Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus

Ali Meftah, Mustafa Qamhan, Yousef A. Alotaibi, Mohammed Zakariah

College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia.

Email: {ameftah, mqamhan, yaalotaibi, mzakariah}@ksu.edu.sa

***Abstract* - Speech is an active source to prompt emotions and attitudes through language. Discovering the emotional content within a speech signal and then recognizing the type of emotion developed from the uttered speech is a significant task for researchers. In this paper, we present a study of emotion recognition in Arabic speech using the KSUEmotions Arabic speech emotion corpus by applying feature-extraction techniques followed by classification techniques like the K-Nearest Neighbor algorithm (KNN) and Support Vector Machine (SVM). The experiments were performed using the Python programming language. The experiments revealed that KNN is better than SVM for this corpus, and the results of the experiment show the highest accuracy for the emotion of sadness, followed by happiness and then by surprise.**

Keywords—speech processing, speech emotion recognition, Arabic, KNN, SVM

I. INTRODUCTION

Human speech is considered the most natural, fastest, and easiest means of communication. Speech is a complex signal comprising data about the message, the speaker, and his/her emotion while speaking. Human-Computer Interaction (HCI) [1] plays an important role in understanding and conveying each other's purposes more naturally. The main task that HCI accomplishes is to develop the capability to recognize the emotion of the speaker very precisely, which is usually very similar to the capability of human-robot interaction [2]. Based on the fastest and easiest method of communication, speech signals are regarded as the means to identify the emotion of the speaker as well [3]. Based on this [4], it is supposed that speech signals can not only deliver the syntactic and semantic content of the statements but also are capable of revealing the emotional state of the human. Thus, the recognition of the emotional state of a human being is possible with the help of a speech signal, by studying the physical state of the human being automatically from his/her speech [5]. The most common issue in the recognition of emotion from speech signals is the selection of optimal features set from the signal [6]. The majority of the past work on speech emotion recognition (SER) has been dedicated to the analysis of speech prosodic features and spectral information [7]. However, some of the other works in this field consider novel feature parameters and Fourier parameters for SER [8]. As stated in [9], automatic SER is very dynamic research is in HCI. Speech emotion recognition has many applications, some of which are the analysis of mental illness, in-car systems to assess the physical condition of the driver for the safety and security of the passengers, in intelligent toys, and in call centers for the detection of lie [3], [10]. Today, emotion recognition through speech signals is of vast importance. Many studies on this topic have already been performed for several languages, including English, Spanish, Slovenian, French, and German. However, very few works have been reported on the Arabic language. Very recently, the techniques applied for the automatic recognition of emotions from speech signals have matured significantly, especially for real-life scenarios [11], [12], [13], such as call centers [14], [15], supplementary disease analysis [16], [17], [18], and distant education [19], [20]. Nevertheless, current SER technology has not attained very good results, perhaps because of the deficiency of real emotion-related features. To solve this problem, we explore the effect of prominence features in SER. The most challenging task in SER is the extraction of the features which are related to the emotions. Because of the shortage of discriminative acoustic features, traditional methods based on traditional acoustic features could not deliver reasonable results. The main motivation for this work is to explore emotion recognition in Arabic speech, which has not been studied so far. Arabic is a very rich language that is spoken in most Arab countries in the world, and this work would open doors for other researchers willing to work on the Arabic language.

This paper is organized as follows. In Section II, we discuss the past work in this field and the results. In Section III, we discuss the details of the corpus used in this experiment, and in Section IV we discuss the experimental framework, followed by Section V with the results and discussion. In Section VI, the conclusion is provided along with the scope of future work.

II. LITERATURE REVIEW

This section provides a review of various SER techniques found in the literature. Many of them are applied based on the traditional classification method with a difference in feature vectors to achieve better results for the recognition of emotion. A Support Vector Machine (SVM) was utilized in [5], [12], [13] and a hierarchical classifier was utilized in [6], [11]. Few researchers have brought

