Arabic Automatic Speech Recognition Based on Emotion Detection

Engy R. Rady*[1], M. Hesham**[2], N.M. Hassan*[3], A. Hassen***[4]

[*] *Basic Science Department, Faculty of Computers and Information, Fayoum University, El Fayoum, Egypt.*

[1]era00@fayoum.edu.eg

[3]nmh00@fayoum.edu.eg

[**] *Engineering Math & Physics Department, Faculty of Engineering, Cairo University, Giza, Egypt.*

[2]mhesham@eng.cu.edu.eg

[***] *Physics Department, Faculty of Sciences, Fayoum University, El Fayoum, Egypt.*

[4]ash02@fayoum.edu.eg

Abstract: ***This work presents a novel emotion recognition via automatic speech recognition (ASR) using a deep feed-forward neural network (DFFNN) for Arabic speech. We present results for the recognition of the three emotions happy, angry, and surprised. The Arabic natural audio dataset (ANAD) is used. Twenty-five low-level descriptors (LLDs) are extracted from the audio signals. Different combination of extracted features is examined. Also, the effect of using the principal component analysis (PCA) technique for dimensionality reduction is examined. For the classification stage, DFFNN is used. Also, the problem of imbalances samples in the dataset is managed by using the borderline-synthetic minority over-sampling technique (B-SMOTE). It is shown from the results that the best accuracy is obtained when applying PCA on the extracted features is 98.56 %. Also, the accuracy is 98.33 % when using the combination of all the extracted features. This result is not too much different from the accuracy of using PCA. It is followed by the accuracy of using MFCC and LSF which is 97.79 %. It is noticed that the accuracy is 95.63 % when using LSF features which shows that they are dominant features. The obtained results showed an improvement compared to previous studies.***

Keywords: ***Arabic speech emotion recognition, Arabic natural audio dataset, deep feed-forward neural network, borderline-synthetic minority over-sampling technique, PCA***

1 INTRODUCTION

Emotion is a complex psychological state that governs our daily lives. Understanding emotion and explore how people react enriches the interaction [1]. To provide better natural Human-machine interaction, the machine should be trained to understand emotion as well as speech. Automatic speech emotion recognition (ASER) is the process of recognizing human emotion from speech. ASER is a portion of the field of ASR that is increasing enormously in recent years. ASER has a wide range of applications: health monitoring, customer feedback that could help voice agents, call centers, banking, virtual reality (VR), and E-learning.

Emotion is detected by several features, including acoustic features from speech, body gestures, biosignals, and facial expressions. In our study, we work on the features of emotion in speech. We focus on identifying the best speech features and models for recognizing three emotional states from Arabic speech: happy, angry, and surprised.

The research activities on ASER face many challenges like the various dialects, the complexity of phonetic rules of the Arabic language. Also, the lack of resources for Arabic speech emotion recognition is a vital problem facing this field [2].

Emotions are represented in two approaches[3]: discrete emotional approach: classifying emotions in discrete labels like happiness, boredom, anger surprised, etc. and dimensional emotional approach: representing emotions with dimensions such as arousal (describe the strength of the emotion), valence (emotion is positive or negative), and power (describe the strength or weakness of the person). In the present study, we work on a discrete emotional approach.

In general, the ASER system consisting of two major parts: feature extraction and classification. Feature extraction is the process of extracting the most effective and dominant characteristics that well represent the given speech. Several studies have proposed acoustic features from a speech which well suited the emotional information from speech, such as pitch, energy, formant frequency [4], linear prediction cepstrum coefficients (LPCC) [5], LSP, MFCC [6]–[9], and Mel filter bank (MFB) [10]. Most researchers prefer to combine more than one type of feature set that is containing more emotional information.

The early classification algorithms for ASER based on the conventional machine learning algorithms such as k-Nearest Neighbor (KNN) [11], support vector machine (SVM) [12], artificial neural network (ANN) [13], Gaussian Mixture Model (GMM) [14], and Hidden Markov Model (HMM) [15] have been used by many researchers. Recently classification methods based on deep learning algorithms have been used in ASER. Deep learning techniques like deep belief networks

(DBN) [16], recurrent neural networks (RNN) [17], convolutional neural networks (CNN) [18]–[20], deep feed-forward neural networks (DFFNN) [21], [22] are used for many applications like computer vision, pattern recognition, ASR, natural language processing, image recognition as well as ASER.
This paper aimed to describe the proposed model, including the corpus, feature extraction, and the recognition process. Besides, it summarizes some of the works related to emotion recognition in speech. The outcome results are discussed, analyzed, and compared to previous studies.

