IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 6, 2021, accepted April 9, 2021, date of publication April 16, 2021, date of current version April 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3073786

Digital Audio Forensics: Microphone and Environment Classification Using Deep Learning

MUSTAFA A. QAMHAN, HAMDI ALTAHERI, (Member, IEEE), ALI HAMID MEFTAH, GHULAM MUHAMMAD, (Senior Member, IEEE), AND YOUSEF AJAMI ALOTAIBI, (Senior Member, IEEE)

Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Ghulam Muhammad (ghulam@ksu.edu.sa)

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

ABSTRACT The recording device along with the acoustic environment plays a major role in digital audio forensics. We propose an acoustic source identification system in this paper, which includes identifying both the recording device and the environment in which it was recorded. A hybrid Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) is used in this study to automatically extract environments and microphone features from the speech sound. In the experiments, we investigated the effect of using the voiced and unvoiced segments of speech on the accuracy of the environment and microphone classification. We also studied the effect of background noise on microphone classification in 3 different environments, i.e., very quiet, quiet, and noisy. The proposed system utilizes a subset of the KSU-DB corpus containing 3 environments, 4 classes of recording devices, 136 speakers (68 males and 68 females), and 3600 recordings of words, sentences, and continuous speech. This research combines the advantages of both CNN and RNN (in particular bidirectional LSTM) models, called CRNN. The speech signals were represented as a spectrogram and were fed to the CRNN model as 2D images. The proposed method achieved accuracies of 98% and 98.57% for environment and microphone classification, respectively, using unvoiced speech segments.

INDEX TERMS Acoustic environments, microphones, classification, digital audio, forensics, deep learning, CNN, LSTM, Arabic speech.

I. INTRODUCTION

Forensics refers to the science that uses scientific methods or expertise to investigate crimes or examine an evidence which may be presented in a court of law. Digital media forensics is a branch of forensic science that involves ensuring that the digital content is accurate and authentic [1]. It focuses on analyzing the evidence to identify any manipulations and counterfeiting, and proving the integrity and authenticity of the digital information, as well as its sources.

Using digital information as evidence in criminal investigations has become very popular in the court. Hence, proving the authenticity and integrity of digital media is important for its consideration as evidence in the court of law. Digital media can be represented in different forms, such as text, audio, video, and image. Much research has been carried out in the field of image forensics [2]. However, the field of audio forensics is comparatively less advanced.

Digital audio forensic includes different activities, such as identifying speakers from the audio, identifying the environment or the recording device, and checking the integrity of the audio content. The methods for authentication of digital audio can be broadly divided into container-based and content-based authentication. Container-based authentication analyzes the description of the audio file and its structure, such as its file format, MAC times (i.e., Modification/Access/Creation times of a file), and hash (i.e., a unique fingerprint of content, e.g., file) analysis. Content-based authentication analyzes the actual content of the audio recording, including the Electric Network Frequency (ENF) analysis, acquisition device, and environment identification.

This paper focuses on acoustic source identification for the purpose of audio authentication which includes identifying both the recording device and the environment in which it

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

TABLE 1. A summary of studies in environment classification.

Study	Year	Features	Classifier	No of Classes	Dataset	Accuracy %
Mushtaq et al. [46]	2020	Mel spectrogram, MFCC, Log-Mel	DCNN	50 10 10	*ESC-50* [47] *ESC-10* [47] *UrbanSound8K* [48]	94.9 89.3 95.4
Huang et al. [49]	2020	MFCC and GFCC	2-DenseNet	15 10	DCASE 2016 [50] *UrbanSound8K* [48]	84.8 85.2
Patole et al. [19]	2019	Decay Rate, and MFCCs, RT	SVM, KNN, LDA,QDA, Ensemble, and ANN	7	**Dataset 1**: 210 sine sweep signal of 25 seconds duration. 7 environment classes. **Dataset 2**: Two anechoic speech of the same speaker. 7 environment.	DS1: 98.7 DS2: 99.5
Narkhede et al. [24]	2019	MFCC, STE, spectral roll-off, SC, and SF	SVM, ANN	7	One speaker. 7 environments. 5 recordings per environment.	90.9
Hossain et al. [31]	2018	Mel-spectrogram	CNN	8	300 GB of audio clips including environmental sound and/or human speech	95 (without human speech) 90 (with human speech)
Jleed et al. [51]	2017	DHT	HMM	10 15	DCASE 2013 [52] DCASE 2016 [50]	D-2013: 77 D-2016: 75.6
Ali et al. [53]	2017	Critical band spectral	GMM	3	King Saud University Arabic Speech Database (KSU-ASD) [3].	99.2
Boddapati et al. [25]	2017	Spectrogram, MFCC, CRP	CNN	50 10 10	*ESC-50* [47] *ESC-10* [47] *UrbanSound8K* [48]	86 73 93
Mafra et al. [54]	2016	Averaged Mel-log spectrogram	Linear SVM	10	DCASE 2013 [52].	75
Muhammad et al. [55]	2011	MFCCs and MPEG-7	HMM	10	200 recordings, 30 seconds per record, 10 environments	above 80 % for all environments
Kraetzer et al. [13]	2007	Time domain and Mel-Cepstral	Naive Bayes	10	400 audio files, 4 microphones, 10 different rooms.	41.5

MFCC: Mel-frequency Cepstral coefficients, GFCC:Gammatone frequency cepstral coefficients, DCNN: deep convolutional neural networks, 2-DenseNet: 2-order dense convolutional network, RT: reverberation time, STE: short time energy, SC: spectral centroid, SF: spectral flux, SVM: support vector machine, KNN: k-nearest neighbors, HMM: hidden Markov model, LDA: linear discriminant analysis, QDA: quadratic discriminant analysis, ANN: artificial neural network, CNN: convolutional neural network. GMM: Gaussian mixture model, DHT: Discrete Hartley transform.

was recorded. It has been shown that the acquisition system, encoding process, and acoustic environment create artifacts in the recorded audio, such as acoustic reverberations, background noise, and device-related noise. Such artifacts can be used to identify the acquisition device and acoustic environment and then verify the authenticity and integrity of the digital audio. Many techniques have been proposed to extract these features and utilize them for microphone and environment classification. A broad categorization of different digital audio forensic techniques has been explored in this respect [1].

Our goal in this paper is to classify the environments and the microphones using spectrogram features with two deep neural networks, namely Convolutional Neural Network (CNN) and Convolutional Recurrent Neural Network (CRNN), i.e. A hybrid CNN and LSTM. Furthermore, we investigated the effect of using voiced and unvoiced segments of the speech on the accuracy of environment and microphone classification. We also studied the effect of background noise on microphone classification in different environments and the effect of microphone quality on the environment classification. A public real-world speech corpus, namely KSU-DB [3], was used in this study. The selected dataset contains 3 environments, 4 classes of recording devices, 136 speakers (68 males, 68 females), and 3600 recordings of words, sentences, and continuous speech.

The main contribution of this paper is to propose a hybrid CNN and LSTM architecture with few parameters, which performed excellent and robust classification accuracy with generalization ability regardless of environmental noise or recording device quality. This research comprehensively studied the problem of microphones classification and environment classification from different perspectives. The effects of speech segments, speaker gender, background

TABLE 2. Summary of studies in microphone classification.

Study	Year	Features	Classifier	No of Classes	Dataset	Accuracy %
Baldini et al. [4]	2020	Different entropy measures	SVM, KNN, Decision Tree	34	34 mobile microphones stimulated by three different types of audio recordings	98
Lin et al. [5]	2020	Audio Spectrogram	CNN Model With Subband Attention	20	20 models of cell-phones made by 5 major manufacturers	95
Baldini et al. [6]	2019	Audio spectrum	CNN	34	Microphone responses from 34 different mobile phones (4 Models/ manufacturers)	80
Jiang et al. [7]	2019	MFCC	GSV, SVM, SRC	4	Ahumada-25 [8]. 25 speakers. 4 different microphones	89.04
Eskidere et al. [9]	2015	Multitaper MFCC	GMM	16	40 speakers, 16 microphones. 120. One room.	99.27
Eskidere et al. [10]	2014	LPCC, PLPC, MFCC	GMM	16	40 speakers, 16 microphones. 120. One room.	99.58
Aggarwal et al. [11]	2014	MFCC	SVM with SMO	26	26 cell phones (5 different manufacturers)	90
Hanilçi et al. [12]	2014	MFCC, LFCC	GMM and SVMs	14	14 cell phones	98.39
Kraetzer et al. [13]	2007	Time domain and Mel-Cepstral	Naive Bayes	4	400 audio files, 4 microphones, 10 different rooms.	75.99

MFCC: Mel-frequency cepstral coefficients, LPCC: Linear Prediction Cepstral Coefficient, PLPC perceptually based linear predictive coefficient, linear frequency cepstral coefficients (LFCC), SVM: support vector machine, KNN:K Nearest Neighbor, CNN: convolutional neural network. GMM: Gaussian mixture model, SRC: Sparse Representation based Classifier, GSV Gaussian Supervector, SMO: Sequential minimal optimization.

noise, and acquisition device quality are investigated to answer questions about the relationships between these factors and microphone and environment classification. Based on our knowledge, there is no previous study that used the proposed model to classify environments and the microphones using the KSU-DB corpus. In addition, this is the first study where environment classification is performed taking into consideration the quality of microphones, and microphones classification is performed taking into consideration the environment (background noise).

