King Saud University Emotions Corpus: Construction, Analysis, Evaluation, and Comparison

ALI HAMID MEFTAH, MUSTAFA A. QAMHAN, YASSER SEDDIQ, YOUSEF A. ALOTAIBI, (Senior Member, IEEE), AND SID AHMED SELOUANI, (Senior Member, IEEE)

1 College of Computer and Information Sciences, King Saud University, Riyadh 11574, Saudi Arabia
2 The National Center for Electronics and Photonics Technology, King Abdulaziz City for Science and Technology (KACST), Riyadh 11442, Saudi Arabia
3 Department of Information Management, Université de Moncton, Shippagan, NB E8S 1P6, Canada

Corresponding author: Ali Hamid Meftah (ameftah@ksu.edu.sa)

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

ABSTRACT Emotional speech recognition for the Arabic language is insufficiently tackled in the literature compared to other languages. In this paper, we present the work of creating and verifying the King Saud University Emotions (KSUEmotions) corpus, which was released by the Linguistic Data Consortium (LDC) in 2017 as the first public Arabic emotional speech corpus. KSUEmotions contains an emotional speech of twenty-three speakers from Saudi Arabia, Syria, and Yemen, and includes the emotions: neutral, happiness, sadness, surprise, and anger. The corpus content is verified in two different ways: a human perceptual test by nine listeners who rate emotional performance in audio files, and automatic emotion recognition. Two automatic emotion recognition systems are experimented with: Residual Neural Network and Convolutional Neural Network. This work also experiments with emotion recognition for the English language using the Emotional Prosody Speech and Transcripts Corpus (EPST). The current experimental work is conducted in three tracks: (i) monolingual, where independent experiments for Arabic and English are carried out, (ii) multilingual, where the Arabic and English corpora are merged in as mixed corpus, and (iii) cross-lingual, where models are trained using one language and tested using the other. A challenge encountered in this work is that the two corpora do not contain the same emotions. That problem is tackled by mapping the emotions to the arousal-valance space.

INDEX TERMS Arabic language, corpus, CRNN, digital speech processing, emotion, ResNet, spectrogram.

I. INTRODUCTION

Digital emotional speech processing is an essential area of digital speech processing to solve two main problems: comprehension of speech emotions and synthesizing them [1]. Speech corpora play a significant role in emotional speech processing. A corpus can be created with spontaneous speech, but it is challenging since it is not easy to find people who express real emotions during recording. There are acted speech corpora created by actors whose performance is very close to genuine emotions. Besides, there are elicited speech corpora created by stimulating speakers to evoke some target emotions [2]. The other factors that categorize emotional speech corpora are the spoken languages, the number of emotions considered, and the number of speakers involved. A corpus usually covers between two and seven types of emotions and sometimes more [3]. Some corpora are created with only one speaker [4], and others contain more than 600 speakers [5].

Most emotional speech corpora have some shortcomings that limit the expected benefits when utilizing them in speech emotion recognition (SER). That is, some necessary contextual information and phonetic transcriptions are insufficiently provided in those corpora. Some deficits are also related to the low quality of the audio signal or emotional acting performance. Inadequately small-sized corpora are also among these shortcomings [1], [6].

Many surveys on emotional speech corpora have been reported. Ververidis et al. [7] reviewed 32 different corpora, language, and emotions. Anagnostopoulos [8] listed...
23 speech corpus links for the corpus available on the web (either with full free access or under a license agreement) along with the respective access link. Sailunaz et al. [9] listed emotions, number of speakers, corpus type, and name for 13 different speech corpora. Rao et al. [10] investigated 32 corpora for English, Chinese, German, Japanese, Spanish, Swedish, Italian, and Russian. Besides a review of 37 different speech corpora, Kasuriya et al. [11] presented the design, construction, annotation process, and analysis of the Thai emotional speech corpus EMOLA.

Table 1 proposes a selected SER survey from 2006 till date, which includes the number of surveyed corpora, used language, the minimum and the maximum number of speakers, and selected emotions in surveyed corpora. From these surveys and as shown in Table 1, we observe the following:

- The most considered corpora are those in Spanish, English, German, and Danish, followed by Chinese, Japanese, Italian, and Hebrew.
- Most corpora are for acted speech created by professional or nonprofessional actors. That is because some legal and ethical issues may prevent researchers from recording real emotions.
- Emotions that exist in the majority of the surveyed corpora are anger, sadness, happiness, fear, disgust, surprise, and neutral.
- The average number of emotions per corpus is seven emotions. The minimum and maximum are two and fifteen emotions, respectively.
- The number of speakers per corpus ranges from one speaker up to a few hundred.
- The majority of the created emotional speech corpora are not available for public use.

Mustafa et al. [12] extracted and analyzed 260 articles from well-known online databases for 12 years from 2006 to 2017. They tried to answer the following question: what are the research focus on SER over the past 12 years based on databases, speech features, and classification? Regarding databases, they found that 76% of the databases used for SER are in European languages, with more than 44% using German emotional databases, followed by English (17%) emotional speech databases. Anger with (17%), sadness with (14%), neutral with (13%), happiness with (12%), fear with (10%), disgust and boredom with (8%), and surprise with (3%) are the emotions that are regularly investigated.

A. ARABIC SPEECH EMOTION CORPORA

From the above reviews, the absence of an Arabic emotional speech corpus is obvious. We find some studies and a few significant emotional databases.

Lamiaa [13] proposed an Egyptian Arabic speech emotion (EYASE) created from the award-winning Egyptian drama series and contains four emotions: angry, happy, neutral, and sad. The EYASE database was recorded by three males and three females lead professional actors. Lamiaa et al. [14] recorded four basic emotions: anger, fear, happiness, and sadness uttered by 32 Egyptian bilingual...
speakers (16 males and 16 females) to create the elicited bilingual Arabic/English speech emotion proprietary database. In the Algerian dialect, Dahmani et al. [15] presented natural Arabic language resources for emotion recognition. The corpus consists of 14 speakers with 1,443 utterances, which are complete sentences. They investigated 15 emotions with five that are dominants: enthusiasm, admiration, disapproval, neutral, and joy. Samira Klaylat et al. [16] collected a realistic speech corpus from the Arabic TV shows (Egyptian, Gulf, Jordan, and Lebanese) where the videos are labeled by their perceived emotions: happy, angry, or surprised. Eighteen human labelers were asked to listen to the videos and label each of them as happy, angry, or surprised, and the average result is used to label each video.