2 RELATED WORK

In [23], the speech rate, pitch, formants, and intensity acoustic features are extracted from the KSUEmotions corpus. Three sentences spoken by eight native Arabic speakers (4 male and 4 female) are selected from the corpus. Two evaluation processes are performed Perception using human listeners and emotion recognition. The perception evaluation results in scores of 87% for males, 84% for females, and 85% for both. A Multilayer Perception (MLP) Classifier based on a backpropagation algorithm is used to classify 5 emotions (neutral, happy, sad, surprise, anger). The highest results for the emotion recognition evaluation are 83%, 56%, and 78% for males, females, and both, respectively.
In [24], a speech emotion recognition is studied in Arabic spoken data for the first time. The authors collected a speech corpus (ANAD) from Arabic TV shows. They labeled the videos by their perceived emotions; namely happy, surprised, or angry. The low-level descriptors include: zero crossing rates (ZCR), intensity, MFCC (1–12), F0 (Fundamental frequency), LSP, F0 envelope, and the probability of voicing are extracted from speech, and thirty-five classification methods include Sequential Mean Optimization (SMO), Random Forest, Simple Logistic, Logistic Model Trees, Attribute Selected with J48, Random Sub Space with RepTree, Multiclass Classifier Updatable , Random Committee with RandomTree, Bagging with RepTree, Logit Boost performs additive logistic regression to Decision Stump, Filtered Classifier with J48, Iterative Classifier Optimizer, Classification via regression, K-Nearest Neighbour, PART, RepTree, JRIP, J48, Multiclass classifier, Decision Table, Random Tree, Logistic, Adaptive Boosting for Decision Stump, OneR, CVParameterSelection with ZeroR, Bayes Net, Hoeffding Tree, Naive Bayes, Naive Bayes Updatable, Naive Bayes Multinomial, Decision Stump, Randomizable Filtered Classifier with K-nearest neighbour, ZeroR, Weighted Instances Handler Wrapper with ZeroR, InputMapped Classifier with ZeroR are applied. Results showed that the SMO method gives the highest classification accuracy of 95.52 %.
In [25], the author performed an ASER system. A dataset called EYASE has been created. It is a semi-natural from an Egyptian TV series. The EYASE dataset contains utterances from six professional actors (3 females and 3 males) for four emotions: sad, angry, happy, and neutral. Spectral, prosodic, wavelet features, and long-term average spectrum (LTAS) are extracted from the utterances. Two experiments are performed speaker-dependent and speaker-independent for three cases: (1) multi-emotion classifications, (2) neutral versus emotion classifications, (3) valence and arousal classifications.
MFCCs and modulation spectral (MS) are extracted from speech signals in [18]. Seven emotions: neutral, happy, sad, calm, fearful, surprise, and disgust are classified using a decision tree, SVM, random forest, and CNN. CNN has shown the highest accuracy 78.20% for recognizing emotions from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDEES) dataset.
In [20], the speech signals are transformed into a 2-D representation based on Short Time Fourier Transform (STFT). For classification, CNNs, Long Short-Term Memory (LSTM), and time distributed CNNs architectures are used. The public dataset Berlin is used in this work. It contains 535 German utterances of 10 different statements by ten actors conveying seven emotions (Anger, Neutral, Happy, Fear, Disgust, Sadness, and Boredom). The time distributed CNNs networks show the highest accuracy of 86.65 %.
In [26], various feature extraction techniques include: log-Mel Spectrogram, MFCCs, pitch, and energy were considered. The extracted features were compared by applying LSTM, CNNs, HMMs, and Deep Neural Networks (DNNs). The RAVDESS dataset includes the 14-class (2 genders x 7 emotions) are used in this work. The highest accuracy achieved with four layers 2-D CNN using the Log-Mel Spectrogram features is 68%.
In [27] a new discrete wavelet transform (DWT) feature set is extracted from speech. The English dataset RAVDESS is used. An artificial neural network (ANN) with 77 nodes as input and 100 nodes in the hidden layer is used for classifying seven emotions (surprise, fear, sad, clam, happy, angry, and disgust). Two emotion labels are considered at a time as a binary classification. 10-fold cross-validation is used to test the accuracy of the model. The results are compared with other classification methods support vector classifier (SVC), Gaussian naive bayes (GNB), and KNN. Average accuracy of more than 90 % and 80 % is achieved. The proposed ANN model shows better performance than the other techniques.