The rest of this paper is structured according to the following. The literature review and the chosen speech corpora are presented in Sections II and III respectively. Data preparation and our experimental work are defined in Section IV. Sections V and VI address the proposed models and experiments, respectively. The main findings are summarized in Sections VII and, finally, we present our conclusions in Section VIII.

II. LITERATURE REVIEW

There have been several bio-inspired techniques used for voice analysis in the literature. A linear predictive coding-based formant analysis was done for automatic speech recognition of dysphonic patients in [14]. For dysphonic patients, the frequency of voiced sounds differs from that of normal persons because of biological distortion in vocal folds. The frequency change was thoroughly studied in [15]. Auditory spectrum using the all-pole model was studied for voice and speech analysis in [16]. A bio-inspired algorithm-based vocal tract irregularity measurement was proposed to quantify the voice pathology in [17].

The ENF signal, which results from the combination of the digital recording system and the power line frequency, has been used in some studies for digital recording authentication [18]. The ENF-based method is one of the most reliable audio forensic approaches; however, it is not applicable in some cases, like when the audio equipment is battery-operated [1]. The study in [13] performed microphone and environment classification using the Naïve Bayes classifier. Although the accuracy of the classification was not high, the work showed the ability to identify the acquisition device and environment based on the captured audio. After that, different features and classifiers have been proposed to represent the information of acquisition device and environment and to classify them.

A digital audio recording typically consists of a direct speech signal, indirect or reflected signals (also known as reverberations), background noises, and acquisition device noises. Reverberations and background noises are used to describe the recording's acoustic environment, while acquisition device-related noises are used to identify the recording microphone. Various studies use reverberation in the speech for audio authentication and environment identification [19]–[22]. The study in [20] extracted reverberation

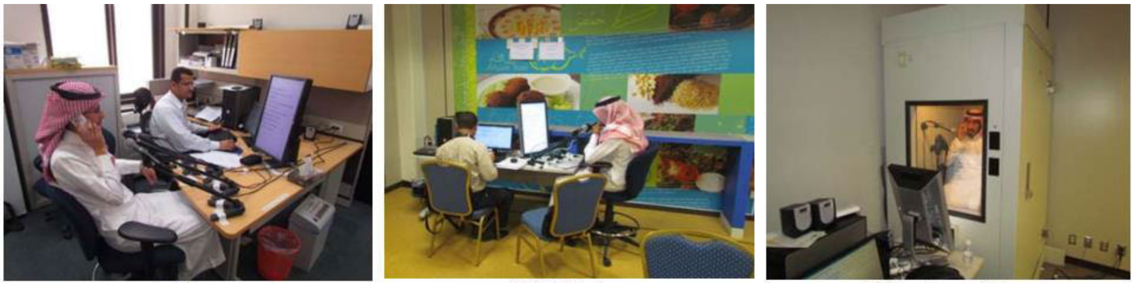

a) Office (b) Cafeteria (c) Sound-proof Room

FIGURE 1. The three environments used to record KSU-DB dataset.

information using decaying tail and temporal peaks, and used it for acoustic scene classification. In [21], reverberation was estimated based on the spectral subtraction and inverse filtering and was used for background noise estimation.

Speech recording is typically divided into a sequence of frames to obtain a sequence of frame-level features. The frame-level features are then concatenated to form a single feature vector for the recording audio. The Mel-frequency Cepstral Coefficients (MFCCs) is one of the frame-level features commonly used in speaker verification and speech recognition [23]. Several studies employed MFCC for microphone and environment classification [7], [19], [24], [25]. Several other frame-level features have also been evaluated for microphone recognition, such as Multi-taper MFCC [9], or Linear Prediction Cepstral Coefficient (LPCC) [10]. The researchers in [10] classified 16 different microphones recorded in one silent room using a Gaussian Mixture Model (GMM) classifier. They studied 3 feature representations: LPCC, MFCC, and Perceptually-based Linear Predictive Coefficients (PLPCs). They showed that the LPCC features outperform the other features (MFCC and PLPC). The audio spectrum is used directly instead of LPCC or MFCC in some studies for telephone identification, such as Sketches of Spectral Features (SSFs) [26], Random Spectral Features (RSFs) [27], and Labeled Spectral Features (LSFs) [28]. The signal spectrogram is another representation of audio signals that is used in many studies ranging from phoneme identification [29] to speech and speaker recognition [30]. The spectrogram of the audio signal was used in [25] for environmental sound classification. The authors divided the speech sound spectrogram into 2D images and classified them using a CNN. The study in [31] used Mel-spectrogram features and CNN for urban environment classification, which has also been used for noise cancellation in smartphones. Tables 1 and 2 present a summary of studies based on environment and microphone classification.

Speech information, which interferes with acoustic source information (i.e., device and environment), is almost useless for microphone and environment recognition. However, distinguishing the source information from the speech information is a difficult task, since neither one is defined in advance. Thus, some researchers attempted to extract features from the near-silence segments of the audio signal [32], noise signal [11], or non-speech segments [12]. These semi-silence signals, however, are unstable because they are quite similar to noise and are not sufficient to train a strong classifier when the audio signal is filled with speech information.

CNNs have achieved outstanding success in extracting local and spatial features and patterns directly from raw data, such as images [33], videos [34], and speech [30]. RNNs can extract temporal features and patterns from time-series data, which makes them useful in video and speech applications. We utilized CNN in this study along with RNN to automatically extract environments and microphone features from the speech sound. This research combines the advantages of both CNN and RNN models.

III. SELECTED SPEECH CORPUS

The King Saud University speech database (KSU-DB) [35] is an Arabic language database of 91879 speech sounds recorded by 257 male and female speakers from 29 Arab and non-Arab countries in three recording sessions. The database was recorded using a sampling rate of 48 kHz with a resolution of 16-bits. Different forms of text were chosen for recording this corpus such as numbers, individual words, sentences, paragraphs, phonetically balanced sentences, phonetically rich words, and responses to questions. The maximum duration of each record is 120 seconds. Figure 1 shows the three different environments where the database was recorded; a soundproof room representing a quiet environment, an office room representing a low noise environment, and a cafeteria representing a noisy environment. The KSU-DB was recorded using different types of microphones; high-quality microphones, medium-quality microphones, and a mobile, as shown in Figure 2. In order to track inter-session differences of the speakers, the database was recorded in 3 sessions with a delay of around 6 weeks. The richness of this corpus makes it

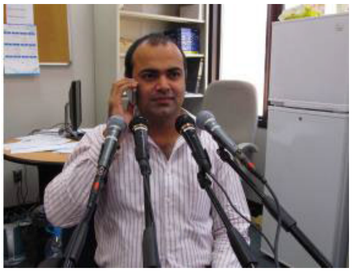


FIGURE 2. Different channels used to record KSU-DB dataset.

TABLE 3. Distribution of the selected KSU_DB dataset.

		Environment						
		Silent Room (very quiet)		Office (quiet)		Cafeteria (noisy)		
		females	males	females	males	females	males	
Recording device	High-quality	200	200	150	150	150	150	**1000**
	Medium-quality-1	0	200	150	150	150	150	**800**
	Medium-quality-2	200	0	150	150	150	150	**800**
	Mobile-Mic	200	200	150	150	150	150	**1000**
		600	600	600	600	600	600	**3600**
		1200		**1200**		**1200**		

suitable for many speech processing investigations, including microphone and environment classification.

IV. DATA PREPARATION AND EXPERIMENTAL WORK

A. DATA PREPARATION

We selected a subset of the KSU-DB corpus containing 3 environments, 4 classes of recording devices, 136 speakers (68 males and 68 females), and 3600 recordings of words, sentences, and continuous speech in this investigation. The rerecorded files were distributed equally between the speakers, the microphones, and the environments, as demonstrated in Table 3. The three types of environments were silent room (very quiet), office room (quiet), and cafeteria (noisy). For the recording systems, the four classes of acquisition devices were [35]:

- A Yamaha_Mixer which we will refer to as "High-quality" in this paper, is comprised of 2 professional microphones (SHURE, Beta 58A, Chicago, United States) connected to a high-quality mixer (Yamaha, MW12CX, Hamamatsu, Japan).
- A Mic_CreativeSB, which we will refer to as "medium-quality-1" in this paper, is a medium-quality microphone (Sony, F-V220, Tokyo, Japan) connected to a sound card (Creative, Creative 5.1 Surrounding Jurong East, Singapore).
- A Computer_Mic_Front, which we will refer to as "medium-quality-2" in this paper, is also a medium-quality microphone (Sony, F-V220) connected directly to a computer (without the external sound card).
- A Mobile_CreativeSB, which we will refer to as "Mobile Mic" in this paper, is a mobile (Nokia, N97, Espoo, Finland) connected to a sound card (Creative Surrounding 5.1).