In the Tunisian dialect, Meddeb et al. [17] proposed their corpus REGIM_TES, which contains the following emotions neutral, sadness, fear, anger, and happiness, recorded by 12 speakers (six males and six females). Abdo et al. [18] developed an audio-visual phonetically annotated Arabic corpus for expressive text to speech recorded by seven speakers who read 500 sentences with the following six emotions: happiness, sadness, fear, anger, inquiry, and neutral. Al-Faham and Ghneim [19] developed an Arabic emotional speech corpus, covering five emotions: happiness, anger, sadness, surprise, and neutral, recorded by six speakers (three males and three females).

B. GOAL AND MOTIVATION

There are a few studies on emotional speech in the Arabic language compared to other languages. Regarding the emotional speech corpora, there are a few more or less significant emotional corpora. Besides, all the corpora that we found are not public and cannot be shared between researchers. There is no public Arabic emotional speech corpus except the KSUEmotions corpus, the subject of this study, which is published in the Linguistic Data Consortium (LDC) [20]. There is an urgent need to develop Arabic linguistic resources in the light of the recommended emotional corpora design criteria to be utilized for emotional speech recognition and synthesis [6].

This paper aims to report a methodology of developing, analyzing, and evaluating a new public Arabic emotional speech corpus (i.e., KSUEmotions) [20]. The study involves the design, development, and verification of this corpus. A past published effort results using human perceptual evaluation [21] is considered as the baseline for this work and to crosscheck with past and current accuracies.

The remainder of the paper is organized as follows. The major processes of designing prompt texts, selection of emotions and speakers, audio file recording, and characteristics are reported in Section 2. Section 3 presents the first corpus content assessed by applying a blind human perceptual test. The second corpus content assessed by using automatic emotion recognition systems based on Convolutional Neural Network (CNN), Convolutional Recurrent Neural Network (CRNN), and Residual Neural Network (ResNet) is presented in Section 4. A comparison to a similar English emotion corpus is conducted based on the application of CRNN and ResNet. Section 5 contains the results. Finally, Section 6 presents the conclusion.

II. KSUEmotions CORPUS DESIGN

A new emotional speech corpus for the Arabic language called KSUEmotions is presented in this paper. Corpus development is conducted in two phases. In Phase 1, an initial content is created and evaluated utilizing a human perceptual test where the recorded emotional speech is played for listeners who are asked to describe what type of emotion they perceive and to what degree in a numerical score (the details of this test is the topic of Section 3 below).

During that initial phase, the design of the current corpus relied on the past Text-To-Speech Database (KTD) corpus of King Abdulaziz City for Science and Technology [22] that we consider as a good baseline for our corpus. The purpose of the human perceptual test is to enhance the quality of outcomes of Phase 2. That is, the selection of text, emotions, and speakers for Phase 2 is made based on the results of the human perceptual test made in Phase 1. Thus, both subsets of Phase 1 and Phase 2 are included in the final release of the corpus. The process of designing and building the corpus is discussed in the subsequent subsections. Those steps are summarized in Table 2.

A. DESIGN OF PROMPT TEXTS

Sixteen sentences chosen from the KTD will be used in the creation of the new KSUEmotions corpus. These sentences were derived from newspaper articles of various media outlets. These sentences and their content vary from typical sad and happy news stories. They also differ in length, with the short sentences containing only four words, and the long sentences with about 16 words. For the term “questioning,” the word “لا/ بال” is added, which is a word that infers a question in the Arabic language. Based on the human perceptual test analysis results performed during Phase 1, only sentences that obtained the highest recognition rate have been selected to be used in Phase 2.

The shortest used sentence in Phase 1 contains four words, so we added the shortest two words to the content of Phase 2: “Yes” and “No” words for diversification and the possibility of studying and comparing the effect of short and long sentences as well as single words. Because they are very short words, the speakers had encountered difficulties in emulating the desired emotion.

B. SELECTION OF EMOTIONS

During the first building phase of KSUEmotions, we consider the five emotions of the baseline corpus (KTD): neutral, sadness, happiness, surprise, and questioning.

Based on the human perceptual test analysis results, questioning was given the highest score by the listeners, whereas happiness has the lowest score. The reason for the high score of the questioning emotion is due to the presence of the
question word /\textit{\textbf{?}}/, which influences the listener’s decision regardless of the speaker’s performance. On the other hand, the low evaluation in the case of happiness could be due to the meaning and content of the sentences, which play a significant role in the quality of emotional acting by speakers. For instance, we cannot express sincere happiness when the subject of the text is death, and particularly when the speakers are not professional actors. When the sentence is about sad occasions, it is difficult for the speaker to pronounce it in a happy way, for example, in case of disaster, pain, and/or death news.

In order to be consistent with other corpora in the field, and following the approaches adopted by many other experts and papers’ reviewers, such as those presented in Table 1, Questioning is not an emotion and can be considered as “confused” or “perplexed” emotion. Therefore, during Phase 2, we exclude the questioning emotion and incorporate the anger emotion. Thus, neutral, sadness, happiness, surprise, and anger are included in Phase 2.

### C. SPEAKER SELECTION

During Phase 1, 20 speakers record 16 sentences expressing the five different emotions. Information about the speakers is presented in Figure 1 and Table 3. The speakers include ten males aged between 20 and 37 years and ten females aged between 19 and 30 years. All speakers are either undergraduate or graduate students, except for one female who was still attending secondary school. During Phase 2, according to the results of the human perceptual test of Phase 1, we select the best acting seven male speakers and the best acting four female speakers among those who participated in Phase 1. The selection of speakers after Phase 1 is not purely based on a high score since it is highly desirable to maintain the diversity of speakers’ nationalities in both phases. Thus, we consider another factor, which is a nationality, since it is important to ensure that all nationalities of Phase 1 also exist during Phase 2. Moreover, we invite three Yemeni female speakers to participate in Phase 2 to compensate for the absence of female Yemeni speakers in Phase 1. These three new female speakers are between 20 and 25 years; two have undergraduate degrees, and the third reached secondary school. The new three female speakers are subject to an assessment process by performing initial recordings. Then, they are evaluated through the human perceptual test in order to ensure they are as good as the old speakers who are selected from Phase 1. That activity of initial recording and evaluation for the new female speakers is equivalent to the effort made by the speakers of Phase 1. Hence, that pre-activity for those new speakers is considered as a valid compensation for their absence in Phase 1.

