices of both classifiers, we observe that most phonemes were confused with their minimal pair counterparts, which increased the error rates of the classifiers. For example, the empathic phoneme sb10 was frequently confused with its corresponding non-emphatic phoneme ss10.

TABLE 7. Element-wise statistics of the extracted DPF elements.

DPF	Total number of 0s	Total number of 1s	Total	DBN				CRNN			
				Zero error	Toggled from 1 to 0	Toggled from 0 to 1	Total	Zero error	Toggled from 1 to 0	Toggled from 0 to 1	Total
affricative	5212	87	5299	5212	87	0	5299	5255	26	18	5299
alveodental	3010	2289	5299	4503	402	394	5299	4917	200	182	5299
alveoplatal	5102	197	5299	5135	158	6	5299	5223	49	27	5299
anterior	5128	171	5299	5128	169	2	5299	5218	47	34	5299
aspirated	5046	253	5299	5055	235	9	5299	5150	97	52	5299
bilabial	1574	3725	5299	5117	41	141	5299	5242	17	40	5299
consonant	603	4696	5299	4912	174	213	5299	5120	87	92	5299
continuant	4985	314	5299	4985	314	0	5299	5085	103	111	5299
coronal	2703	2596	5299	4771	332	196	5299	5082	132	85	5299
emphatic	5112	187	5299	5112	187	0	5299	5202	78	19	5299
fricative	5037	262	5299	5037	262	0	5299	5117	137	45	5299
glottal	5212	87	5299	5212	87	0	5299	5236	44	19	5299
high	5216	83	5299	5220	75	4	5299	5275	19	5	5299
interdental	5195	104	5299	5195	104	0	5299	5237	52	10	5299
labiodental	5124	175	5299	5146	144	9	5299	5239	38	22	5299
labiovelar	5219	80	5299	5224	74	1	5299	5275	16	8	5299
lateral	5110	189	5299	5182	92	25	5299	5264	20	15	5299
nasal	4726	573	5299	4855	214	230	5299	5117	73	109	5299
palatal	4767	532	5299	5085	141	73	5299	5222	34	43	5299
pharyngeal	5136	163	5299	5138	160	1	5299	5254	29	16	5299
plosive	5212	87	5299	5212	87	0	5299	5246	41	12	5299
rounded	4309	990	5299	4835	278	186	5299	5174	61	64	5299
semivowel	5022	277	5299	5025	272	2	5299	5179	82	38	5299
short	5232	67	5299	5232	67	0	5299	5261	15	23	5299
trill	1177	4122	5299	4883	164	252	5299	5170	69	60	5299
unvoiced	3888	1411	5299	5211	51	37	5299	5275	15	9	5299
uvular	1457	3842	5299	4606	131	562	5299	5011	126	162	5299
velar	2178	3121	5299	4412	312	575	5299	4919	195	185	5299
voiced	4708	591	5299	4723	545	31	5299	5020	133	146	5299
vowel	3940	1359	5299	5198	40	61	5299	5241	17	41	5299

TABLE 8. Overall accuracy performances of the DBN and CRNN (%).

	DBN	CRNN
DPF	51.33*	78.68*
	79.61**	92.53**
Phonemes	81.75	84.02

***Hamming distance = 0, ** Hamming distance = 3**

As shown in Table 5, CRNN was generally more accurate than DBN, but the vs10, ds10, zs10, tb10, gs10, ks10, hs10, and us21 phonemes were more accurately classified by DBN. When the poorly recognized phonemes in the CRNN classifier intersected with the poorly recognized DPF elements, the outcomes were usually ones (i.e., plusses) (see Table 9 and similar cases for the DBN classifier in Table 10).

In other words, when CRNN performed poorly, the "+" symbols in the DPF tables were reversed to "−" symbols. The misclassified and erroneous DPF elements degraded the accuracy of classifying the corresponding phonemes. The inverse situation also appeared: the most accurately classified phonemes and DPF elements intersected as "−" symbols in Table 9. This result is consolidated in other tables and figures; for example, Table 7 and

TABLE 9. The best and worst DPF and phoneme classifications by DBN.

DPF		Phonemes												
		Worst (< 60%)						Best (≥ 90%)						
		bs10	**db10**	**tb10**	**zb10**	**ds10**	**ss10**	hb10	zs10	js10	as10	us10	is10	us20
Worst	**Alveodental**	0	1	1	0	1	1	0	1	0	0	0	0	0
	Coronal	0	1	1	1	1	1	0	1	1	1	0	0	0
	Uvular	0	0	0	0	0	0	0	0	0	0	0	0	0
	Velar	0	0	0	0	0	0	0	0	0	0	0	0	0
	Voiced	1	1	0	1	1	0	0	1	0	1	1	1	1
Best	**High**	0	1	1	1	0	0	0	0	0	0	0	0	0
	Labiovelar	0	0	0	0	0	0	0	0	0	0	0	0	0
	Unvoiced	0	0	1	0	0	1	1	0	1	0	0	0	0

TABLE 10. The best and worst DPF and phoneme classifications by DBN.