Table 3 presents a full picture of the selected files for the environments and the microphones with the respective distribution of males and females.

In the experiments, first, we will perform a human perceptual test to study human performance on the environment and microphone classification as a baseline for our system.

Second, we will investigate the effect of voiced and unvoiced phonemes and the speakers' gender for the classification of microphones and environments. Third, the effect of environmental background noise on microphone classification will be investigated. Forth, the quality of recording devices in environment classification will be studied. The following experiments will be performed to achieve these targets, as shown in Figure 3:

1) ENVIRONMENT CLASSIFICATION

Spectrograms will be extracted from 3600 audio files representing the 3 environment types in this experiment, regardless of the type of microphone used.

2) MICROPHONE CLASSIFICATION

Spectrograms will be extracted from 3600 audio files recorded with the 4 types of microphones in this experiment, regardless of the recording environment.

3) MICROPHONE CLASSIFICATION IN A VERY QUIET ENVIRONMENT

Spectrograms will be extracted from all 1200 audio files recorded in a silent room for this experiment.

4) MICROPHONE CLASSIFICATION IN THE QUIET ENVIRONMENT

Spectrograms will be extracted from all 1200 audio files recorded in an office room for this experiment.

5) MICROPHONE CLASSIFICATION IN THE NOISY ENVIRONMENT

Spectrograms will be extracted from all 1200 audio files recorded in a cafeteria room for this experiment.

6) MICROPHONE CLASSIFICATION IN EACH ENVIRONMENT SIMULTANEOUSLY

Spectrograms will be extracted from all 3600 audio files recorded in all environments at the same time for this experiment.

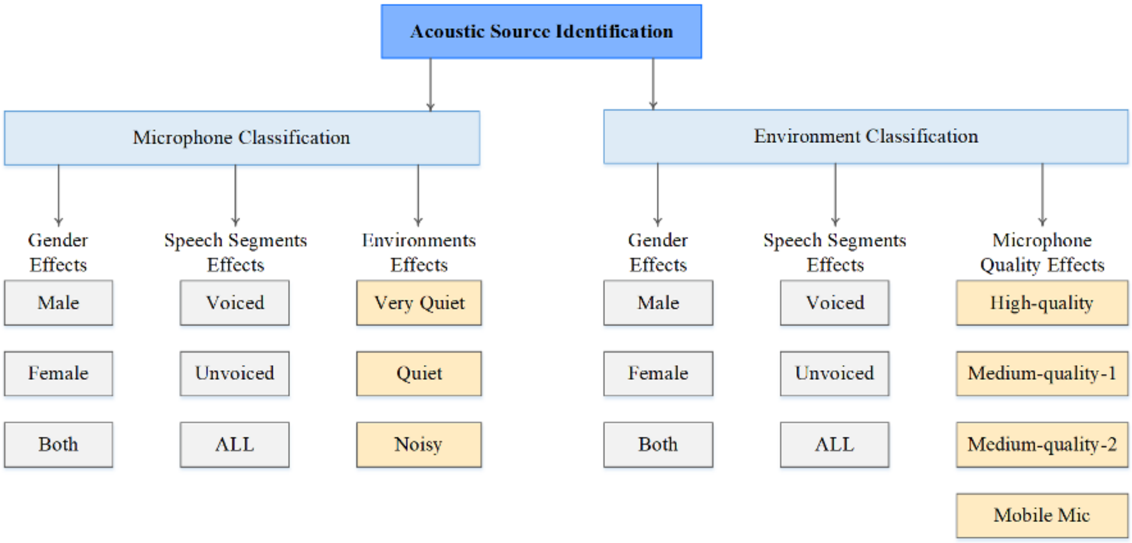


FIGURE 3. **The experiments performed in this study (Very quiet: silent room, Quiet: office, Noisy: Cafeteria, High-quality: Yamaha_Mixer, Medium-quality-1: Mic_CreativeSB, Medium-quality-2: A Computer_Mic_Front, and Mobile Mic: Mobile_CreativeSB).**

7) ENVIRONMENT CLASSIFICATION USING RECORDING DEVICES WITH DIFFERENT QUALITIES

Spectrograms will be extracted from 3600 audio files recorded with the 4 classes of recording devices for this experiment.

B. EXPERIMENTAL WORK

1) HUMAN PERCEPTUAL TEST

The human perceptual test was performed on a subset of 300 audio files randomly selected from the dataset, 30 audio files per subject. The number of volunteers for the classification task was 10, ranging in the age of 18 to 40 years. Each of them was given a number of 120 labeled files for training, which contained the three different environments and four types of recording devices. They were asked to listen to a sufficient number of files to be able to differentiate between the different categories. After that, they were asked to classify 30 audio files that were randomly selected from the testing dataset. Different audio files were given to each subject.

2) SPECTROGRAMS (SELECTED FEATURES)

A spectrogram is a visual representation of sound. It is commonly depicted as an image with two dimensions representing frequency and time on the vertical and horizontal axes, respectively. The color intensity of a spectrogram represents the signal amplitude of a given frequency at a specific time. The light blue color represents the lowest amplitude, with brighter colors indicating higher amplitudes, and the highest amplitude is represented by dark red. A Short-Time Fast Fourier Transform (FFT) was used to generate speech signal spectrogram. The method to generate a spectrogram for a speech signal at a given point is described as follows. First, a speech signal was divided into a sequence of frames (30 ms per frame in this study). Next, the Hamming window was multiplied by each frame. The spectra of the windowed frames were then generated by applying FFT according to (1)

$$X_k = \sum_{n=0}^{N-1} w_n x_n e^{-i2\pi kn/N} \quad k = 0, \ldots, N-1 \tag{1}$$

where w_n is the Hamming window function, x_n is the original speech signal, and k refers to the frequency. The log-power representation can be calculated from a power spectrogram by applying (2)

$$S_k = 10 \log |X_k| \tag{2}$$

Figure 4 shows sample spectrograms (corresponding to different microphones in different environments) generated by using the described method.

3) VOICED AND UNVOICED SEGMENTATION

Speech signals can be separated into several voiced and unvoiced segments to provide preliminary acoustic segmentation for different speech processing applications, such as speech recognition, speech synthesis, and speech enhancement. Approximately two-thirds of speech is voiced and this type of speech is also what is the most important for characterizing intelligibility [36]. A speech signal is composed of three segments: unvoiced (U), voiced (V), and silent (S). Classifying speech into V/U/S segments is a fundamental process for many speech processing tasks, such as speaker identification, speech synthesis, and speech recognition. Various approaches have been proposed to segment a speech signal into V/U/S using energy, zero crossings, or pitch [37]. We used pitch information in this investigation, which depends on the vibration frequency in the vocal folds, for classifying the speech signal into V/U segments, as shown in Figure 5. Framing with a frame size of 30 ms and overlapping with a step size of 10 ms were implemented in all the speech files.