3 MODEL SETUP

In this work, the influences of the use of different combinations of the extracted features for the ANAD Arabic dataset are examined, then classified the three emotions using DFFNN. The main proposed model architecture for our work consists of 4 modules depicted in Fig. 1. The first module is the feature extraction process. It is followed by the over-sampling algorithm, then the classification and evaluation. Also, as the used dataset is imbalanced (the numbers of instances in the three classes are not equally represented) as shown in Fig. 2, we used the B-SMOTE [28] over-sampling technique. It is an over-sampling technique that synthetically augments samples of minority classes based on the borderline between the majority and minority classes. It has the same concept of SMOTE that is the nearest neighbors of the same class are calculated for every minority sample, then new synthetic samples are generated along the line between the minority sample and its chosen nearest neighbors. The only difference between the B-SMOTE and SMOTE is that the B-SMOTE only synthetic samples along the borderline of the minority classes.

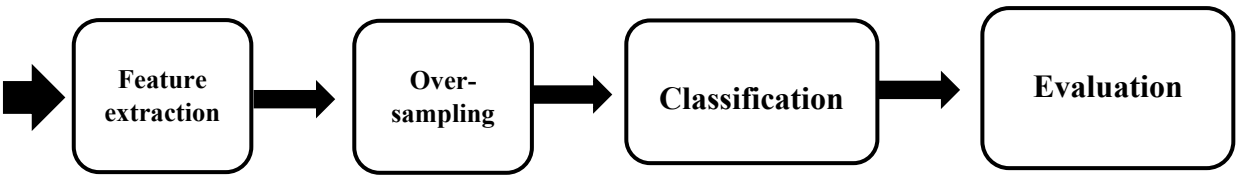


Figure 1: Block diagram of the proposed model

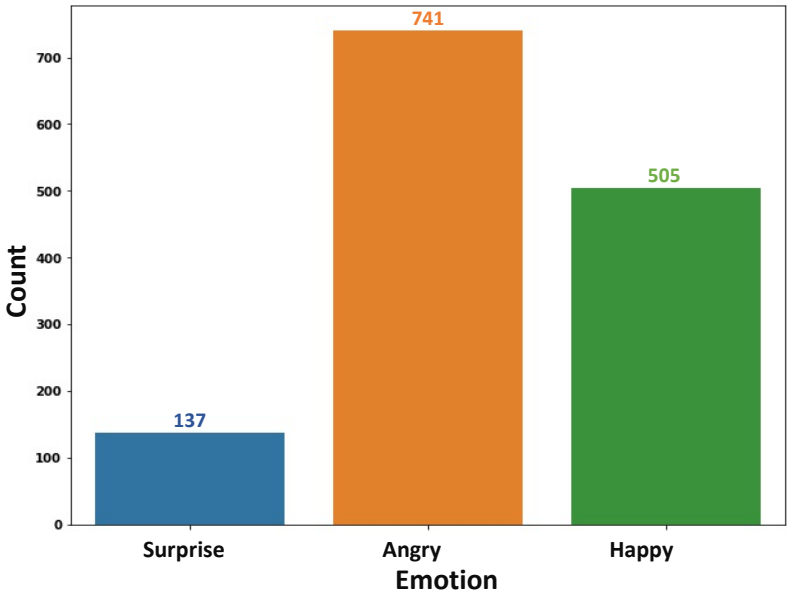


Figure 2: Emotion classes distribution for ANAD dataset

A trial to reduce the number of features by using principal components analysis (PCA) is considered. PCA [29] is an unsupervised dimensionality reduction technique that constructs relevant features through linear combinations of the original features while maintaining the majority of the important information.

PCA uses statistical tools to identify noise and redundancy in the dataset [30]. It keeps the necessary parts that have more variation of the data and removes the unnecessary parts with fewer variations, therefore speeding up the training and testing time of the machine learning algorithm. It is shown from the cumulative variance curve in Fig. 3 that all the variance is represented by only about 260 features out of 844 features, which represent 99.9 % of the total variance in the original features. So, the 260 features are constructed as linear combinations of the original 844 features that are sufficient to explain all the variance in the original features.