various approaches to the traditional classifier for the classification of emotion. The hidden Markov model was also applied for the recognition of emotion [21], as well as Gaussian mixtures model [22], SVM [23], artificial neural network [24], K-nearest neighbor [25]. Among the above, the most widely used learning algorithms are SVM and HMM for speech-related applications [26], [27]. However, the experimental results show that the accuracy of each classifier is dependent on the domain and the quality of the data. Most of the experiments are based on a single classifier, but some systems have implemented multiple classifiers to improve the accuracy of the SER [28], which is called deep-feature-based SER for smart effective services. The most important task in emotion detection is the extraction of appropriate features and the selection of a good classifier to determine the exact emotion. Because there is no fixed formula to select the classifier, it all depends on the geometry of the input vector. There are various types of classifiers for the recognition of speech emotion systems, including Artificial Neural Network (ANN), Gaussian Mixtures Model (GMM), Hidden Markov Model (HMM), SVM, K-Nearest Neighbor (KNN), which are some of the most extensively used classifiers in emotion recognition systems. These classifiers have their advantages and disadvantages over each other. It has been said that the human mind can barely recognize the emotion from speech, with a success rate of up to 60% for unknown speakers, whereas the researchers could reach up to 99% accuracy in speaker-independent speech.

In previous research, the results obtained were as follows. For emotion recognition, the accuracy with HMM concerning the speaker-dependent classification reached 76.12%, and concerning speaker-independent classification, the accuracy reached 64.77%. Additionally, the accuracy rate for speaker-dependent GMM was about 89.12%, whereas that for the speaker-independent case reached 75%. Similar experiments using ANN were conducted that yielded an accuracy of about 52.87% for the speaker-independent case, which is considerably less compared to that of other classification techniques, and of about 51.19% for speaker-dependent classification Further classification experiments were conducted using the KNN classification technique, which had an accuracy rate of 64% for four different emotional states using feature vectors like energy contours and pitch [1], [2], [3], [5] [14].

III. DATASET

The KSUEmotions corpus [29] was created for Modern Standard Arabic (MSA) using 23 speakers (10 males and 13 females) from three Arabic countries: Yemen, Saudi Arabia, and Syria. The recording took place in two phases. The total number of files are shown in the TABLE I. In Phase 1, 10 male speakers were selected from Saudi Arabia, Yemen, and Syria, and 10 female speakers were selected from Saudi Arabia and Syria. All speakers read the 16 MSA sentences selected from the original corpus, King Abdulaziz City for Science and Technology Text-To-Speech Database (KTD) [30]. In this phase, neutral, sadness, happiness, surprise, and questioning emotions were selected. The following are the codes used for the emotions in the experiments: Neutral (E00), Happiness (E01), Sadness (E02), Surprise (E03), and Anger (E05). The questioning was considered an emotion because it was incorporated in the corpus originally used (KTD). To evaluate the Phase 1 recordings, a blind human perceptual test was performed. Nine listeners (6 males and 3 females) were involved to listen to the recorded files to determine whether they were able to recognize the recorded emotion. According to the results of the human perceptual test, and by avoiding defective speakers and/or files and to ensure uniformity among different variables, such as the speakers' gender, Phase 2 was produced.

TABLE I. STATISTICAL DETAILS OF THE KSUEMOTION CORPUS

		Total No. of files
Male Speakers	**Phase 1**	800
	Phase 2	840
	All	1640
Female Speakers	**Phase 1**	800
	Phase 2	840
	All	1640
Phase 1		1600
Phase 2		1680
Phase 1+ Phase 2		3280

Seven male speakers from Phase 1 and 7 female speakers (4 from Phase 1 and 3 new female speakers from Yemen) and 10 sentences were chosen for Phase 2. In this phase, the questioning emotion was excluded and the anger emotion was added for consistency with other similar corpora in the field (e.g., [31]). In Phase 2, each sentence was spoken over two trials. The total duration of all recorded files was 2 h and 55 min for Phase 1 and 2 h and 15 min for Phase 2. Again, a blind human perceptual test was performed for Phase 2 with the same nine listeners who reviewed Phase 1. PRAAT software [32] was used for the KSUEmotions corpus recording process.

IV. EXPERIMENTAL FRAMEWORK

A. Data Preparation

During the first experiment, only Phase 1 of the KSUEmotions corpus was taken for both training and testing. In Experiment 2, only Phase 2 of the corpus was taken for training and testing. In experiment 3, the full Phase 1 data from the KSUEmotions corpus was taken for training and that from Phase 2 was taken for testing. In experiment 4, Phase 2 of the corpus was used for training and Phase 1 was used for testing. In the final experiment, both Phase 1 and

Phase 2 of the corpus were mixed and then partitioned into training and testing. For all the experiments, both classifiers SVM and KNN were applied and the results were matched. Confusion matrix tables were also developed. Figure 1 shows the experiment flowchart.