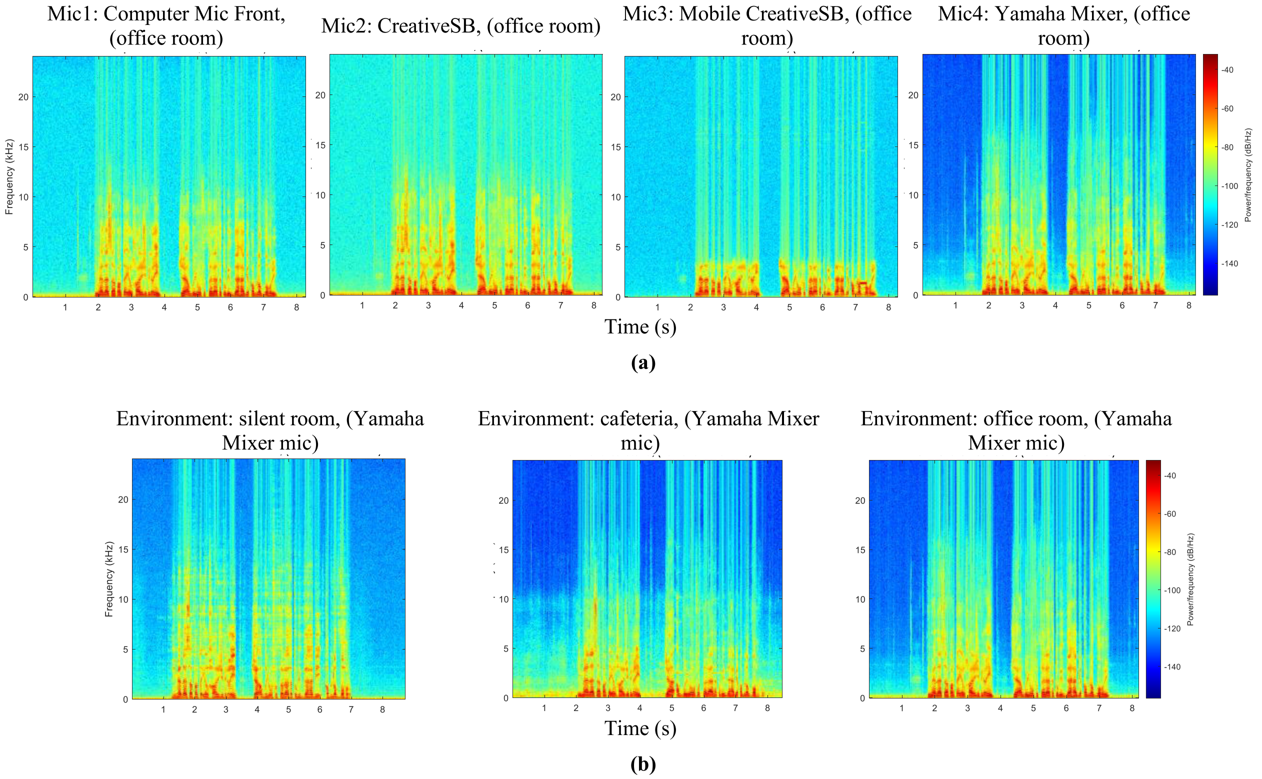


FIGURE 4. **Spectrogram examples of one speaker using the same sentence, a) recorded in one environment (office) using four different recording devices, b) recorded using one type of recording devices (Yamaha mixer) in three different environments.**

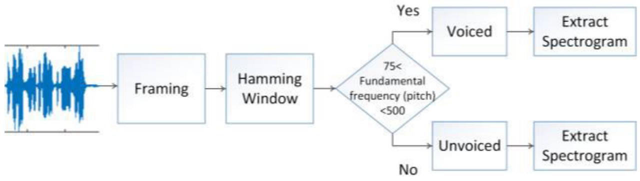


FIGURE 5. **Spectrogram generation.**

The spectrograms obtained from the audio speech files based on the above method contain a variable number of frames. To solve this problem, spectrograms were divided into segments of fixed-size, which were used for the proposed models as inputs. The number of frames for the last segments varied between audio speech files. If the last segment frame size exceeded 50% of the default segment, a sufficient number of frames from the same file could fill the remaining portion of the segment. The segment was otherwise ignored. Equations (3) and (4) shows the procedure for the segmentation process, where F_x refers to an audio file x, N_s is the number of segments in the file F_x, N_{F_x} is the number of frames in the file F_x, S_s refers to the size of the segment frames, and P_f refers to the size of the padding frames. The result of this process is N_s segments with fixed-size frames from each of the audio speech files.

$$N_s = round\left(\frac{N_{F_x}}{S_s}\right) \tag{3}$$

$$P_f = max(0, N_{F_x} - N_s S_s) \tag{4}$$

V. PROPOSED CNN AND CRNN MODELS

CNN's design is a collection of neural networks organized of different-size layers in a certain sequence, in which each layer makes a specific contribution. The former levels are less profitable and the deeper layers are more advanced features, such as the speaker in a dialogue, or the target in a photograph. A traditional CNN model comprises of many structural building blocks such as convolution layers and pooling layers [38]. A convolutional layer is an integral part of the infrastructure of CNN that extracts features. A pooling layer gives the network computing reduced by traditional downsampling activity. The output feature maps of the last pooling (or convolution) layer are typically flattened and connected to one or more fully connected layers.

The convolution layer contains many filters (kernels) that are convolved across the inputs from previous layers. The convolution process can be expressed by the following equation:

$$y_{i'j'} = \sum_{i,j=0}^{n} w_{ij} x_{i+i',j+j'} \tag{5}$$

where $y_{i'j'}$ refers to the output feature map computed from the $i'j'$ position in the input matrix x_{ij}, w_{ij} is the kernel matrix variable, $x_{i+i',j+j'}$ is the spatial area element from the input, i, j denotes the row and column for the kernel's current elements, and n denotes the kernel's elements. The used activation function in this analysis is exponential linear

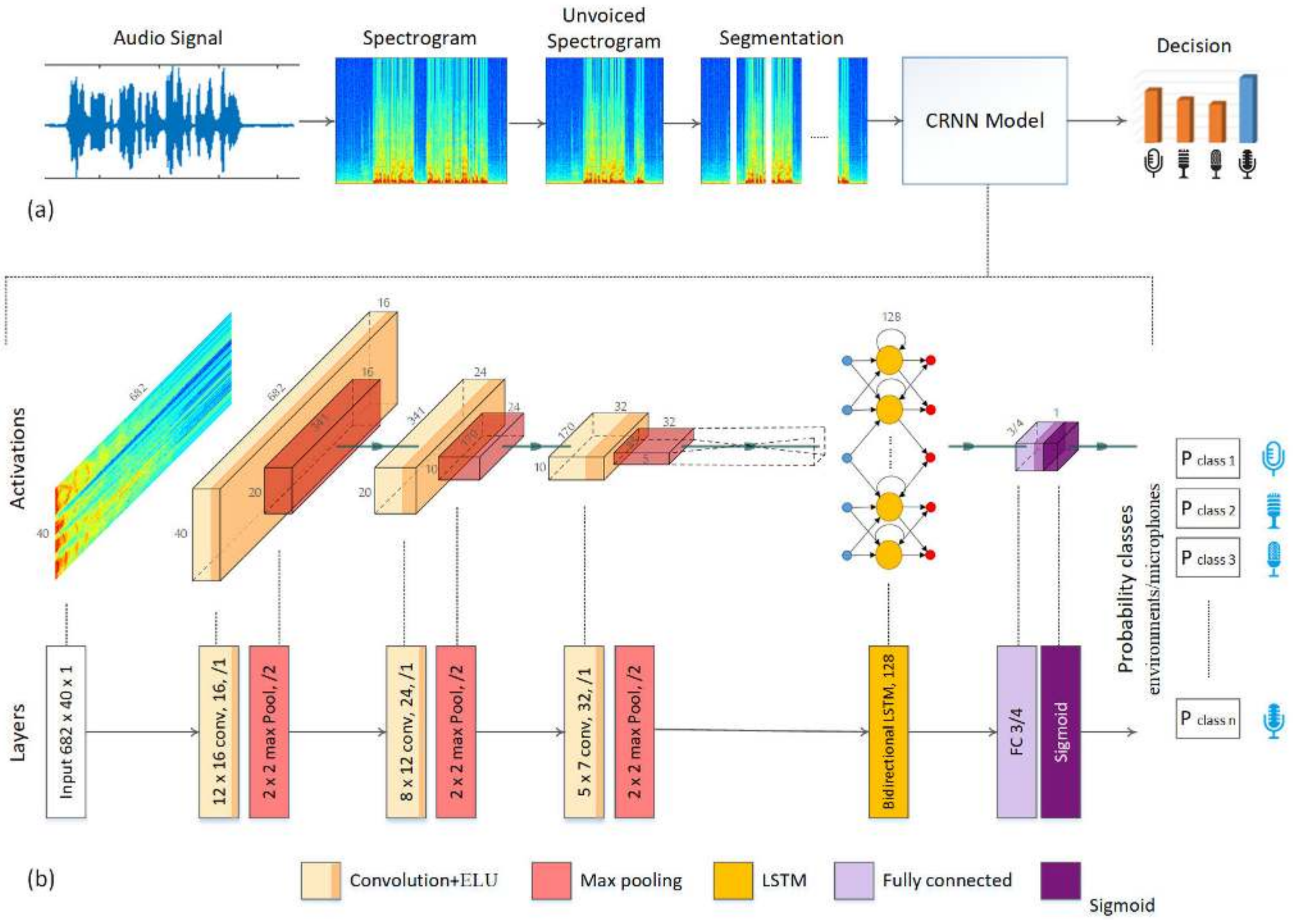


FIGURE 6. The proposed system architecture: a) a diagram of the proposed microphone and environment classification system from audio; b) the deep neural network architecture using CRNN.

units (ELU), as expressed in the equation below.

$$g(x) = e^{x}-1 \; for \; x \leq 0 \quad \text{and}$$
$$g(x) = x \; for \; x > 0 \tag{6}$$

The CRNN model consists of a CNN network followed by an RNN network, which utilizes both the spatial and temporal features of an audio signal. LSTM model is a type of RNN, well-suited to learn patterns in time-series data and capable of learning long-term relationships. LSTM overcomes the vanishing gradient problem in traditional RNN networks. The proposed CRNN model combines the advantages of both CNN with LSTM architectures. This hybrid configuration is discussed in a following subsection.