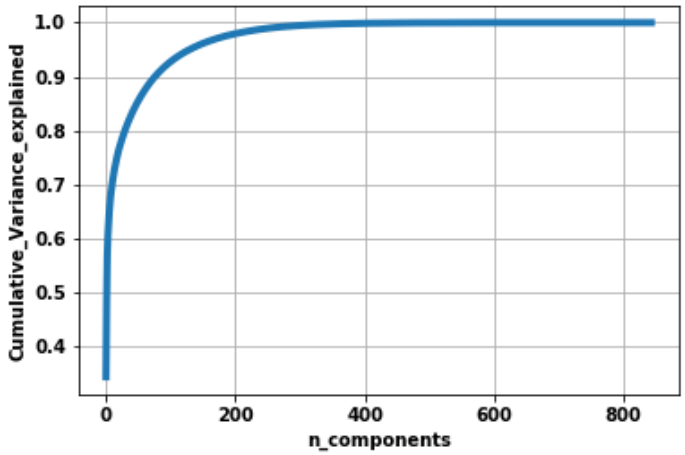


Figure 3: Cumulative variance curve

A. Corpus

The first Arabic Natural Audio Dataset (ANAD) [31] is used to evaluate our proposed model. Three discrete emotions: angry (A), surprised (S), and happy (H) are used. Eight videos of live calls between an anchor and a human outside the studio are downloaded from online Arabic talk shows. Eighteen labelers are asked to listen to the videos and label each one of them as angry, surprised, or happy. The results are averaged to label each video. Each video is then divided into callers and receivers. We removed the laughs, noise, and silent parts. Every chunk was then automatically divided into 1-sec speech units. The corpus contains 1384 records with 741 angry, 137 surprised, and 505 happy units. The properties of the used videos are depicted in Table 1. By using the B-SMOTE algorithm the three classes become of the same number of samples.

TABLE 1
CORPUS DETAILS

Id	Dialect	Gender	Length (s)	# of segments	Emotion perceived
1	Egyptian	Male	114	9	Happy
2	Egyptian	Male	78	6	Surprised
3	Gulf	Female	73.8	6	Happy
4	Jordan	Male	210	17	Angry
5	Gulf	Male	198	34	Angry
6	Egyptian	Female	23.4	2	Surprised
7	Lebanese	Female	504	24	Angry
8	Egyptian	Female	430.8	87	Happy

B. Feature Extraction

Feature extraction is the process of extracting the most important characteristics from speech signals [32]. The low-level descriptors features (LLDs) include: the probability of voicing, intensity, fundamental frequency F0, F0 envelope, LSF (0-7), MFCC (1-12), ZCR are extracted from speech. Table 2 showed the description of each feature. The following statistical functions are calculated on each feature: minimum (min), maximum (max), range (max-min), the absolute position of the max, the absolute position of the min, linear regression A (the difference of linear approximation and the contour), linear regression Q (quadratic error between the linear approximation and the contour), linear regression 1 (the slope of linear approximation of contour), linear regression 2 (offset of linear approximation of contour), standard deviation, kurtosis, skewness, Quartiles 1, 2, 3, inter-quartile ranges 1–2, 2–3, 1–3. Then, the delta coefficient for each LLD is also computed as an estimate of the first derivative.

Finally, the ineffective features are removed by testing the Kruskal-Wallis non-parametric test resulting in a features vector of length 844.

TABLE 2
SPEECH FEATURES DESCRIPTION

Name	Description
Intensity	The loudness of the sound.
Fundamental frequency (F0)	The frequency of the (quasi-) periodic structure of voiced speech signals.
F0 envelope	The amplitude envelope measured from voiced speech imposed on the signal.
Probability of voicing	The percentage of voiced energy for each harmonic.
Zero crossing rate (ZCR)	The number of times the signal crosses the zero-amplitude line by passage from a positive to negative or vice versa.
LSF	Line spectral frequencies
MFCC	Mel Frequency Cepstral Coefficients

C. Deep Feed-forward Neural Network (DFFNN) Classifier

DFFNN is a fully connected feed-forward neural network that has many hidden layers with many nodes (neurons) [33]. This means that the nodes within a layer are densely connected to the nodes in the adjacent layer. DFFNN applications include automatic language identification [34], ASER [35] [36], data compression [37], and Handwritten Characters Recognition [38]. The used DFFNN model depicts in Fig. 4 contains four hidden layers with rectified linear units (ReLu) activation function. In the end, a fully connected Softmax layer is added to output the probability of each class. We trained the model with stochastic gradient descent (SGD) as an optimizer with a learning rate of 0.001 to find the optimal weights to minimize the error and categorical cross-entropy loss function with a batch size of 32 for 300 epochs.