### D. RECORDING AND FILENAME FORMAT

As we mentioned in the introduction, one of the big challenges facing researchers in speech emotion recognition is the difficulty of obtaining real emotion classes due to the difficulty of finding people who express real emotions during recording and due to ethical permissions reasons. To achieve audio recording environments that are very close to real-life conditions. Thus, a studio environment is not recommended in our case due to its idealized conditions that do not serve our purpose. Instead, the speakers are requested to record in a normal home, lab, or office environment using a SHURE.
A. H. Meftah et al.: KSUEmotions Corpus: Construction, Analysis, Evaluation, and Comparison

TABLE 3. Information about the speakers.

|                        | Phase 1 | Phase 2 | Full corpus |
|------------------------|---------|---------|-------------|
| Number of male speakers| 10      | 7       | 10          |
| Number of female speakers| 10     | 7       | 13          |
| Ages of male speakers | between 20 and 37 years | between 20 and 35 years | between 20 and 37 years |
| Ages of female speakers | between 19 and 30 | between 19 and 30 | between 19 and 30 |
| Nationalities of male speakers | 6 Yemenis | 4 Yemenis | 6 Yemenis |
|                         | 2 Syrians | 1 Syrian | 2 Syrians |
|                         | 2 Saudis  | 2 Saudis | 2 Saudis |
| Nationalities of female speakers | 8 Syrians | 3 Yemenis | 3 Yemenis |
|                         | 2 Saudis  | 2 Syrian | 2 Syrian |
|                         |          | 2 Saudis | 2 Saudis |
| Total number of speakers | 20      | 14      | 23          |
| Number of audio files by male speakers | 799    | 840     | 1639        |
| Number of audio files by female speakers | 797    | 840     | 1637        |
| Total number of audio files | 1596   | 1680    | 3276        |
| Total duration of recorded speech | 2 hours and 55 minutes | 2 hours and 15 minutes | 5 hours and 10 minutes |

FIGURE 1. Summary statistics about speakers.

58A high-quality microphone and an XPS 14Z Dell laptop running Windows 7, at a sampling rate of 16 kHz. Mono recording processing is applied using PRAAT software [23]. The speakers are asked to do many practice readings for the written sentences before recording.

The audio files are named using the $D_{xx}E_{xx}P_{gxx}S_{xx}T_{xx}$ format, as contained in Table 4. The $D_{xx}$ part indicates the serial number of the corpus-based on its order in the lab portfolio. This corpus is assigned the number 05; hence, all audio files in the corpus carry file names starting with “D05.” The $E_{xx}$ part encodes the emotion type (E00, E01, etc.). Speaker gender and code are indicated as $P_{gxx}$, where $g$ carries either 0 or 1 to indicate male or female, respectively, and $xx$ is a two-digit code representing the given speaker’s ID. The $S_{xx}$ code refers to the sentence (S01, S02…S18), and finally, $T_{xx}$ indicates the trial number (T01, T02, etc.). For instance, an audio file with the name D05E03P104S01T01 contains the first trial recording of Sentence 1 as recorded by female speaker number 4, targeting a surprise emotion (emotion number 03).

The corpus material of Phase 1 and Phase 2 can be primarily distinguished by directory names. That is, Phase 1 and Phase 2 files are organized under the two separate directories named “Phase_1” and “Phase_2”, respectively, to make it possible to distinguish between them. It is also worth mentioning that, with some exceptions, the trial number $T_{xx}$ in the file name can give some hint about phase number since all the file names in Phase 1 end with T01, while the file names in Phase 2, generally, end with T02 or T03. However, there are some exceptions since there are file names in Phase 2 that happen to end with T01, but those files still could be associated with Phase 2 as they belong to emotion or speakers that do not exist in Phase 1 (i.e., angry emotion and speakers P111, P112, and P113). Again,
Table 4. Filename format details.

| Field  | Description                  | Value(s)       | Decoding               |
|--------|------------------------------|----------------|------------------------|
| Dxx    | Corpus serial number         | D05            | KSUEmotions speech corpus |
|        |                              | E00            | Neutral                |
|        |                              | E01            | Happiness              |
|        |                              | E02            | Sadness                |
|        |                              | E03            | Surprise               |
|        |                              | E04            | Questioning            |
|        |                              | E05            | Anger                  |
| Exx    | Emotion type                 | P001           | Male speaker 1         |
|        |                              | P101           | Female speaker 1       |
|        |                              | P002           | Male speaker 2         |
|        |                              | P102           | Female speaker 2       |
|        |                              | P010           | Male speaker 10        |
|        |                              | P113           | Female speaker 13      |
| Pgxx   | Speakers’ identity code:     | S01            | Sentence 1             |
|        | g=gender xx=speaker ID       | S02            | Sentence 2             |
|        |                              | S18            | Sentence 18            |
| Sxx    | Sentence number              | T01            | 1st trial              |
|        |                              | T02            | 2nd trial              |
|        |                              | T03            | 3rd trial              |

Figure 2. Distribution of audio files with respect to speakers’ gender, emotion types and phases.

E. Final Content

The outcome after conducting the above steps is the KSUEmotions corpus consisting of 2 hours and 55 minutes in Phase 1 (1596 audio files), and 2 hours and 15 minutes in Phase 2 (1680 audio files) along with their labeling and alignment data. As illustrated earlier in Table 3, the share of male and female speakers out of the total number of audio files are 1639 and 1637, respectively. The details of these numbers are presented in the mirror chart in Figure 2(a), showing the distribution of audio files concerning speakers’ gender and emotion types. The chart shows that both genders have balanced distributions. There are two emotion types (i.e., questioning and anger) that have nearly half as many audio files as the other emotion types. This is because each one of these two emotions appears only in one phase (i.e., questioning in Phase 1 and anger in Phase 2), unlike the other emotions that are included in both phases as illustrated in Figure 2(b).

Figure 3 shows the contribution of each speaker in recording emotional audio files. The data related to the male speakers are visually depicted as a stacked area chart in Figure 3(a).
The chart shows the share of each one of the ten male speakers in recording the six types of emotions. The figure shows balanced distribution since the layers are of almost equal thickness except for three speakers who appear in thinner layers that are highlighted as dark-colored streaks (i.e., speakers P002, P007, and P010). The reason for the low contribution of these three speakers is that they are only involved in Phase 1. They are included in Phase 2 since they do not pass the human perceptual test. This explains why those speakers disappear for the anger emotion, which is only considered in Phase 2. Similarly, the female speaker’s data is shown in Figure 3(b). As mentioned earlier, four female speakers have contributed throughout the two phases. Namely, these are P101, P102, P109, and P110 who appear in the chart as the four layers (light-colored) that continue across all six emotions. However, the speakers from P103 to P108 inclusive appear in thin dark streaks that eventually disappear at the anger emotion. They are the speakers who are included during Phase 1 but excluded during Phase 2 based on their lowest score in the human perceptual test. As mentioned earlier, three new Yemeni female speakers participated in Phase 2 to augment the diversity of nationalities. These three speakers are P111, P112, and P113 who are visualized as dot-shaded streaks. Since they have contribution only in Phase 2, they appear as thinner layers and do not exist at the questioning emotion.