B. Features Used and Classifiers

The following are the feature extraction techniques applied to the datasets, zero-crossing rate, short-term energy, MFCC's and delta features. The parameters applied for these features are short_term_window = 0.036, short_term_step = 0.012, mid_term_window = 1.3, mid_term_step = 0.65, perc_train = 0.75. after features were extracted with the above parameter setting then the experiments were conducted with SVM and KNN classifiers. Python programming language was used for the implementation of this work.

E. Experiments

Five experiments were conducted as listed in Table II.

TABLE II. EMOTION RECOGNITION RESULTS IN KSUEMOTIONS CORPUS (PH1: PHASE 1, PH2: PHASE 2)

Experiments	Training Subset	Testing Subset	Results	
			SVM	KNN
1	PH1	PH1	67.92%	69.38%
2	PH2	PH2	78.96%	87.04%
3	PH1	PH2	59.5%	53.57%
4	PH2	PH1	48.51%	42.71%
5	PH1&PH2 Training (75%)	PH1&PH2 Testing (25%)	68.75%	75.49%

Experiment 1: During this experiment, the dataset used was the KSU Emotions corpus, from which Phase 1 was selected for both training and testing with a ratio of 70% to 30%, respectively. After the features were extracted using the above discussed feature-extraction technique, these features were applied for the classification techniques. Both SVM and KNN were applied.

***Experiment 2*:** Phase 2 dataset from the KSUEmotions corpus taken for both training and testing with a ratio of 70% to 30%, respectively. The features extracted were applied to the SVM and KNN classifiers.

***Experiment 3*:** The full Phase 1 dataset was used for training and the full Phase 2 dataset was used for testing. The SVM and KNN classifiers were applied for the classification of emotions.

***Experiment 4*:** The full Phase 2 dataset was used for training and the full Phase 1 dataset was used for testing with the application of both the SVM and KNN classifiers for classification.

Experiment 5: In this experiment, the Phase 1 and Phase 2 datasets were mixed and then distributed among the training and testing datasets with a ratio of 70% to 30%, respectively.

V. RESULTS

The following are the results of the experiments. During experiment 1, the entire set of Phase 1 data is partitioned with 70% for training and 30% for testing, but this dataset lacks the emotion of anger (E05). The accuracy was 69.38% using the KNN technique, whereas it was only 67.92% using the SVM technique.

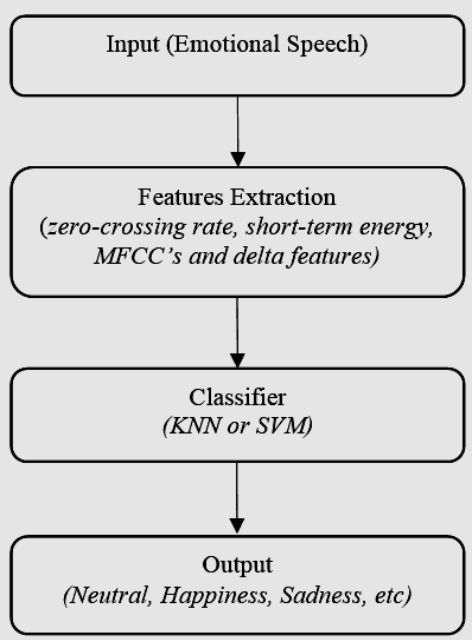

Fig. 1 Flowchart of the experiment.