The LSTM model is composed of LSTM cells [39]. The LSTM cell is controlled by three gates: input, forget, and output gates. The entries to LSTM gates are current time data and the hidden previous time data. The values of the input, forget, and output gates are calculated by three fully connected layers with the sigmoid function. Stacked LSTM cells can form an LSTM layer. These LSTM layers together can form either unidirectional or bidirectional LSTM. The equations that control the transition of the LSTM states are:

$$\mathrm{I_t} = \sigma\left(X_t W_{xi} + H_{t-1} W_{hi} + b_i\right) \tag{7}$$
$$\mathrm{F_t} = \sigma\left(X_t W_{xf} + H_{t-1} W_{hf} + b_f\right) \tag{8}$$
$$\mathrm{O_t} = \sigma\left(X_t W_{xo} + H_{t-1} W_{ho} + b_o\right) \tag{9}$$
$$\tilde{\mathrm{C}}_\mathrm{t} = tanh\left(X_t W_{xc} + H_{t-1} W_{hc} + b_c\right) \tag{10}$$
$$\mathrm{C_t} = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t \tag{11}$$
$$\mathrm{H_t} = O_t \odot \tanh(C_t) \tag{12}$$

where h refers to the hidden units, d denotes the number of inputs, n is the batch size, $X_t \in \mathbb{R}^{n\times d}$ refers to the input, $H_t \in \mathbb{R}^{n\times h}$ refers to the hidden state, $H_{t-1} \in \mathbb{R}^{n\times h}$ denotes the hidden state of the previous time step, $I_t \in \mathbb{R}^{n\times h}$ is the input gate, $F_t \in \mathbb{R}^{n\times h}$ is the forget gate, $O_t \in \mathbb{R}^{n\times h}$ is the output gate, $\tilde{C}_\mathrm{t} \in \mathbb{R}^{n\times h}$ is the candidate memory cell, $\mathrm{C_t} \in \mathbb{R}^{n\times h}$ is the final memory cell, and $W_{xi}, W_{xf}, W_{xo}, W_{xc} \in \mathbb{R}^{d\times h}$, $W_{hi}, W_{hf}, W_{ho}, W_{hc} \in \mathbb{R}^{h\times h}$, $b_i, b_f, b_o, b_c \in \mathbb{R}^{1\times h}$ are the weight and bias parameters.

In a bidirectional LSTM, two layers are working in forwarding and backward time direction respectively. These layers help to learn bidirectional long-term dependencies between time steps. For bidirectional LSTM, given a minibatch input $X_t \in \mathbb{R}^{n \times d}$, where n refers to the number of samples and d refers to the number of inputs in each sample, and let f the activation function of hidden layer. For any time step t, the forward and backward hidden states are $\vec{H}_\mathrm{t} \in \mathbb{R}^{n \times h}$ and $\overleftarrow{H}_\mathrm{t} \in \mathbb{R}^{n \times h}$, respectively, where h is the number of hidden units. The parameters of the forward and backward hidden state are updated as follows:

$$\vec{H}_\mathrm{t} = f\left(X_t W_{xh}^{(f)} + \vec{H}_{t-1} W_{hh}^{(f)} + b_h^{(f)}\right) \tag{13}$$

$$\overleftarrow{H}_\mathrm{t} = f\left(X_t W_{xh}^{(b)} + \overleftarrow{H}_{t+1} W_{hh}^{(b)} + b_h^{(b)}\right) \tag{14}$$

where, $W_{xh}^{(f)} \in \mathbb{R}^{d \times h}, W_{hh}^{(f)} \in \mathbb{R}^{h \times h}, W_{xh}^{(b)} \in \mathbb{R}^{d \times h}$, and $W_{hh}^{(b)} \in \mathbb{R}^{h \times h}$ are the weights of the model, and $b_h^{(f)} \in \mathbb{R}^{1 \times h}$ and $b_h^{(b)} \in \mathbb{R}^{1 \times h}$ are the biases.

Next, we concatenate the forward and backward hidden states $\vec{H}$ and $\overleftarrow{H}_\mathrm{t}$ to obtain the hidden state $H_\mathrm{t} \in \mathbb{R}^{n \times 2h}$ to be fed into the output layer. The output layer is calculated as:

$$O_\mathrm{t} = H_\mathrm{t} W_{hq} + b_q \tag{15}$$

where, $W_{hq} \in \mathbb{R}^{2h \times q}$ and $b_q \in \mathbb{R}^{1 \times q}$ refer to the weight matrix and bias parameters of the output layer, respectively.

A. ARCHITECTURE OF THE PROPOSED MODELS

The CNN network consists of 3 convolution layers (Conv) and one fully connected (FC) layer followed by a classification layer. The 3 convolutional layers consisted of 16, 24, and 32 filters of dimensions 12 × 16, 8 × 12, and 5 × 7 with one-pixel stride, respectively. Each of them was attached by an exponential linear unit (ELU) as a non-linear activation function and a 2 × 2 max-pooling layer with a stride of 2. Adding non-linearity using ELU helps the CNNs train much faster. The last convolutional layer (Conv3) was followed by a FC layer with the number of neurons equal to the number of environments or microphones in the selected corpus. A dropout layer with a probability of 0.4 was used to generalize the network and prevent overfitting [40]. A sigmoid function was used at the end of the FC layer to produce the classification outputs in the form of prediction accuracy for different environments or microphones.

The CNN model received the speech spectrogram as segments of size Ss × 682, where Ss specifies the size of the segment and 682 is the number of frequency points generated by the short-time FFT.

The CRNN model consisted of the same convolutional layers in the proposed CNN architecture linked to a single bidirectional LSTM layer as demonstrated in Figure 6. A bidirectional LSTM layer with 128 units was tailed after the last convolutional layer (Conv3) followed by a FC layer with the number of neurons equal to the number of environments or microphones in the selected corpus. After the LSTM layer, a dropout layer with a probability of 0.4 was used to generalize the network and prevent overfitting [40]. At the end of the fully connected layer, a sigmoid function was used to produce the classification outputs in the form of prediction accuracy for different environments or microphones.

TABLE 4. Human perceptual test confusion matrix for environment classifications.

		Predicted Classes		
		Silent room	Office	Cafeteria
Actual Classes	Silent room	**70.10**	29.90	0.00
	Office	24.00	**60.00**	16.00
	Cafeteria	4.85	8.74	**86.41**

TABLE 5. Human perceptual test confusion matrix for microphone classifications.

		Predicted Classes			
		Medium-quality-1	Medium-quality-2	Mobile	High-quality
Actual Classes	Medium-quality-1	**85.94**	7.81	1.56	4.69
	Medium-quality-2	7.02	**56.14**	7.02	29.82
	Mobile	1.12	10.11	**74.16**	14.61
	High-quality	2.22	30.00	11.11	**56.67**

The CNN and CRNN models were implemented using TensorFlow [41] with Keras as a front-end system [42]. The models were trained by one GPU, Nvidia GeForce RTX-2080-Ti 11 GB, with an Intel Xeon E5-2600 CPU and 32 GB RAM. Each experiment was trained 10 different times for a duration of 200 epochs with an early stopping mechanism that ended the training process if the loss does not decrease after 20 epochs. The batch size was set to 128 samples. An Adam adaptive gradient descent optimizer with a learning rate of 0.001 was used to train the models [43].

VI. EXPERIMENTS

The work was carried out as follows. First, for each audio file, we extracted overlapping frames. We computed the spectrogram of each frame using PRAAT software [44]. Then, we further segmented each spectrogram into segments, with a size of Ss × 682. We set the value of Ss to 40 frames upon examining several values for the segment size. We carried out 10 runs in the experiment, and for each run, we split data into 80% for training and validation and 20% for testing.

A. OUTCOME PREDICTION CALCULATION

The system we developed predicts input as either "environment", or "microphone". Every separate spectrogram input represents the belief value for the corresponding output (i.e., either environment or microphone), and hence, we developed our model accordingly. We applied a probabilistic evaluation

TABLE 6. System accuracies for environments and microphones classification using different deep learning techniques with voiced, unvoiced, and whole audio signal. M: Male, F: Female, All: Both Male and Female.