Figure 4 shows the distribution of the different sentences over speakers’ gender and phases. From Figure 4(a), for any particular sentence, the number of audio files that are recorded by male and female speakers is almost equal. However, the audio file count is not uniform for all sentences. This is because not all the 18 sentences of the corpus are recorded in the two phases. Figure 4(a) also shows that there are four corrupted files for the following sentences S02, S05, S08, and S12. As shown in Figure 4(b), the sentences that are common in Phase 1 and Phase 2 are from S05 to S12 inclusive, S15, and S16. It justifies the higher number of audio files recorded for those sentences compared to the rest of the sentences,
FIGURE 4. Distribution of audio files with respect to sentences, speakers’ genders and phases.

which are either recorded only in Phase 1 (i.e., from S01 to S04 inclusive, S13 and S14) or recorded only in Phase 2 (i.e., S17 and S18).

### III. HUMAN PERCEPTUAL TEST VERIFICATION

An assessment process is conducted to evaluate the performance of the emotional acting of speakers. That is, a perceptual test is conducted by asking listeners, upon listening to audio files, to identify the type of emotion that best describes an audio file. Besides, they evaluate to what degree they can feel other emotions in the played audio. A test session involves a listener and an attendant who plays audio files and records the listener’s feedback.

A listener is allowed to replay an audio file before giving the evaluation. However, a listener is not allowed to compare the performance of two speakers before giving an evaluation to the first one. A listener can take breaks without constraints. The human perceptual test is applied to the nine listeners (six males and three females). All listeners are Arabs except one male, who is an Indian, yet masters the Arabic language. He is in the 40s of age while the rest of the listeners are in their 20s. Audio files are played to listeners in random selection. Recall that each audio file represents, theoretically, only one emotion type out of five types. Thus, each listener provides five different ratings for one audio file, where each rating corresponds to one of the five possible emotions. Each rating (given as a percentage) measures how much particular emotion a listener can feel in the audio file.

The percentage scale is converted to a five-level mean opinion score (MOS) with levels described in Table 5 to make it easier to figure out the function of the well-known MOS scale. Therefore, for one particular audio file, each one of the nine listeners provides five ratings. As a result, each emotion type is evaluated nine times (i.e., one time by one listener), which are averaged to one score. Consequently, the audio file has five different average scores (i.e., one for each possible emotion). Finally, one emotion type corresponding to the highest score is selected as the final test result for that particular file. Table 6 shows a sample test record. As shown in this table, some files have been categorized as not recognized (NR). For the listeners, there is no option of an “unclassified label” or NR, we considered this option in the listeners’ outcome analysis phase if there is no clear decision among human testers and emotions. For example, as shown in Table 6, the file D05E00P001S18T02 when the file is recognized as two (or more) equal weight emotions at the same time, it is NR. Table 7 presents the human perceptual test results for Phases 1 and Phase 2 [21]. It is obvious from Table 7, that the overall accuracy of Phase 2 is much better than the overall accuracy of Phase 1. With no doubt, the experience that the speakers gained in Phase 1 positively contributes to this improvement in Phase 2 in addition to the selection process as that we presented in section II and also

| MOS | Quality       | Distortion              |
|-----|---------------|-------------------------|
| 1   | Bad           | Very annoying and objectionable |
| 2   | Poor          | Annoying, but not objectionable |
| 3   | Fair          | Perceptible and slightly annoying |
| 4   | Good          | Just perceptible, but not annoying |
| 5   | Excellent     | Imperceptible            |

TABLE 5. Mean opinion score (MOS) [35].
TABLE 6. Sample test record of the human perceptual test (Avg. average, NR. not recognized).

| Audio File        | Targeted | Average evaluation for 9 listener (R1 to R9) |
|-------------------|----------|---------------------------------------------|
|                   |          | Avg. Neutral | Avg. Happiness | Avg. Sadness | Avg. Surprise | Avg. Anger | Max Average | Result   |
| D05E00P001S12T02  | Neutral  | 4            | 1            | 2            | 1            | 1          | 4          | Neutral  |
| D05E00P001S12T03  | Neutral  | 5            | 1            | 1            | 1            | 1          | 5          | Neutral  |
| D05E00P001S15T02  | Neutral  | 4            | 1            | 2            | 1            | 1          | 4          | Neutral  |
| D05E00P001S15T03  | Neutral  | 4            | 1            | 2            | 1            | 1          | 4          | Neutral  |
| D05E00P001S16T02  | Neutral  | 5            | 1            | 1            | 1            | 1          | 5          | Neutral  |
| D05E00P001S16T03  | Neutral  | 5            | 1            | 1            | 1            | 1          | 5          | Neutral  |
| D05E00P001S17T02  | Neutral  | 5            | 1            | 1            | 1            | 1          | 5          | Neutral  |
| D05E00P001S17T03  | Sad      | 2            | 1            | 3            | 1            | 1          | 3          | Sad      |
| D05E00P001S18T02  | NR       | 3            | 1            | 3            | 1            | 1          | 3          | NR       |

TABLE 7. Confusion matrix results for human perceptual tests [21].

|                | Phase 1 | Phase 2 |
|----------------|---------|---------|
| Neutral        | 95.3%   | 93.2%   |
| Happiness      | 48.1%   | 77.4%   |
| Sadness        | 76.0%   | 92.6%   |
| Surprise       | 82.5%   | 92.6%   |
| Question       | 98.7%   | -       |
| Anger          | -       | 86.0%   |
| Overall Accuracy | 79.94% | 88.3%   |
|                | 84.24%  |         |

in Table 2. Although the work that is done Phase 1 serves the purpose of enhancing the quality of Phase 2, the former show satisfactory results in the human perceptual test (79.94%). Such a score is high enough to qualify the content of Phase 1 to be part of the final corpus beside Phase 2. Indeed, that will give the researcher more choices and will enrich the corpus.

IV. AUTOMATIC VERIFICATION

The methodology for SER includes two stages: feature extraction (e.g., source-based excitation features, prosodic features, vocal traction factors, and other hybrid features) and feature classification stage using linear and nonlinear classifiers (e.g., Bayesian Networks and Support Vector Machine for linear classifier and Gaussian Mixture Model and Hidden Markov Model for nonlinear classifier). Deep learning has been considered a rising research field in machine learning and has attracted significant attention. In image and video processing, Deep Neural Networks (DNNs) and CNN are used and provide efficient results while in speech classification, SER, and natural language processing, Recurrent Neural Networks and Long Short-Term Memory (LSTM) are more effective [24]. CNN has made significant progress in image processing. Recent studies show that CNN has demonstrated superior performance than traditional methods. This progress shows its superior capacity in capturing the structure of natural images. In general, regular CNN consist of four essential structure layers: convolution layer, nonlinear mapping layer or activation layer, pooling layer, and batch normalization layer [25].