DPF		Phonemes												
		Worst (< 60%)									Best(≥ 90%)			
		db10	**us10**	**bs10**	**ss10**	**is10**	**sb10**	**zb10**	**vb10**	**ms10**	zs10	as10	is10	us20
Worst	**Alveodental**	1	0	0	1	0	1	0	0	0	1	0	0	0
	Coronal	1	0	0	1	0	1	1	1	0	1	1	0	0
	Uvular	0	0	0	0	0	0	0	0	0	0	0	0	0
	Velar	0	0	0	0	0	0	0	0	0	0	0	0	0
	Voiced	1	1	1	0	1	0	1	1	1	1	1	1	1
Best	**Affricative**	0	0	0	0	0	0	0	0	0	0	0	0	0
	High	1	0	0	0	0	1	1	0	0	0	0	0	0
	Labiovelar	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4 show more erroneous 1-to-0 toggles than 1-to-0 toggles.

In general, the CRNN outperformed the DBN in DPF-elements extraction and phoneme classification. The performance was measured by the numbers of correct matches between the extracted elements and the reference entries of the lookup table. Both extractors were fed with the test subset of KAPD (Table 8). To our knowledge, we present the first attempt to combine the spectrogram as a feature of DNNs in Arabic DPF modeling, DPF extraction, and phoneme classification. In our previous work [9], we examined several combinations of acoustic cues, and determined which combination can efficiently represent a DPF elements. The responsiveness of the acoustic cues was assessed from the MLP performance. Among the top-level extractors, the DNN-based extractor outran the MLP-based extractor.

V. CONCLUSION

This study examined the advantages of deep learning in the DPF modeling and extraction of Arabic language phonemes. Experiments were performed on 30 DPF elements of MSA. Two models (one based on DBN, the other on CRNN) were designed for extracting the DPF elements and classifying the phonemes. The extraction and classification tasks were experimentally assessed on spectrogram data. Finally, DPF-vector extraction was applied to the resulting models. For this task, two extractors were developed. The detailed phoneme-matching rates demonstrated the higher effectives of the CRNN extractor than the DBN extractor. Beside achieving a lower error rate in general, the CRNN extractor generated fewer confusion errors than the DBN extractor, and more robustly generated error-free vectors. The CRNN is widely applied in digital speech processing, and the present study demonstrated its additional advantage in acoustic-to-phonetic conversion. The study further confirmed the representativeness of the spectrogram cues as DPF elements.

The present results have improved our understanding of DPFs and lay a foundation for more advanced applications in this context. The high-performance DPF extractors proposed in this work provide promising perspectives towards the effective integration of phonetic features thereby solving many problems in Arabic digital speech processing.

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, pp. 1426–1439, 2013, doi: 10.3906/elk-1112-29.

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

[3] P. L. Garvin, R. Jakobson, C. Gunnar, M. Fant, and M. Halle, "Preliminaries to speech analysis: The distinctive features and their correlates," *Language*, vol. 29, no. 4, p. 472, Oct. 1953, doi: 10.2307/409957.

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

[5] A. Wentlent, *Alfred'S IPA Made Easy: A Guidebook for the International Phonetic Alphabet*. Los Angeles, CA, USA: Alfred Publishing Company, 2014.

[6] M. M. Alghamdi, "KACST arabic phonetics database," *Congr. Phonetics Sci.*, vol. 15, no. 1, pp. 7–10, 2003.

[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, 2003, pp. 25–28, doi: 10.1109/icassp.2003.1202285.

[8] S. M. R. Kabir, F. Hassan, F. Ahamed, K. Mamun, M. N. Huda, and F. Nusrat, "Phonetic features enhancement for Bangla automatic speech recognition," in *Proc. Int. Conf. Comput. Inf. Eng. (ICCIE)*, Nov. 2015, pp. 25–28, doi: 10.1109/ccie.2015.7399309.

[9] 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, doi: 10.1109/ACCESS.2019.2924014.

[10] A. B. Ibrahim, Y. M. Seddiq, A. H. Meftah, M. Alghamdi, S.-A. Selouani, M. A. Qamhan, Y. A. Alotaibi, and S. A. Alshebeili, "Optimizing arabic speech distinctive phonetic features and phoneme recognition using genetic algorithm," *IEEE Access*, vol. 8, pp. 200395–200411, 2020, doi: 10.1109/ACCESS.2020.3034762.