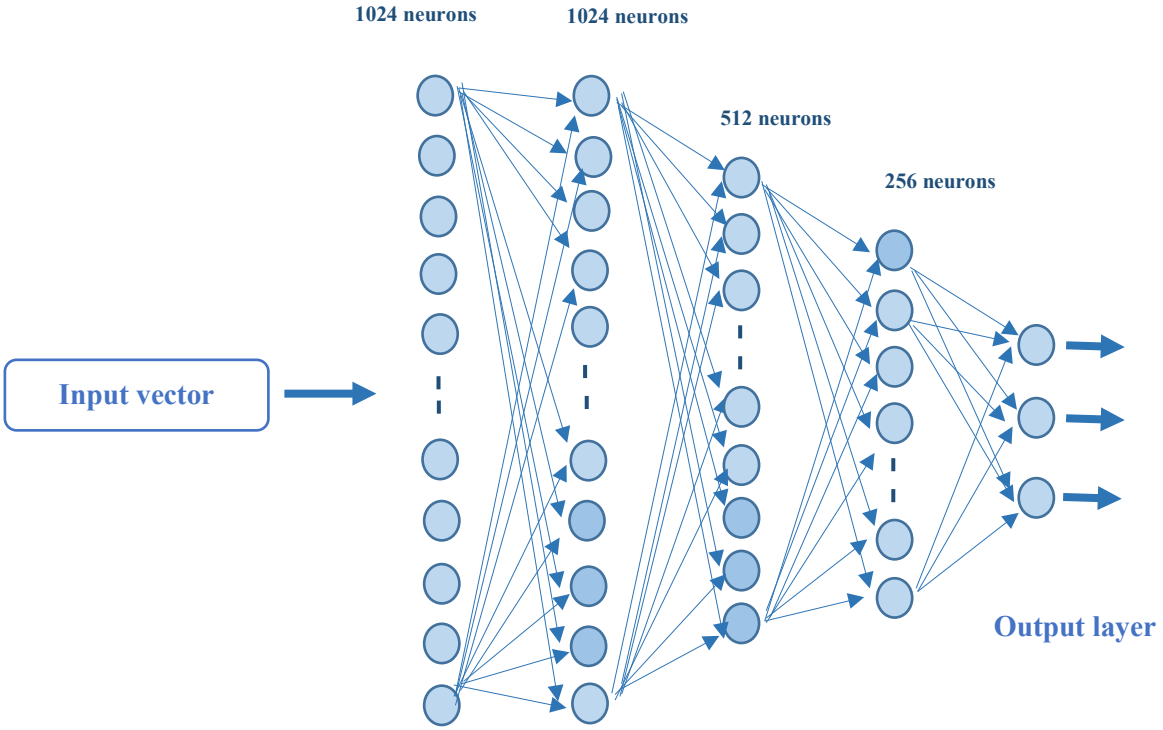


Figure 4: DFFNN architecture

4 RESULTS AND DISCUSSION

The recognition accuracy of using various combinations of the extracted features with DFFNN classifiers is reported. The vector size for all features is depicted in Table 3. The proposed model was trained and evaluated using 10-fold cross-validation, so we can make predictions on all of our data. That is randomly dividing the data into 10 subsets (folds) of equal size. Each fold is treated as a validation set once, and the remaining 9 folds are used to train the model. The process is repeated 10 times, then the average error across all 10 trials is computed. The training is achieved using an SGD optimizer

with a learning rate of 0.001 and a categorical cross-entropy loss function with a batch size of 32 for 300 epochs. Table 4 shows the confusion matrix, precision, recall, f1-score, and average accuracy for each combination of the extracted features with DFFNN.

TABLE 3
FEATURES VECTOR SIZE

Features combination	Vector size
ALL (all features)	844
MFCC	405
LSF	287
MFCC-LSF (MFCC and LSF)	692
NO-MFCC (all features without MFCC)	439
NO-LSF (all features without LSF)	557
NO-MFCC-LSF (all features without MFCC and LSF)	152
PCA	260

The highest accuracy obtained is 98.56 % when applying the PCA technique on the extracted features. It is followed by the accuracy when using all the extracted features and when using LSF-MFCC, 98.33 %, and 97.79 % respectively. The accuracy is 96.85 % when removing MFCC features from the extracted features. It is also shown that the accuracy of 95.63 % is obtained when using the LSF features. Also, it is realized that the accuracy of recognition is inclined to 92.58 % when removing LSF features. It is concluded from the obtained comparison that the LSF features are the most dominant and important features.
Figures 5 depict the comparison of the obtained results using different combinations of the extracted features with DFFNN.

TABLE 4
CONFUSION MATRIX, PRECISION, RECALL, F1-SCORE, AND AVERAGE ACCURACY OF DIFFERENT COMBINATIONS OF THE EXTRACTED FEATURES WITH DFFNN