In our previous preliminary work [26], an experiment of automatic SER was conducted using DNN. The automatic recognizer was designed using CRNN built with two sub-modules CNN, followed by LSTM. The input to the recognizer is the linearly-spaced spectrogram of the audio waveform computed from Phase 2 of the KsuEMotions corpus. In the final results of this system, the “surprise” emotion achieved the worst accuracy while the “sadness” emotion achieved the best accuracy with the system accuracy reach 84.6%. In this study, we applied ResNet and CNN only, without LSTM.

A. RESIDUAL NEURAL NETWORK (ResNet)

ResNet was presented at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2015, which won with a performance in the top-5 error rate (3.57%).

ResNet introduced an architecture, which trains a deeper network up to 152 layers 8x deeper than VGG nets with lower complexity [27]. The main idea of ResNet employs identity shortcut connections, which is a strategy of shortcuts or jump connections that skip pairs of groups of convolutional layers. The idea of shortcuts in ResNet is to prevent a very deep network from dying due to the gradients evolving through the stacking of mappings of identities. ResNet uses, at a given point, a signal that is the sum of the signal produced by the
two previous convolutional layers and the signal transmitted directly from the previous point to these layers, joining a processed signal with a previous step, as shown in Figure 5.

B. CONVOLUTIONAL NEURAL NETWORK (CNN)

CNNs are generally used for pattern recognition and give better information characterization. These systems have little size neurons present on each layer of the structured model engineering that cycle the information as open fields [24]. The CNN architecture is an arranged group of neural networks in a particular order with layers of different sizes, where each layer performs a particular contribution. The earlier layers learn low-level features, while the deeper layers learn high-level features that represent, for example, the speaker in a speech or object in an image. A typical CNN model consists of several convolutional layers, which transform the spectrogram of speech into several feature maps. Time-frequency maps are extracted through the convolution of speech waves with a wide range of filters during the training phase. Pooling layers are a type of CNN layer that is placed after convolution layers to abstract the network. Pooling layers reduce the spatial resolution of convolutional feature maps, which reduces network computation.

V. EXPERIMENTAL SETUP

A. EMOTIONAL PROSODY SPEECH AND TRANSCRIPTS CORPUS (EPST)

EPST is a public corpus produced by the Linguistic Data Consortium (catalog number LDC2002S28). Three males and five females in their twenties (professional actors) recorded EPST by producing a series of semantically neutral utterances (dates and numbers) using fourteen different emotions ranging from hot anger, cold anger, panic, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt, and neutral. The recorded speech files were sampled at 22.05 kHz in a 2-channel interleaved 16-bit PCM format [3]. Table 8 compares statistics taken from the datasheets of each corpus.

| TABLE 8. Corpora comparison. |
|-----------------------------|
| KSUEmotions | EPST          |
| Type         | Acted | Acted |
| Number of Speakers | 23 | 8        |
| Number of speech files | 3280 | 4787    |
| Average Audio File Duration | 4 s | 1.5 s    |
| Sampling Rate | 16 kHz | 22.05 kHz |

B. DATA PREPARATION AND SELECTED FEATURES

The KSUEmotions corpus contains five emotions, whereas the EPST corpus contains 14 emotions in addition to the neutral. The two corpora have got four emotions in common, which are neutral, sadness, happiness, and anger. There are two types of anger in the EPST corpus: hot anger and cold anger. We select hot anger only and ignored the cold one because the anger emotion in KSUEmotions is considered hot. We downsample the data of EPST corpus to 16kHz.

The following three main experiments are carried using the two corpora: (i) monolingual SER, (ii) multilingual SER, and (iii) cross-lingual SER for the common emotions. In the monolingual experiment, the SER is applied to each corpus in independently. In the multilingual experiment, the common emotions between KSUEmotions Phase 2 and EPST that are neutral, sadness, happiness, and hot anger are mixed and grouped in a new corpus called KSUEPST. In the monolingual and multilingual experiments, each corpus (KSUEmotions, EPST, or KSUEPST) is randomly split into training and testing subsets.

The training and testing subsets are split as follows: 80% for training and validation, and 20% for testing. For every experiment run, those two subsets are randomly selected such that speakers or sentences in the training and testing subsets are not necessarily mutually exclusive. Also, in each experiment, we conduct ten runs where the training phase is carried out for a duration of 200 epochs with a batch size of 128 samples. The Adam adaptive gradient descent optimizer with a learning rate of 0.001 is used [28].

In the third experiment (i.e., the cross-lingual experiment), the training subset is built by taking all audio files for the targeted emotions purely from only one corpus (either KSUEmotions or EPST) while testing data is purely taken from the other corpus. This experiment involves three parts (to be described shortly), where each part is carried out twice by switching the training and testing corpora each time.

The three main parts of the cross-lingual experiment are as follows: First, cross-lingual emotion recognition conducted considering the four emotions that are common between the two corpora. The second part is similar to the first, but the emotion classes are subject to arousal-valance mapping covering the arousal classes (high and low), and the valance classes (positive, and negative). In the third part of this experiment, the work is extended to cover all the emotions in the
two corpora (not just the common emotions), also, applying arousal-valance mapping on those emotions.

For the three aforementioned experiments, the speech signal is sliced into overlapping frames of length equal to 30 ms each. A new frame is started every 10 ms step. Frames are qualified by their voicing, which is measured by pitch activity. That is, frames with valid pitch values are selected for feature extraction. Next, a 512-point spectrogram is calculated for the qualified frames.

Figure 6 shows the change in spectrograms for the same sentence (S05) which is spoken by the same speaker (P001) in the different five emotions. Then, spectrograms are grouped into segments consisting of 120 frames. The shortage is compensated with frame padding copied from the same segment since the last segment of each audio file could be shorter than 120 frames.

However, if the shortage is more than half the segment length, the segment is rather discarded. The resultant segments are fed as input to the proposed model. Figure 7 (a) shows the general proposed system.

The two proposed models are implemented using TensorFlow [29] with Keras as a front-end system [30]. To perform that, we used an 11 GB NVIDIA GeForce RTX 2080 Ti GPU to run the proposed CRNN model.

C. ResNet MODEL ARCHITECTURE

The ResNet model receives the speech spectrogram as segments of size $S_s \times 682$, where $S_s$ specifies the size of the segment and 682 is the number of frequency points generated by the short-time FFT, as shown in Figure 7 (a).

Figure 7 (b) shows the architecture of the proposed ResNet model. The kernel size of convolution layers is $3 \times 3$ with stride 1. Each convolutional layer is followed by batch normalization and nonlinear unit (ReLU). The model consists of three blocks. At the first of any block, the feature map size is reduced to half using a convolutional layer with a stride of two. At the same time, the number of filters is doubled. The skip connections are established in every two convolution operations of the plain network. Finally, we add an average pooling layer with a size of eight, followed by a fully connected (FC) layer output.