This was followed by experiment 2, wherein the entire Phase 2 dataset is applied and partitioned among training and testing with the same ratio. The outcome of the experiment, as shown in Table II, is 87.04% accuracy using KNN and 78.96% using SVM. Experiment 3 was conducted by taking the entire Phase 1 and Phase 2 datasets as training and testing, respectively, but to maintain the stability in the datasets, the files with only similar emotions among Phase 1 and Phase 2 were taken and the rest were excluded. The

Phase 1 question emotion is excluded in the training, as it was not present in Phase 1, and for Phase 2, anger is excluded as it was not present in Phase 1, to maintain similar emotions for the experiment. The results after applying SVM and KNN are shown in Table II. The accuracy of SVM and KNN was 59% and 53%, respectively. Similarly, in experiment 4, the phases are shuffled as they were applied in Experiment 3 with Phase 2 for training and Phase 1 for testing. Dissimilar emotions were excluded. The outcome of the experiment, as shown in Table II, is 48.51% accuracy for SVM and 42.71% for KNN. Finally, in experiment 5, both

Phases are mixed and then partitioned into training and testing. Phase 1 and Phase 2 datasets were mixed and partitioned with 70% of the data going into training and the remaining 30% to testing. The outcome of the experiment, as shown in Table II, is 68.75% for SVM and 75.49% for KNN.

The confusion matrix for Experiments 1, 2, and 5 are displayed in tabular format. For experiment 1, the confusion matrix is shown in Table III, where the accuracy of 75.83% is the highest for sadness. For experiment 2 the confusion matrix is shown in

TABLE III. PH1 CONFUSION MATRIX USING SVM MODEL

(E00: Neutral, E01: Happiness, E02: Sadness,and E03: Surprise)

Emotions	E00	E03	E01	E02
E00	**64.17**	10.83	10.83	14.17
E03	7.5	**61.67**	22.5	8.33
E01	15.0	11.67	**70.0**	3.33
E02	18.33	5.0	0.83	**75.83**

TABE IV. PH2 SINGLE CONFUSION MATRIX USING KNN MODEL (E05: ANGER)

Emotions	E00	E05	E03	E01	E02
E00	**88.51**	2.3	0.0	5.75	3.45
E05	1.15	**89.66**	5.75	2.3	1.15
E03	3.45	4.6	**80.46**	10.34	1.15
E01	9.2	5.75	8.05	**77.01**	0.0
E02	0.0	0.0	0.0	0.0	**100.0**

Table IV with the highest accuracy of about 100%, which also goes to the emotion of sadness. Finally, concerning experiment 5 with the mixture of both Phase 1 and Phase 2 datasets, the highest accuracy shown in Table V also goes to the emotion of sadness. For all the experiments, the highest accuracy is achieved for detecting the emotion of sadness.

TABLE V. PH1&PH2 CONFUSION MATRIX USING KNN MODEL

Emotions	E00	E03	E01	E02
E00	**80.92**	10.53	4.61	3.95
E03	9.87	**69.74**	13.82	6.58
E01	12.5	19.08	**67.11**	1.32
E02	13.82	0.66	1.32	**84.21**

VI. CONCLUSION

In this work, we recognized emotion within Arabic speech data. The Arabic speech corpus was developed at King Saud University. We have constructed this speech emotion corpus and verified its reliability by conducting different experiments. Various feature-extraction techniques were applied to the dataset, including the zero-crossing rate, short-term energy, MFCCs, and delta feature, and then were followed by classifiers like KNN and SVM with the tuning of the parameters. Finally, we discovered that the emotion of sadness is recognized with the highest accuracy. We have compared the results with two classifiers: SVM and KNN. The results achieved are the highest for sadness, followed by surprise. The Python programming language was used to implement this work. KNN has shown better accuracy than SVM for this corpus. Phase 2 of the corpus is better than Phase 1. In the future, we would like to implement deep learning for this corpus to recognize emotions.

ACKNOWLEDGMENT

This project was funded by the National Plan for Science, Technology, and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (11-INF1968-02).

REFERENCES

[1] M. Song, M. You, N. Li, and C. Chen, "A robust multimodal approach for emotion recognition," Neurocomputing, vol. 71, no. 10–12, pp. 1913–1920, 2008.

[2] P. Salovey and J. D. Mayer, "Emotional Intelligence," Imagin. Cogn. Pers., vol. 9, no. 3, pp. 185–211, 1990.

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

[4] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artif. Intell. Rev., pp. 1–23, 2012.

[5] G. R. Javier, "Speech emotion recognition in emotional feedback for Human-Robot Interaction," Int. J. Adv. Res. Artif. Intell., vol. 4, no. 2, pp. 20–27, 2015.

[6] S. Steidl et al., "The INTERSPEECH 2010 Paralinguistic Challenge," INTERSPEECH 2010, no. September, pp. 2794–2797, 2010.