Exp.	Features		S	A	H	Precision	Recall	F1-score	Average accuracy
#1	PCA	S	726	4	11	0.986	0.979	0.983	**98.56 %**
		A	6	730	5	0.991	0.985	0.988	
		H	4	2	735	0.979	0.991	0.985	
#2	ALL	S	721	6	14	0.989	0.973	0.981	98.33 %
		A	3	731	7	0.989	0.986	0.987	
		H	5	2	734	0.972	0.990	0.981	
#3	LSF-MFCC	S	718	9	14	0.986	0.969	0.977	97.79 %
		A	3	729	9	0.983	0.983	0.983	
		H	7	3	731	0.969	0.986	0.978	
#4	NO-MFCC	S	709	13	19	0.983	0.956	0.969	96.85 %
		A	5	718	18	0.972	0.969	0.970	
		H	7	8	726	0.952	0.979	0.965	
#5	LSF	S	709	17	15	0.972	0.956	0.964	95.63 %
		A	9	714	18	0.943	0.963	0.952	
		H	11	27	703	0.956	0.948	0.951	
#6	NO-LSF	S	652	63	26	0.939	0.879	0.907	92.58 %
		A	41	677	23	0.906	0.913	0.908	
		H	4	8	729	0.937	0.983	0.959	
#7	MFCC	S	629	76	36	0.920	0.848	0.881	90.77 %
		A	48	666	27	0.890	0.898	0.892	
		H	10	8	723	0.921	0.975	0.947	
#8	NO_LSF-MFCC	S	425	181	135	0.751	0.573	0.648	73.05 %
		A	126	534	81	0.691	0.720	0.704	
		H	18	58	665	0.678	0.804	0.735	

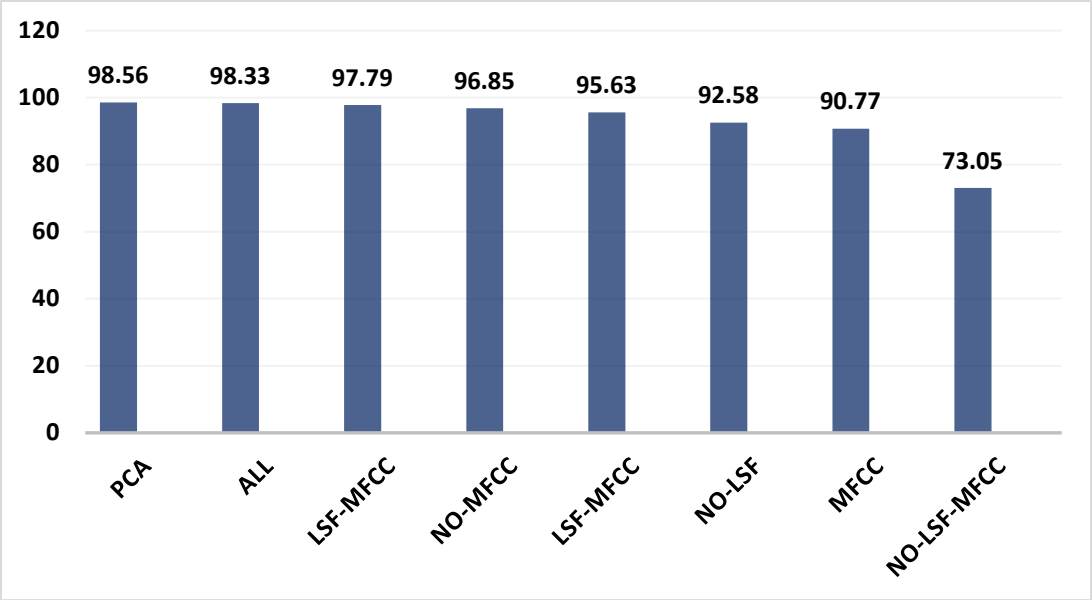


Figure 5: Recognition rate % for different features combination with DFFNN

The proposed model gives better recognition accuracy (98.56 %) than that (95.52 %) reported earlier in [24]. This means that the best recognition accuracy can be achieved when PCA with DFFNN is used.

5 CONCLUSION

Recognizing emotion from speech utterances enriches automatic speech recognition, which indeed enhances the human-machine interaction. Arabic speech emotion recognition system using a different combination of features based on DFFNN classifier is performed. It is shown that DFFNN achieves the highest accuracy when applying PCA on the extracted features with an accuracy of 98.56 %. Through the experiments in this study, it is found that the accuracy is 98.33 % when using the whole 844 extracted features. Besides, the accuracy is 97.79 % when the LSF and MFCC features are used. Moreover, the accuracy of using LSF is 95.63 %, which proves that the LSF are dominant features.

REFERENCES

[1] K. S. Rao and S. G. Koolagudi, Emotion Recognition using Speech Features, 1st edition, *Springer-Verlag New York*, 2013.

[2] E. Alsharhan and A. Ramsay, "Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition," *Lang Resources & Evaluation*, Vo. 54, pp. 975–998, Oct. 2020.

[3] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116., pp. 56–76, 01-Jan-2020.

[4] A. A. Khulage, "Extraction of pitch, duration and formant frequencies for emotion recognition system," in *Fourth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom2012)*, Bangalore, pp. 7-9, India, 2012.