			Accuracy (%) ± Standard Deviation		
			Voiced Phonemes	Unvoiced Phonemes	Whole Audio (Voiced and Unvoiced)
Environments Classification	CNN	M	96.17 ± 1.135	99.7 ± 0.285	98.29 ± 0.481
		F	99.13 ± 0.641	99.4 ± 0.343	98.66 ± 0.18
		All	95.37 ± 1.103	97.91 ± 0.973	98.4 ± 0.374
	CRNN	M	96.95 ± 0.389	99.31 ± 0.44	98.32 ± 0.94
		F	97.77 ± 0.735	99.43 ± 0.429	99.11 ± 0.757
		All	96.0 ± 0.62	98.0 ± 1.537	98.9 ± 0.395
Microphones Classification	CNN	M	99.28 ± 0.315	99.49 ± 0.322	99.54 ± 0.559
		F	98.23 ± 0.691	98.21 ± 0.758	97.72 ± 0.474
		All	98.04 ± 0.532	98.12 ± 0.529	98.39 ± 0.654
	CRNN	M	98.98 ± 0.658	99.72 ± 0.006	99.84 ± 0.4
		F	98.13 ± 0.12	98.34 ± 0.22	98.26 ± 1.584
		All	98.31 ± 0.648	98.57 ± 0.434	98.76 ± 0.337

for system prediction using mean prediction-based reasoning procedure. The accuracy of model prediction is considered acceptable if more than 50% of the audio files are predicted correctly. We also believed that using multiple spectrograms would lead to higher accuracy. Based on the gathered predictions of the deep models, we calculated the probabilities of each prediction.

B. EVALUATION OF SYSTEM ACCURACY

To determine the accuracy of the system, we have two options, the first is the evaluation based on the segment-level, and the second by adopting the file-level. In this work, we considered the file level. Each audio file contained different numbers of segments (S). The accuracy of each segment was calculated to compute the whole audio file accuracy as represented in equations (16), (17), and (18).

The total sum of all output probabilities for any segments is 1 as demonstrated in (16), where $\boldsymbol{j}$ is the predicted output, $\boldsymbol{n}$ is the number of environments or microphones, $\boldsymbol{S}$ is the total number of segments in the processed file, $\boldsymbol{i}$ represent any segment from 1 *to* S, and the probability that segment i is assigned to output j denoted by $\boldsymbol{p_{ij}}$.

$$\sum_{j=1}^{n} p_{ij} = 1, \quad i = 1 \ldots S \tag{16}$$

In the following step, we calculated the average probability ($\overline{\boldsymbol{p_j}}$) of all segments in the file for the targeted output $\boldsymbol{j}$ (environment or microphone) as demonstrated in (17).

$$\overline{p_j} = \frac{1}{S}\sum_{i=1}^{S} p_{ij}, \quad j = 1 \ldots n \tag{17}$$

Finally, we assigned the highest average probability to the output of the file as in (18)

$$\text{argmax}_{1 \leq j \leq n} \overline{p_j} \tag{18}$$

The highest average probability value determines if a file is correctly or falsely recognized, based on the averages of the number of targeted output (environments or microphones).

VII. RESULTS AND DISCUSSION

The main target of any classification system is to achieve the highest accuracy with the lowest processing time. Many experiments were carried out in this study to achieve our goal in several different ways.

A. HUMAN PERCEPTUAL TEST RESULTS

The average human-based classification accuracy for all subjects was 72% and 68% with a standard deviation of 10.8 and 17.4 for the environments and microphones classification, respectively. As shown in Table 4, the cafeteria achieved the highest classification accuracy due to background noise which helped to distinguish it from the other environments. The volunteers also had difficulty differentiating between a silent room and an office, as can be seen in the confusion matrix. As for the microphones classification, the minimum accuracy was obtained for both Medium-quality-2 and High-quality microphones with a confusion rate around 30% between them, on the other hand, the Medium-quality-1 achieved the highest accuracy with an accuracy of 85% as shown in Table 5.

B. CRNN AND CNN SYSTEMS RESULTS

CRNN and CNN were applied for automatic verification using only the spectrogram as a feature. The effect of the speaker's gender, in addition to the effect of using voiced or unvoiced phonemes in the system accuracy results are presented. Furthermore, the classifications of environments and microphones are presented in different situations. These results were generated by the designed system in over 10 runs for each experiment with similar system parameters to reach the highest robust accuracy results through the proposed system and features used.

Many experiments were performed to investigate the effect of the speakers' gender, as shown in Table 6, which illustrates the average of 10 runs of each experiment in different cases. In these experiments, voiced phonemes, unvoiced phonemes, and the whole audio signal were used as input for the classification of environments and microphones. We noticed that the system accuracy for the female speakers was a little higher than for the male speakers in environment classification, in contrast with microphone classification where system accuracy for male speakers was a little higher than for female speakers.

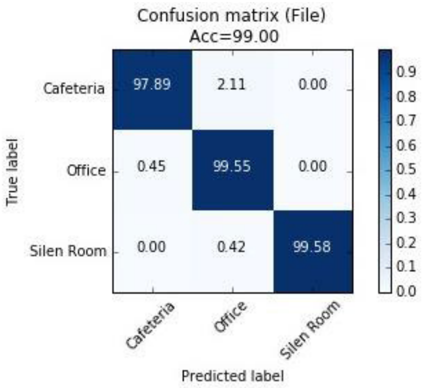
The effect of the voiced or unvoiced phonemes used in system accuracy was investigated to specify the most valuable segment in the audio file for the classification process. We conducted experiments using only voiced segments, unvoiced segments, and the entire audio file, regardless of segment type, voiced or unvoiced. Table 6 presents detailed results for environments and microphone classification using CRNN and CNN models with voiced, unvoiced, and whole audio signals. For further clarification. As shown in Table 6 the CRNN and CNN models achieved good results for environment classification or microphone classification by using only the spectrogram feature. In addition, the results showed that the use of unvoiced phonemes achieved excellent results compared to the use of only voiced phonemes in most of the results from both environment and microphone classification. Besides that, it was clear that using the whole audio signal did not add significant improvement to the classification and, in many cases, accuracy was decreased. In general, the results show that CRNN outperformed CNN in both environment and microphone classification.

Seddiq *et al.* in [45] presented the modern standard Arabic voiced and unvoiced phonemes where the number of voice phonemes is 21, and the number of unvoiced phonemes is 11, in addition to 2 phonemes (/ʔ/, and /h/) which are not classified as voiced or unvoiced.

Figure 7 illustrates the Arabic phonemes distribution percentage into voiced and unvoiced. As shown in this figure, Arabic voiced phonemes represent around 2-thirds of the Arabic speech phonemes, while the unvoiced phonemes represent only the third. Depending on these results, if we used only unvoiced phonemes, the required time for the classification process will be reduced to one third only of the required time for the classification process by using the whole audio signal or to half the required time for the classification process by using voiced phonemes. Therefore, to save time and to get higher accuracy, we performed our remaining experiments by using unvoiced phonemes only.

With respect to the first and second experiments (environments and microphones classification), which we referred to in section IV (data preparation and experimental setup), Table 7 shows the confusion matrix of the environment classification (one run). As shown in the table, the overall system accuracy reached 99%. The silent environment showed the highest accuracy followed by the office environment and finally cafeteria environment. In other words, noise plays an important role in this situation where when the noise is too

TABLE 7. Environment classifications confusion matrix.

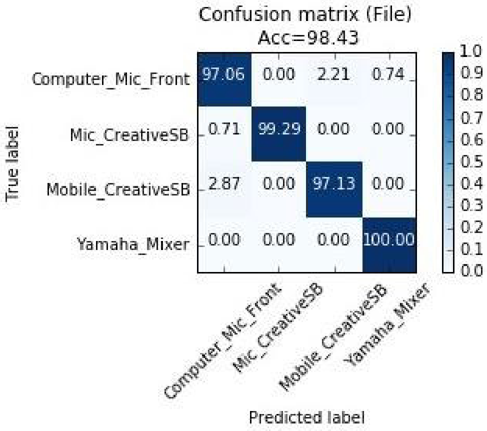


Confusion matrix (File)
Acc=99.00

True label \ Predicted label	Cafeteria	Office	Silen Room
Cafeteria	97.89	2.11	0.00
Office	0.45	99.55	0.00
Silen Room	0.00	0.42	99.58

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

TABLE 8. Microphone classifications confusion matrix.



Confusion matrix (File)
Acc=98.43

True label \ Predicted label	Computer_Mic_Front	Mic_CreativeSB	Mobile_CreativeSB	Yamaha_Mixer
Computer_Mic_Front	97.06	0.00	2.21	0.74
Mic_CreativeSB	0.71	99.29	0.00	0.00
Mobile_CreativeSB	2.87	0.00	97.13	0.00
Yamaha_Mixer	0.00	0.00	0.00	100.00

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

TABLE 9. Microphone classification accuracy (± standard deviation) in three different environments.

	Environment		
	Silent room	**Office**	**Cafeteria**
Medium-quality-2	100 ± 0	95 ± 2.549	98.06 ± 1.685
Medium-quality-1	100 ± 0	100 ± 0	97.79 ± 2.021
Mobile-Mic	100 ± 0	94.23 ± 1.951	100 ± 0
High-quality	100 ± 0	100 ± 0	98.9 ± 1.598
Microphone classification (Average)	**100 ± 0**	**97.36 ± 0.781**	**98.78 ± 0.708**

TABLE 10. Environment classification accuracy (± standard deviation) using four different qualities of recording devices.