D. CNN MODEL ARCHITECTURE

Similar to the ResNet model, the CNN model also receives the speech spectrogram as segments of size $S_s \times 682$, as shown in Figure 7 (a). The CNN model consists of three convolutional layers, followed by an FC layer with sigmoid activation, as shown in Figure 7 (C). The three convolutional layers consist of 16, 24, and 32 filters of dimensions $12 \times 16$, $8 \times 12$, and $5 \times 7$ with one-pixel stride, respectively. Each of them is attached by an exponential linear unit (ELU) as a nonlinear activation function and a $2 \times 2$ max-pooling layer with a stride of two. Adding non-linearity using ELU helps the CNNs train much faster. The last convolutional layer (Conv3) is followed by an FC layer with several neurons equal to the number of emotions in the selected corpora. At the end of
A. H. Meftah et al.: KSUEmotions Corpus: Construction, Analysis, Evaluation, and Comparison

The FC layer, a sigmoid function was used to produce the classification outputs in the form of prediction accuracy for different emotions.

The topologies of the ResNet and the CNN models that we propose in this work were tuned experimentally, and by evaluating the performance of each experimented topology, we selected the models reported here.

E. EVALUATION OF SYSTEM ACCURACY

The performance of the automatic recognizer of emotions is evaluated for each audio file. Recall that the system handles each file as a set of segments where each segment is recognized independently from the other segments in the same file. Therefore, the automatic recognizer could predict different emotions for different segments belonging to the same audio file.
file since each one of these segments is processed individually. Thus, the situation where different types of emotions are found in the predictions belonging to a certain file is possible, which contradicts the fact that one file is associated with exactly one type of emotion. To resolve that discrepancy, the overall prediction of a file is reached based on the set of predictions of all segments constituting that file (Figure 8). That is, the output layer of the recognizer model has five outputs corresponding to the five possible labels. The highest output value appears at the label of the predicted emotion given the input of the current segment. Thus, by taking the average of all segments across each emotion label, we have an average score for each emotion in that file. The final overall decision is made by taking the maximum average among those. The example in Figure 8 illustrated the evaluation method. The file in the figure consists of \( n \) segments: \( S_1, S_2, S_3, \ldots, S_n \). Each column shows the five outputs of each input segment. By calculating the average across each emotion (rows), we end up with five averages representing the five emotions (last column). Taking the maximum average value, we decide the emotion of that particular file.

VI. RESULTS AND DISCUSSION
We performed the following three sets of experiments: monolingual, multilingual (merge two corpora KSUEmotions and EPST), and cross-lingual (train on one corpus, test on the other one).

A. MONOLINGUAL EMOTION RECOGNITION
1) KSUEmotions CORPUS
This section proposes the results generated by the designed systems over ten runs with similar system parameters to each corpus separately. Table 9 presents the proposed ResNet system average results of the ten runs, overall accuracy, and standard deviation. ResNet system achieved 85.53% with a standard deviation of 1.87%. As presented in Table 9, the highest recognized emotion is anger followed by neutral and sadness, while the happiness emotion is considered the worst emotion among the five.

Table 10 presents the same average results of ten runs, overall accuracy, and standard deviation for the CNN-based system. The overall accuracy (83.31%) is less than the overall accuracy achieved by the ResNet system with a standard deviation of 4.14%. Similar to the ResNet system, happiness emotion is the worst recognized emotion, while sadness is the best-recognized emotion by 76.93% and 89.66% respectively, as shown in the table.

As we mentioned above, the human perceptual [21] test is our baseline system in addition to the previous study [26], where we applied CRNN. Figure 9 illustrates the average
TABLE 10. Average confusion matrix of ten runs of the proposed CNN system for KSUEmotions.

|          | Neutral | Happiness | Sadness | Surprise | Anger |
|----------|---------|-----------|---------|----------|-------|
| Neutral  | 84.58   | 6.52      | 5.81    | 2.78     | 0.30  |
| Happiness| 4.88    | 76.93     | 0.04    | 14.98    | 3.18  |
| Sadness  | 7.22    | 1.30      | 89.66   | 1.43     | 0.39  |
| Surprise | 2.58    | 11.73     | 1.16    | 78.31    | 6.22  |
| Anger    | 0.68    | 5.03      | 0.00    | 7.88     | 86.42 |

Overall accuracy = 83.31, and STD 4.14

FIGURE 9. Overall accuracy for the emotions recognition for averaged ten runs by ResNet and CNN with recalled results of the human perceptual test and CRNN.

TABLE 11. Average confusion matrix of ten Runs of the proposed ResNet system for EPST corpus.

|          | Neutral | Happiness | Sadness | Anger |
|----------|---------|-----------|---------|-------|
| Neutral  | 61.68   | 12.22     | 10.78   | 15.32 |
| Happiness| 13.49   | 52.50     | 15.40   | 18.61 |
| Sadness  | 27.88   | 16.19     | 54.07   | 1.86  |
| Anger    | 11.75   | 17.16     | 4.13    | 66.96 |

Overall accuracy = 57.66 with an STD of 3.95

TABLE 12. Average confusion matrix of ten Runs of the proposed CNN system for EPST corpus.

|          | Neutral | Happiness | Sadness | Anger |
|----------|---------|-----------|---------|-------|
| Neutral  | 58.63   | 8.77      | 22.72   | 9.89  |
| Happiness| 12.68   | 50.70     | 18.99   | 17.63 |
| Sadness  | 22.35   | 13.66     | 59.42   | 4.56  |
| Anger    | 16.48   | 21.47     | 5.48    | 56.58 |

Overall accuracy = 54.2 with an STD of 5.9

results of ten runs of CRNN (combined of CNN and LSTM), ResNet, and CNN models in addition to the human perceptual test. The ResNet model’s overall accuracy is the nearest to the human perceptual test overall accuracy, followed by CRNN and CNN models.

As shown in Figure 9, the negative emotions (anger and sadness) and neutral are easy to recognize with high accuracy results by the three different models, while the positive emotion (happiness or surprise) is a big challenge.

2) EPST CORPUS

Tables 11 and 12 present the average results of the ten run, overall accuracy, and standard deviation by the proposed systems ResNet and CNN for EPST corpus.