[5] S. Ajibola Alim and N. Khair Alang Rashid, "Some Commonly Used Speech Feature Extraction Algorithms," in *From Natural to Artificial Intelligence - Algorithms and Applications*, *London, U.K.: IntechOpen*, pp. 4-22, 2018.

[6] B. J. Mohan and N. Ramesh Babu, "Speech recognition using MFCC and DTW," in *2014 International Conference on Advances in Electrical Engineering (ICAEE),* pp. 1-4, Vellore, India, 2014.

[7] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, "Emotion Detection Using MFCC and Cepstrum Features," *Procedia Computer Science*, vol. 70, pp. 29–35, 2015.

[8] J. C. Wang, J. F. Wang, and Y. S. Weng, "Chip design of MFCC extraction for speech recognition," *Integration*, vol. 32, no. 1–2, pp. 111–131, Nov. 2002.

[9] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," in *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing*

and Networking (WiSPNET), pp. 2257–2260, Chennai, India, 2018

[10] S. Ravindran, C. Demiroglu, and D. V. Anderson, "Speech recognition using filter-bank features," in *the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, vol. 2, pp. 1900–1903, Pacific Grove, CA, USA, 2003.

[11] R. B. Lanjewar, S. Mathurkar, and N. Patel, "Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques," in *Procedia Computer Science*, vol. 49, no. 1, pp. 50–57, 2015.

[12] C. Yu, Q. Tian, F. Cheng, and S. Zhang, "Speech emotion recognition using support vector machines," in *Shen G., Huang X. (eds) Advanced Research on Computer Science and Information Engineering. CSIE 2011. Communications in Computer and Information Science*, vol. 152, pp 215-220, 2011.

[13] V. Anoop, P. V. Rao, and S. Aruna, "An effective speech emotion recognition using artificial neural networks," in *Reddy M., Viswanath K., K.M. S. (eds) International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications. Advances in Intelligent Systems and Computing,* vol. 628. Springer, Singapore, pp. 393–401, 2018.

[14] X. Cheng and Q. Duan, "Speech emotion recognition using gaussian mixture model," in *2nd international conference on computer application and system modeling, Published by Atlantis Press,* pp. 1222-1225, Paris, France, 2012.

[15] Z. Liu and S. Wang, "Emotion recognition using hidden Markov models from facial temperature sequence," in *D'Mello S., Graesser A., Schuller B., Martin JC. (eds) Affective Computing and Intelligent Interaction. ACII 2011. Lecture Notes in Computer Science*, vol. 6975. Springer, Berlin, Heidelberg, pp. 240–247, 2011.

[16] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, "Random Deep Belief Networks for Recognizing Emotions from Speech Signals," *Comput. Intell. Neurosci.,* vol. 2017, no. 2, pp. 1-9, Mar. 2017.

[17] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention,*"* in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227-2231, LA, New Orleans, 2017.

[18] A. Christy, S. Vaithyasubramanian, A. Jesudoss, and M. D. A. Praveena, "Multimodal speech emotion recognition and classification using convolutional neural network techniques," *Int. J. Speech Technol.*, vol. 23, no. 2, pp. 381–388, Jun. 2020.

[19] A. Bin Abdul Qayyum, A. Arefeen, and C. Shahnaz, "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition," in *IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON),* pp. 122-125, Dhaka, Bangladesh, 2019.

[20] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1-4, Jeju, South Korea, 2016.

[21] M. F. Alghifari, T. S. Gunawan, and M. Kartiwi, "Speech emotion recognition using deep feedforward neural network," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 10, no. 2, pp. 554–561, May 2018.

[22] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," in *Fifteenth annual conference of the international speech communication association,* pp. 223–227, Singapore, 2014.

[23] A. Meftah, S. A. Selouani, and Y. A. Alotaibi, "Preliminary Arabic speech emotion classification," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT),* pp. 000179-000182, Noida, 2014.

[24] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, "Emotion recognition in Arabic speech," *Analog Integr. Circuits Signal Process.*, vol. 96, no. 2, pp. 337–351, Aug. 2018.

[25] L. Abdel-Hamid, "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features," *Speech Communication*, vol. 122, pp. 19–30, Sep. 2020.

[26] K. Venkataramanan and H. R. Rajamohan, "Emotion Recognition from Speech," *arXiv Prepr. arXiv1912.10458.*, pp. 1–14, Dec. 2019.

[27] T. Roy, T. Marwala, and S. Chakraverty, "Speech Emotion Recognition Using Neural Network and Wavelet Features," in *Chakraverty S., Biswas P. (eds) Recent Trends in Wave Mechanics and Vibrations. Lecture Notes in Mechanical Engineering. Springer,* pp. 427-438, Singapore, 2019.