	Different qualities of recording devices			
	High-quality	**Medium-quality-1**	**Medium-quality-2**	**Mobile mic**
Cafeteria	99.63 ± 1.11	100 ± 0	100 ± 0	98.91 ± 1.129
Office	99.69 ± 0.94	99.82 ± 0.547	99.21 ± 1.307	98.87 ± 1.416
Silent room	100 ± 0	100 ± 0	100 ± 0	99.33 ± 0.891
Environment classification (Average)	**99.78 ± 0.441**	**99.92 ± 0.227**	**99.72 ± 0.453**	**99 ± 0.559**

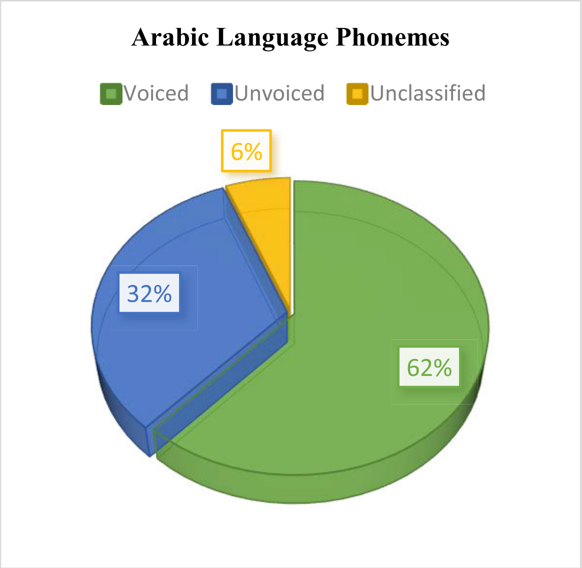


FIGURE 7. Distribution of the Arabic phonemes in the class of voiced and unvoiced segments.

low the accuracy is too high. Table 8 shows the microphone classification confusion matrix regardless of the recording environments, the gender of the speaker, or the background noise. The overall system achieved 98.43% accuracy results.

Regarding other experiments from the third experiment to the sixth experiment, which relate to microphones classification in each environment separately, and the effect of the environment in the microphones classification systems accuracies, Tables 9 presents these experiments' average 10 runs results. This table is not a confusion matrix, it is a summary and average of tens of experiments performed to achieve these experiments targets. As shown in the table, the system was able to fully classify all types of microphones in the silent room without any error in addition to 98.8% of the microphones in the cafeteria room. Mobile-Mic achieved the highest accuracy in the cafeteria while in the office room on the contrary it achieved the lowest accuracy.

For more reliability and more investigation, the four microphone classification in the three different environments at the same time. The results also show that a silent room is the best for all the microphone types followed by the office room. Also, the High-quality microphone achieved the best performance, the reason for this is what characterizes this type of microphone which is known as high-quality microphones.

Concerning environment classification using recording devises with different qualities, which is the last experiment, similar to Table 9, Table 10 presents this experiment's results.

The proposed system can classify any type of microphone without regarding the recorded environment where the average accuracy is more than 99% for all microphones, as shown in this table. Also, the environment can be classified regardless of the recording device where the lowest accuracy reaches 98.87%.

TABLE 11. Comparison of the proposed CRNN environment classification model with other studies using different datasets with different number of classes.

Method	features	Classifier	Dataset			Accuracy (%)
			Name	No of Classes	Classes	
Bae et al [56]	MFCC and spectrum	Combination of LSTM and CNN	DCASE-2016 [50]	17	Beach, bus, cafe/restaurant, car, city-center, forest-path, grocery-store, home, library, metro-station, office, park, residential-area, train, and tram	79.15
Zöhrer et al [57]	spectrogram	3-layer GRNN				79.1
Mafra et al [54]	averaged Mel-log spectrogram	Linear SVM	DCASE 2013 [52]	10	Bus, busy-street, office, open-air-market, park, quiet-street, restaurant, supermarket, tube, and tube station.	75
Ghulam et al [55]	MFCCs, MPEG-7 audio features, and combination of MFCCs and MPEG-7	HMM	Recorded in Riyadh City	10	Restaurant, Crowded Street, Quiet Street, Shopping Mall, Car-Open Window, Car-Closed Window, Corridor, Office Room, Desert, and Park	above 80 for all environments
Human Perceptual Test	-	-	**KSU_DB** [3]	4	**Silent room (very quiet), Office room (quiet), and Cafeteria (noisy)**	**72%**
Proposed System	**Spectrogram**	**CRNN**	**KSU_DB** [3]	**3**	**Silent room (very quiet), Office room (quiet), and Cafeteria (noisy)**	**98.0**

TABLE 12. A Comparison of the proposed CRNN microphone classification model with other studies using different datasets.

Method	features	Classifier	Dataset			Accuracy (%)
			Name	No of Classes	Classes	
Y. Jiang et al [7]	MFCC	GSV, SVM, SRC	Ahumada-25 [8]	4	25 speakers. 4 different microphones	89.04
Human Perceptual Test	-	-	**KSU_DB** [3]	**4**	**4 different recording devices**	**68%**
Proposed System	**Spectrogram**	**CRNN**	**KSU_DB** [3]	**4**	**136 speakers. 4 different recording devices**	**98.57**

The main target of this paper is to classify environments and microphones using spectrogram features with deep neural networks for the Arabic language, for that, we used the King Saud University speech database (KSU-DB). Since, there is no prior work on the same topic on the same database, we performed a human perceptual test to study human performance on the environment and microphone classification using the same dataset and used as a baseline system.

Tables 11 and 12 present other models used to classify different environments or different microphones as compared to our proposed models. In these tables, we clarified the number of classes used in each study and the proposed classifiers and features to make a general comparison of the available studies in this area.

The proposed system achieved excellent classification accuracies and minimized the classification time by using unvoiced phonemes.

VIII. CONCLUSION

In this paper, we have shown that the spectrogram and deep convolutional neural networks can be successfully trained to classify environments and recording devices in speech sound. Spectrograms were extracted from the audio files of the KSU_DB dataset and fed as 2D images to the CNN and CRNN (CNN+LSTM) models. The selected KSU_DB dataset consists of three environments with different noise levels and four types of acquisition devices with different quality levels. Various experiments were performed to study the effect of speaker gender, sound segment, environmental noise, and acquisition device quality in the environment and microphone classifications. The results showed that the effect of gender varies with the classification task. In the environment classification, the accuracy was slightly higher for female speakers, while, in the microphone classification, it was slightly higher for male speakers. Unvoiced phonemes achieved excellent results with less processing time compared to voiced phonemes. The results also showed that using a full audio signal did not add a significant improvement to performance. The proposed CNN and CRNN models achieved high and robust results regardless of environmental noise or recording device quality. In general, CRNN outperformed CNN in both environment and microphone classification. As for future work, we will investigate the performance of the proposed method in a dataset with low inter-class variation such as microphones of the same manufacture and model, and environments in homogeneous rooms.

REFERENCES

[1] M. Zakariah, M. K. Khan, and H. Malik, "Digital multimedia audio forensics: Past, present and future," *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 1009–1040, Jan. 2018.

[2] H. Farid, "Image forgery detection," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 16–25, Mar. 2009.

[3] M. Alsulaiman, G. Muhammad, B. Abdelkader, A. Mahmood, and Z. Ali, *King Saud University Arabic Speech Database LDC2014S02*, Hard Drive Linguistic Data Consortium, Philadelphia, PA, USA, 2014.

[4] G. Baldini and I. Amerini, "An evaluation of entropy measures for microphone identification," *Entropy*, vol. 22, no. 11, p. 1235, Oct. 2020.

[5] X. Lin, J. Zhu, and D. Chen, "Subband aware CNN for cell-phone recognition," *IEEE Signal Process. Lett.*, vol. 27, pp. 605–609, Apr. 2020.

[6] G. Baldini, I. Amerini, and C. Gentile, "Microphone identification using convolutional neural networks," *IEEE Sensors Lett.*, vol. 3, no. 7, pp. 1–4, Jul. 2019.

[7] Y. Jiang and F. H. F. Leung, "Source microphone recognition aided by a kernel-based projection method," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 11, pp. 2875–2886, Nov. 2019.

[8] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguiar, "AHUMADA: A large speech corpus in Spanish for speaker characterization and identification," *Speech Commun.*, vol. 31, nos. 2–3, pp. 255–264, Jun. 2000.

[9] O. Eskidere and A. Karatutlu, "Source microphone identification using multitaper MFCC features," in *Proc. 9th Int. Conf. Electr. Electron. Eng. (ELECO)*, Nov. 2015, pp. 227–231.