[11] M. Algabri, H. Mathkour, M. M. Alsulaiman, and M. A. Bencherif, "Deep learning-based detection of articulatory features in arabic and english speech," *Sensors*, vol. 21, no. 4, pp. 1–23, 2021, doi: 10.3390/s21041205.

[12] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8599–8603, doi: 10.1109/ICASSP.2013.6639344.

[13] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[14] A.-R. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5060–5063, doi: 10.1109/ICASSP.2011.5947494.

[15] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.

[16] A. Fischer and C. Igel, "An introduction to restricted Boltzmann machines," in *Proc. Iberoamerican Congr. Pattern Recognit.*, in Lecture Notes in Computer Science, vol. 7441, 2012, pp. 14–36, doi: 10.1007/978-3-642-33275-3_2.

[17] A.-R. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4273–4276, doi: 10.1109/ICASSP.2012.6288863.

[18] A.-R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," *Scholarpedia*, vol. 4, no. 5, pp. 1–9, 2009, doi: 10.4249/scholarpedia.5947.

[19] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012, doi: 10.1109/TASL.2011.2109382.

[20] G. E. Dahl, M. Ranzato, A. R. Mohamed, and G. Hinton, "Phone recognition with the mean-covariance restricted Boltzmann machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 469–477.

[21] R. Sarikaya, G. E. Hinton, and B. Ramabhadran, "Deep belief nets for natural language call-routing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5680–5683, doi: 10.1109/ICASSP.2011.5947649.

[22] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted Boltzmann machines," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5884–5887, doi: 10.1109/ICASSP.2011.5947700.

[23] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 90, 2013, pp. 42–51, doi: 10.1016/j.ultras.2018.05.007.

[24] Ł. Brocki and K. Marasek, "Deep belief neural networks and bidirectional long-short term memory hybrid for speech recognition," *Arch. Acoust.*, vol. 40, no. 2, pp. 191–195, Jun. 2015, doi: 10.1515/aoa-2015-0021.

[25] M. Alghamdi, *Arabic Phonetics*. Riyadh, Saudi Arabia: A1-Toubah Bookshop, 2001.

[26] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York, NY, USA: IEEE, 2000.

[27] M. Abadi *et al.* (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. Accessed: Nov. 20, 2020. [Online]. Available: http://download.tensorflow.org/paper/whitepaper2015.pdf

[28] A. Gulli and S. Pal,*Deep Learning With Keras*. Birmingham, U.K.: Packt, 2017.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

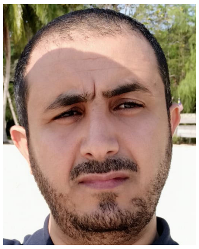
[30] F. J. Valverde-Albacete and C. Peláez-Moreno, "100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox," *PLoS ONE*, vol. 9, no. 1, 2014, Art. no. e84217, doi: 10.1371/journal.pone.0084217.

[31] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013, doi: 10.1016/j.ins.2013.07.007.

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 Researcher Assistant with King Saud University. His main research interests include digital signal processing, speech processing, and artificial intelligence.

YOUSEF AJAMI ALOTAIBI (Senior Member, IEEE) received the B.Sc. degree from King Saud University, Riyadh, Saudi Arabia, in 1988, and the M.Sc. and Ph.D. degrees from the Florida Institute of Technology, USA, in 1994 and 1997, respectively, all in computer engineering. From 1988 to 1992 and from 1998 to 1999, he joined Al-ELM Research and Development Corporation, Riyadh, as a Research Engineer. From 1999 to 2008, he joined as 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. His research interests include digital speech processing, specifically speech recognition and Arabic language, and speech processing.

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 the 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 at 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, where he is also a Researcher Assistant. His research interests include digital speech processing, specifically speech recognition and Arabic language, and speech processing and artificial intelligence.

SID AHMED SELOUANI (Senior Member, IEEE) received the B.E. and M.S. degrees in electronics engineering and the D.Sc. degree in speech processing from the University of Science and Technology Houari Boumediene (USTHB), in 1987, 1991, and 2000, respectively. In 2000, he joined the Algerian-French Double Degree Program, Université Joseph Fourier of Grenoble. He is currently a Full Professor with the Université de Moncton, Shippagan, where he is also the Founder of the Laboratory of Research in Human–System Interaction. He is an Invited Professor with INRS-Telecommunications, Montreal, QC, Canada. His main areas of research interests include artificial intelligence, human–computer interaction, speech recognition robustness, speaker adaptation, speech processing using soft computing and deep learning, dialog systems, ubiquitous systems, intelligent agents, and speech enhancement.