As shown in Tables 11 and 12, the negative emotions (anger and sadness) and neutral are easy to recognize since high accuracy is achieved by both ResNet and CNN, while the positive emotion (happiness) seems not easy to recognize by both systems.
Figure 10 shows the overall accuracy comparison between KSUEmotions and EPST corpora using ResNet and CNN systems for the common emotions. As shown in the figure, the KSUEmotions corpus outperformed the EPST corpus in general. Negative emotions of the two corpora are better identified than the positive emotions except for neutral. From the two tables, the ResNet system has outperformed the CNN system, where the overall accuracy of ResNet is better than CNN. We also observed that CNN’s STD is higher than ResNet’s.

B. MULTILINGUAL AND CROSS-LINGUAL EMOTION RECOGNITION

Regarding the second set of experiments (multilingual), we merged the common emotions of Arabic and English into one data subset. KSUEmotions contain five emotions, while EPST corpus 14 emotions. The neutral, happiness, sadness, and anger emotions are common between the two corpora. Tables 13 and 14 present the average results and overall accuracy for ten-times runs. The results show that the negative emotions (anger or sadness) achieve the highest score, while positive emotions (except neutral) achieve the lowest.

C. CROSS-LINGUAL EMOTION RECOGNITION

Cross-lingual experiments are conducted using only common emotions shared between KSUEmotions and EPST corpora. We use KSUEmotion corpus for the training and EPST for the testing and then vice versa. Table 15 presents the average overall results of the ten runs.
TABLE 15. Cross-lingual accuracy results.

| Training Corpus | Testing Corpus | ResNet | CNN |
|-----------------|----------------|--------|-----|
|                 |                | Accuracy | STD | Accuracy | STD |
| KSUEmotions     | EPST           | 25     | 0.92 | 30.6     | 1.4 |
| EPST            | KSUEmotions    | 32.7   | 2.2  | 36.5     | 4.04 |

As shown in Table 15, the system has failed to recognize the same emotion from another language. It can also be noticed that the accuracy results of the CNN system are slightly higher than those obtained by the ResNet system. These results are expected due to the big contextual differences between the two corpora in terms of language, type of sentences, method of recording, type of microphones, and the recording environment. Therefore, with the purpose to eliminate the effect of those differences between the two corpora, all emotions in the two corpora are mapped into Arousal-Valence space, as shown in Table 16. Then, the ResNet and the CNN models are trained for binary classification of arousal and valence in speech. All emotions presented in the two corpora are mapped into high, low, positive, and negative. Hence, those emotions are transcribed with the new classes (high, low) or (positive, negative). Then, the experiment is carried out considering one complete corpus for training and the other corpus for testing. After that, the two corpora are switched and the experiment is repeated.

Table 17 presents the average overall accuracy of the ten runs of each experiment two cases (arousal-valence mapping is applied in both): (i) for the four common emotions only, and (ii) for all emotions in the two corpora. As shown in this table and compared to Table 15, mapping the emotions into the arousal-valence space improved the system performance. Specifically, the improvement associated with the arousal space mapping is better than that of the valence space. Looking at the effect of training and testing corpora on performance, it is observed that when the models were trained using EPST and tested using KSUEmotions, results are better than the other way around. However, when the two corpora are switched for training and testing, standard deviations over ten runs in each experiment improved. The performance is also affected by the emotions covered in the experiments. That is, the results when we apply the four common emotions are better than when all emotions are included.

Table 18 presents our previous work [31] where the following low-level acoustic features are used: the pitch and intensity (minimum, maximum, range, mean, and standard deviation), the first three formants (minimum, maximum, mean, and standard deviation), the average duration of each emotional state, and the average energy of each emotional state. The system is trained using these features and tested using the proposed dataset. The performance is evaluated by calculating the average accuracy of the ten runs of each experiment. The results show that the proposed acoustic features are effective in recognizing emotions in speech.
mean, and standard deviations), the shimmer, the jitter, and the speech rate. Deep Belief Networks (DBN) and Multi-Layer Perceptron (MLP) are used as classifiers.

Figure 11 presents a comparison between the four models: ResNet and CNN in the current work, and DBN and MLP in [31]. The system accuracy achieve good results with the arousal mapping compared to valance mapping. ResNet and CNN models perform better in the case of the four common emotions than when considering all emotions. On the contrary, the DBN and the MLP achieve better results when all emotions were considered. Also, as mentioned earlier, using EPST for training and KSUEmotions for testing has a positive effect on the performance of the ResNet and the CNN models. In contrast, the DBN and the MLP models achieve better results when trained using the KSUEmotions corpus and tested using EPST. That overall difference between the current and the past results is due to the difference between the features that are extracted in each case.

### VII. CONCLUSION

The researchers in Arabic SER suffering the lack of an Arabic emotional speech corpus. A few significant emotional databases are available, and most of them are not public. KSUEmotions is the first Arabic speech emotion published in the Linguistic Data Consortium by the catalog number LDC2017S12. It was recorded in two phases and contains the following emotions: neutral, happiness, sadness, surprise, and anger. Several ways applied to evaluate the recorded emotion files, such as the human perceptual test and CRNN in our previous study, were applied. In this study, the KSUEmotions corpus construction, analysis, evaluation, and comparison with other published corpus were presented. Besides, we applied ResNet and CNN DNNs models using spectrograms as input features for more verification using monolingual, multilingual, and cross-lingual. The ResNet model’s overall accuracy (85.53%) is the nearest to the overall accuracy of our baseline system human perceptual test (88.3%), followed by CNN models (83.31%), the lowest accuracy in two languages. EPST system results achieved 57.66% and 54.2% by ResNet and CNN, respectively. For the multi-language when mixed between two corpora, the accuracies reached 80.28% and 75.6% by ResNet and CNN models, respectively. Regarding the emotions classes, the negative emotions anger and sadness for Arabic

| Training Set | Testing Set | Arousal | Valence | Arousal | Valence | Arousal | Valence | Arousal | Valence |
|--------------|-------------|---------|---------|---------|---------|---------|---------|---------|---------|
| KSUEmotions  | EPST        | 57.2    | 53.22   | 53.66   | 49      | 67.03   | 51.38   | 65.71   | 52.92   |
| EPST         | KSUEmotions | 42.52   | 54.63   | 48.62   | 50.29   | 62.28   | 50.1    | 61.7    | 53.09   |

### TABLE 18. Accuracy of cross-lingual emotion recognition using different features as in past work [31].
language (KSUEmotions) or English language (EPST corpus) achieved the highest accuracy results, while the positive emotion happiness or surprise achieved the lowest.

REFERENCES

[1] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: A review,” Int. J. Speech Technol., vol. 15, no. 2, pp. 99–117, Jan. 2012.

[2] E. Navas, A. Castelruiu, I. Luengo, J. Sánchez, and I. Hernández, “Designing and recording an audiovisual database of emotional speech in basque,” in Proc. LREC, May 2004, pp. 1387–1390.