[10] Ö. Eskidere, "Source microphone identification from speech recordings based on a Gaussian mixture model," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 22, no. 3, pp. 754–767, Apr. 2014.

[11] R. Aggarwal, S. Singh, A. K. Roul, and N. Khanna, "Cellphone identification using noise estimates from recorded audio," in *Proc. Int. Conf. Commun. Signal Process.*, Apr. 2014, pp. 1218–1222.

[12] C. Hanilçi and T. Kinnunen, "Source cell-phone recognition from recorded speech using non-speech segments," *Digit. Signal Process.*, vol. 35, pp. 75–85, Dec. 2014.

[13] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: A first practical evaluation on microphone and environment classification," in *Proc. 9th workshop Multimedia Secur.*, Sep. 2007, pp. 63–74.

[14] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, M. Alsulaiman, and M. Bukhari, "Formant analysis in dysphonic patients and automatic Arabic digit speech recognition," *Biomed. Eng. Online*, vol. 10, no. 1, pp. 1–12, 2011.

[15] A. Al-nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of voice pathology detection and classification on different frequency regions using correlation functions," *J. Voice*, vol. 31, no. 1, pp. 3–15, Jan. 2017.

[16] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model," *J. Voice*, vol. 30, no. 6, p. 757, Nov. 2016.

[17] G. Muhammad, G. Altuwaijri, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, and A. Al-Nasheri, "Automatic voice pathology detection and classification using vocal tract area irregularity," *Biocybern. Biomed. Eng.*, vol. 36, no. 2, pp. 309–317, 2016.

[18] B. Gerazov, Z. Kokolanski, G. Arsov, and V. Dimcev, "Tracking of electrical network frequency for the purpose of forensic audio authentication," in *Proc. 13th Int. Conf. Optim. Electr. Electron. Equip. (OPTIM)*, May 2012, pp. 1164–1169.

[19] R. Patole and P. Rege, "A comparative analysis of classifiers and feature sets for acoustic environment classification," *J. Audio Eng. Soc.*, vol. 67, no. 12, pp. 939–952, Dec. 2019.

[20] M. Markovic and J. Geiger, "Reverberation-based feature extraction for acoustic scene classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 781–785.

[21] H. Zhao and H. Malik, "Audio recording location identification using acoustic environment signature," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 11, pp. 1746–1759, Nov. 2013.

[22] H. Malik, "Acoustic environment identification and its applications to audio forensics," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 11, pp. 1827–1837, Nov. 2013.

[23] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Low-variance multitaper mel-frequency cepstral coefficient features for speech and speaker recognition systems," *Cognit. Comput.*, vol. 5, no. 4, pp. 533–544, Dec. 2013.

[24] M. Narkhede and R. Patole, "Acoustic scene identification for audio authentication," in *Soft Computing and Signal Processing*. Singapore: Springer, 2019, pp. 593–602.

[25] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia Comput. Sci.*, vol. 112, pp. 2048–2056, Jan. 2017.

[26] C. Kotropoulos, "Telephone handset identification using sparse representations of spectral feature sketches," in *Proc. Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2013, pp. 1–4.

[27] Y. Panagakis and C. Kotropoulos, "Automatic telephone handset identification by sparse representation of random spectral features," in *Proc. Multimedia Secur.*, Sep. 2012, pp. 91–96.

[28] Y. Panagakis and C. Kotropoulos, "Telephone handset identification by feature selection and sparse representations," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2012, pp. 73–78.

[29] A. S. Altamrah, H. Altaheri, M. Alsuliman, O. F. Shehadeh, and Y. A. Alotaibi, "An acoustic analysis and comparison of two unique and almost identical arabic emphatic phonemes," in *Proc. Eur. Model. Symp. (EMS)*, Nov. 2017, pp. 84–88.

[30] A. H. Meftah, H. Mathkour, S. Kerrache, and Y. A. Alotaibi, "Speaker identification in different emotional states in Arabic and English," *IEEE Access*, vol. 8, pp. 60070–60083, 2020.

[31] M. S. Hossain and G. Muhammad, "Environment classification for urban big data using deep learning," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 44–50, Nov. 2018.

[32] R. Buchholz, C. Kraetzer, and J. Dittmann, "Microphone classification using Fourier coefficients," in *Proc. Int. Workshop Inf. Hiding*, Jun. 2009, pp. 235–246.

[33] H. Altaheri, M. Alsulaiman, and G. Muhammad, "Date fruit classification for robotic harvesting in a natural environment using deep learning," *IEEE Access*, vol. 7, pp. 117115–117133, Aug. 2019.

[34] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, and M. A. Mekhtiche, "Hand gesture recognition for sign language using 3DCNN," *IEEE Access*, vol. 8, pp. 79491–79509, 2020.

[35] M. Alsulaiman, Z. Ali, G. Muhammed, M. Bencherif, and A. Mahmood, "KSU speech database: Text selection, recording and verification," in *Proc. Eur. Model. Symp.*, Nov. 2013, pp. 237–242.

[36] J. K. Lee, "Wavelet speech enhancement based on voiced/unvoiced decision," in *Proc. 32nd Int. Congr. Expo. Noise Control Eng.*, Seogwipo, South Korea, Aug. 2003, pp. 4149–4156.

[37] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, Apr. 1993.

[38] F. Alshehri and G. Muhammad, "A comprehensive survey of the Internet of Things (IoT) and AI-based smart healthcare," *IEEE Access*, vol. 9, pp. 3660–3678, Jan. 2021.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[41] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, *arXiv:1603.04467*. [Online]. Available: https://arxiv.org/abs/1603.04467

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

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

[44] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [computer program]. Version 6.1.41," 2021. Accessed: Mar. 25, 2021. [Online]. Available: http://www.praat.org/

[45] 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.

[46] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Appl. Acoust.*, vol. 167, Oct. 2020, Art. no. 107389.

[47] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1015–1018.

[48] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.

[49] Z. Huang, C. Liu, H. Fei, W. Li, J. Yu, and Y. Cao, "Urban sound classification based on 2-order dense convolutional network using dual features," *Appl. Acoust.*, vol. 164, Jul. 2020, Art. no. 107243.

[50] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2016, pp. 1128–1132.

[51] H. Jleed and M. Bouchard, "Acoustic environment classification using discrete hartley transform features," in *Proc. IEEE 30th Can. Conf. Electr. Comput. Eng. (CCECE)*, Apr. 2017, pp. 1–4.

[52] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2013, pp. 1–4.

[53] Z. Ali, M. Imran, and M. Alsulaiman, "An automatic digital audio authentication/forensics system," *IEEE Access*, vol. 5, pp. 2994–3007, 2017.

[54] G. S. Mafra, N. Q. K. Duong, A. Ozerov, and P. Pérez, "Acoustic scene classification: An evaluation of an extremely compact feature representations," in *Proc. Detection Classification Acoust. Scenes Events*, Budapest, Hungary, Sep. 2016.

[55] G. Muhammad and K. Alghathbar, "Environment recognition for digital audio forensics using MPEG-7 and MEL cepstral features," *J. Electr. Eng.*, vol. 62, no. 4, pp. 199–205, Jul. 2011.

[56] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, Sep. 2016, pp. 11–15.

[57] M. Zöhrer and F. Pernkopf, "Gated recurrent networks applied to acoustic scene classification and acoustic event detection," in *Proc. Detect. Classif. Acoust. Scenes Events*, Sep. 2016, pp. 1–5.

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.

HAMDI ALTAHERI (Member, IEEE) received the master's degree in computer engineering from King Saud University, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Computer Engineering, College of Computer and Information Sciences. His research interests include computer vision, machine learning, deep learning, and artificial intelligence.

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. He is currently a Research Assistant with King Saud University. 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, and artificial intelligence.

GHULAM MUHAMMAD (Senior Member, IEEE) received the B.S. degree in computer science and engineering from the Bangladesh University of Engineering and Technology, in 1997, and the M.S. and Ph.D. degrees in electronic and information engineering from the Toyohashi University and Technology, Japan, in 2003 and 2006, respectively. He is currently a Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), Riyadh, Saudi Arabia. He has authored and coauthored more than 250 publications, including IEEE/ACM/Springer/Elsevier journals, and flagship conference papers. He owns two U.S. patents. He has supervised more than 15 Ph.D. and master theses. He is involved in many research projects, as a Principal Investigator and a Co-Principal Investigator. His research interests include signal processing, machine learning, the IoTs, medical signal and image analysis, AI, and biometrics. He received the Best Faculty Award from the Department of Computer Engineering, KSU, from 2014 to 2015. He was a recipient of the Japan Society for Promotion and Science (JSPS) Fellowship from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

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, Florida, 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 the College of Computer And Information Sciences, King Saud University, as an Assistant Professor, and an Associate Professor, from 2008 to 2012, where he has also been a Professor, since 2012. His research interests include digital speech processing, specifically speech recognition and Arabic language, and speech processing.