[7] A. Ingale and D. Chaudhari, "Speech Emotion Recognition," Int'l J. Soft Comput. Eng., vol. 2, no. 1, pp. 235–238, 2012.

[8] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 69–75, 2015.

[9] S. A. Firoz, S. A. Raji, and A. P. Babu, "Automatic Emotion Recognition from Speech Using Artificial Neural Networks with Gender-Dependent Databases," in International Conference on Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09., 2009, pp. 162–164.

[10] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using Support Vector Machines," in 2013 5th International Conference on Knowledge and Smart Technology (KST), 2013, pp. 86–91.

[11] E. Martinelli, A. Mencattini, E. Daprati, and C. Di Natale, "Strength is in numbers: Can concordant artificial listeners improve prediction of emotion from speech?," PLoS One, vol. 11, no. 8, 2016.

[12] W. J. Han, H. F. Li, H. Bin Ruan, and L. Ma, "Review on speech emotion recognition," Ruan Jian Xue Bao/Journal Softw., vol. 25, no. 1, pp. 37–50, 2014.
[13] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Multimodal Approaches for Emotion Recognition: A Survey," Proc. SPIE – Int. Soc. Opt. Eng., pp. 56–67, 2005.
[14] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process. A Rev. J., vol. 22, no. 6, pp. 1154–1160, 2012.
[15] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, 2005.
[16] Y. T. Fan and Y. Cheng, "Atypical mismatch negativity in response to emotional voices in people with autism spectrum conditions," PLoS One, vol. 9, no. 7, 2014.
[17] E. Bal, E. Harden, D. Lamb, A. V Van Hecke, J. W. Denver, and S. W. Porges, "Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state," J. Autism Dev. Disord., vol. 40, no. 3, pp. 358–370, 2010.
[18] D. J. France and R. G. Shiavi, "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE Trans. Biomed. Eng., vol. 47, no. 7, pp. 829–837, 2000.
[19] K. Bahreini, R. Nadolski, and W. Westera, "Towards real-time speech emotion recognition for affective e-learning," Educ. Inf. Technol., vol. 21, no. 5, pp. 1367–1386, 2016.
[20] W. Wang and J. Wu, "Emotion recognition based on CSO&SVM in e-learning," in 2011 Seventh International Conference on Natural Computation, 2011, vol. 1, pp. 566–570.
[21] R. Raman, P. K. Sa, B. Majhi, and S. Bakshi, "Direction Estimation for Pedestrian Monitoring System in Smart Cities: An HMM Based Approach," IEEE Access, vol. 4, pp. 5788–5808, 2016.
[22] S. Yun and C. D. Yoo, "Loss-scaled large-margin Gaussian mixture models for speech emotion classification," IEEE Trans. Audio, Speech Lang. Process., vol. 20, no. 2, pp. 585–598, 2012.
[23] C. M. Bishop, Pattern Recognition and Machine Learning, vol. 4, no. 4. Springer, 2006.
[24] D. Gharavian, M. Sheikhan, A. Nazerieh, and S. Garoucy, "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network," Neural Comput. Appl., vol. 21, no. 8, pp. 2115–2126, 2012.
[25] K. M. Eisenhardt, "Building Theories from Case Study Research.," Acad. Manag. Rev., vol. 14, no. 4, pp. 532–550, 1989.
[26] J. P. A. Ioannidis, "Why most published research findings are false," PLoS Medicine, vol. 2, no. 8. pp. 0696–0701, 2005.
[27] F. Chang et al., "Bigtable: A distributed storage system for structured data," 7th Symp. Oper. Syst. Des. Implement. (OSDI '06), Novemb. 6-8, Seattle, WA, USA, pp. 205–218, 2006.
[28] U. Alon, "How to choose a good scientific problem.," Mol. Cell, vol. 35, no. 6, pp. 726–728, 2009.
[29] A. H. Meftah, Y. A. Alotaibi, and S.-A. Selouani, "King Saud University Emotional speech corpus (KSUEmotions)," Linguistic Data Consortium (LDC), 2017. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2017S12.
[30] King Abdulaziz City for Science and Technology (KACST), "KTD Corpus," Unpubl. Tech. Report.
[31] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," Linguist. Data Consortium, Philadelphia, no. LDC2002S28, 2002.