[3] J. B. M. Liberman, K. Davis, M. Grossman, and N. Martey. (2002). Emotional Prosody Speech and Transcripts, Linguistic Data Consortium. Philadelphia, PA, USA. Accessed: Mar. 12, 2015. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2002S28

[4] E. M. Caldognetto, P. Cosi, C. Drioli, G. Tisato, and F. Cavicchio, “Modifications of phonetic labial targets in emotive speech: Effects of the co-production of speech and emotions,” Speech Commun., vol. 44, nos. 1–4, pp. 173–185, Oct. 2004.

[5] A. Batliner, M. Blomber, S. D’Arcy, D. Elenius, G. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, “The PF_STAR children’s speech corpus,” in Proc. INTERSPEECH, 2005, pp. 2761–2764.

[6] M. El Ayadi, S. M. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011.

[7] D. Ververidis and C. Kotropoulos, “A review of emotional speech databases,” in Proc. Panhellenic Conf. Informat. (PCI), Nov. 2003, pp. 560–574.

[8] C.-N. Anagnostopoulo, T. Ilou, and D. Giannikous, “Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011,” Artif. Intell. Rev., vol. 43, no. 2, pp. 155–177, Feb. 2015.

[9] K. Saitnunaz, M. Dhalwal, J. Rokne, and R. Alhaji, “Emotion detection from text and speech: A survey,” Soc. Netw. Anal. Mining, vol. 8, no. 1, p. 28, 2018

[10] K. S. Rao and S. G. Koolagudi, Emotion Recognition Using Speech Features (SpringerBriefs in Speech Technology), 1st ed. New York, NY, USA: Springer-Verlag, 2013, doi: 10.1007/978-1-4614-5143-3.

[11] S. Kasuriya, T. Theeramunkong, C. Wutiwiwatchai, and P. Sukhummek, “Emotional speech recognition research: An analysis of research focus,” Int. J. Speech Technol., vol. 53, no. 1, pp. 17–55, Mar. 2019.

[12] A. Meftah, Y. Seddiq, A. Alotaibi, and S.-A. Selouani, “Cross-corpus Arabic and English emotion recognition,” in Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT), Dec. 2017, pp. 377–381.

[13] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Commun., vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[14] M. Swain, A. Routray, and K. Babishapathy, “Databases, features and classifiers for speech emotion recognition: A review,” Int. J. Speech Technol., vol. 21, no. 1, pp. 93–120, Mar. 2018.

[15] B. Schuller, B. Vlašenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, “Cross-corpus acoustic emotion recognition: Variances and strategies,” IEEE Trans. Affect. Comput., vol. 7, no. 3, pp. 2665–2669, 2016.

[16] A. Al-Faham and N. Ghneim, “Towards enhanced Arabic speech emotion recognition: Comparison between three methodologies,” Asian J. Sci. Technol., vol. 7, no. 3, pp. 2665–2669, 2016.

[17] A. H. Meftah, Y. A. Alotaibi, and S.-A. Selouani, “Evaluation of an Arabic speech corpus of emotions: A perceptual and statistical analysis,” IEEE Access, vol. 6, pp. 72845–72861, 2018.

[18] "KTD corpus," King Abdulaziz City Sci. Technol., Tech. Rep.

[19] "TTD corpus," King Abdulaziz City Sci. Technol., Tech. Rep.

[20] "speech emotion recognition using deep learning techniques: A review," IEEE Access, vol. 7, pp. 117327–117345, 2019.

[21] A. Al-Faham and N. Ghneim, “Towards enhanced Arabic speech emotion recognition: Comparison between three methodologies,” Asian J. Sci. Technol., vol. 7, no. 3, pp. 2665–2669, 2016.

[22] A. Wendemuth, and G. Rigoll, “Cross-corpus acoustic emotion recognition: Variances and strategies,” IEEE Trans. Affect. Comput., vol. 7, no. 3, pp. 2665–2669, 2016.

[23] P. Boersma and D. Weenink. (2021). Praat: Doing Phonetics by Computer [Computer Program]. Version 6.1.41. Accessed: Mar. 25, 2021. [Online]. Available: http://www.praat.org/

[24] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, “Speech emotion recognition using deep learning techniques: A review,” IEEE Access, vol. 7, pp. 117327–117345, 2019.

[25] S. Wu, S. Zhong, and Y. Liu, “Deep residual learning for image steganalysis,” Multimedia Tools Appl., vol. 77, no. 9, pp. 10437–10453, 2017.

[26] M. A. Qamhan, A. H. Meftah, S.-A. Selouani, Y. A. Alotaibi, M. Zakariah, and Y. M. Seddigh, “Speech emotion recognition using convolutional recurrent neural networks and spectrograms,” in Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE), Aug. 2020, pp. 1–5.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[28] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Dec. 2014, arXiv:1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980

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

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

[31] A. Meftah, Y. Seddiq, Y. Alotaibi, and S.-A. Selouani, “Cross-corpus Arabic and English emotion recognition,” in Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT), Dec. 2017, pp. 377–381.

[32] A. Al-Hamid Meftah et al. received the B.Sc. and M.Sc. degrees in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 2009 and 2015, respectively, where he is currently pursuing the Ph.D. degree. Since 2010, he has been a Researcher with King Saud University. His research interests include digital speech processing specifically speech recognition and Arabic language and speech processing.

[33] A. H. Meftah et al. 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 Telecommunications Company (PTC), Yemen. He is currently pursuing the Ph.D. degree. Since 2010, he has been a Researcher with King Saud University. His main research interests include digital signal processing, speech processing, and artificial intelligence.
Yasser 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. degree in computer engineering from King Saud University (KSU), Riyadh, Saudi Arabia, in 2010, where he is currently pursuing the Ph.D. degree in computer engineering. Since 2004, he has been a Digital Systems Designer with King Abdullah City for Science and Technology (KACST), Riyadh. His research interests include digital signal processing, speech processing, image processing, computer arithmetic, and digital systems design using FPGA.

Yousef A. Alotaibi (Senior Member, IEEE) received the B.Sc. degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 1988, and the M.Sc. and Ph.D. degrees in computer engineering from the Florida Institute of Technology, Melbourne, FL, USA, in 1994 and 1997, respectively. From 1988 to 1992 and from 1998 to 1999, he joined Al-ELM Research and Development Corporation, Riyadh, as Research Engineer. From 1999 to 2008, he joined the College of Computer and Information Sciences, King Saud University, as an Assistant Professor, and from 2008 to 2012, he was an Associate Professor with the College of Computer and Information Sciences, King Saud University. 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.

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. He joined the Algerian-French Double Degree Program, Université Joseph Fourier of Grenoble, in 2000. 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.

---

VOLUME 9, 2021 54219