[28] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Huang DS., Zhang XP., Huang GB. (eds) Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science*, vol. 3644. Springer, pp. 878–887, Berlin, Heidelberg, 2005.

[29] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Comput. Geosci.,* vol. 19, no. 3, pp. 303–342, Mar. 1993.

[30] E. ElMaghraby, A. Gody, and M. Farouk, "Noise-Robust Speech Recognition System based on Multimodal Audio-Visual Approach Using Different Deep Learning Classification Techniques," *The Egyptian Journal of Language Engineering,* vol. 7, no. 1, pp. 27–42, Apr. 2020.

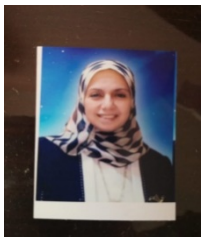
[31] S. klaylat, ziad Osman, R. Zantout, and L. Hamandi, "Arabic Natural Audio Dataset," *Mendeley Data*, vol. 1,

2018.
[32] D. Gupta, P. Bansal, and K. Choudhary, "The state of the art of feature extraction techniques in speech recognition," in *Agrawal S., Devi A., Wason R., Bansal P. (eds) Speech and Language Processing for Human-Machine Communications. Advances in Intelligent Systems and Computing,* vol. 664. Springer, pp. 195–207, Singapore, 2018.
[33] T. K. Gupta and K. Raza, "Optimizing Deep Feedforward Neural Network Architecture: A Tabu Search Based Approach," *Neural Process. Lett.,* vol. 51, no. 3, pp. 2855–2870, Jun. 2020.
[34] I. Lopez-Moreno, J. Gonzalez-Dominguez, D. Martinez, O. Plchot, J. Gonzalez-Rodriguez, and P. J. Moreno, "On the use of deep feedforward neural networks for automatic language identification," *Computer Speech & Language,* vol. 40, pp. 46–59, Nov. 2016.
[35] K. Y. Huang, C. H. Wu, Q. B. Hong, M. H. Su, and Y. H. Chen, "Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5866–5870, Brighton, United Kingdom, 2019.
[36] V. M. Praseetha and S. Vadivel, "Deep learning models for speech emotion recognition," *J. Comput. Sci.*, vol. 14, no. 11, pp. 1577–1587, 2018.
[37] F. Hussain and J. Jeong, "Efficient deep neural network for digital image compression employing rectified linear neurons," *Journal of Sensors*, 2016.
[38] J. Memon, M. Sami, and R. A. Khan, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," *IEEE Access*, vol. 8, pp. 142642–142668, Dec. 2019.

BIOGRAPHY

Engy R. Rady received a B.Sc. (Honor's) degree in Physics from the Faculty of Science, Fayoum University in 2006. She received the M.Sc. degree in speech recognition systems from Faculty of Science, Ain shams University in 2013. She is currently a Ph.D. student at the Faculty of Science, Fayoum University. She is working as Assistant Lecturer at The Basic Science Department, Faculty of Computers and Information, Fayoum University. Her research interest is in signal processing and deep learning.

Mohamed H. Farouk received the B.Sc. in Electronics Engineering from the Faculty of Engineering, Cairo University. Egypt, in 1982. He received the M.Sc. and PhD. of Engineering Physics from the Faculty of Engineering, Cairo University, Egypt, in1988 and 1994 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Cairo University, Egypt in 1984. His Current Position is full Professor, Engineering Maths. & Physics Dept., Faculty of Engineering, Cairo Univ from 2007-Till Now. He is the author and co-author of about 40 papers in national and international conference proceedings and journals.



N. M. Hassan
Dean Faculty of Computers and Information
Prof. of Theoretical Physics-High Energy.
Previewer in National Authority for Quality Assurance And Accreditation of Education

A. Hassen received Ph.D. in Natural Science from the Faculty of Science and Mathematics, University of Augsburg, Germany, in 2003. He was awarded postdoctoral scholarships from Indian National Science Academy (INSA) in 2005 and South Korea (BK21 project), Pusan National University, in 2007. He is a Professor of Materials Science, Department of Physics, Faculty of Science, Fayoum University. He was the Dean (from Aplil 2017 to April 2020) of the Faculty of Science, Fayoum University, Egypt. He published many papers in international peer-reviewed journals and conferences. He is an active reviewer for different high-ranking international journals. He becomes Associate Editor for two international Journals.

التعرف علي العاطفة عن طريق إستخدام التعرف التلقائي علي الكلام العربي

انجي رجائي راضي*[1]، محمد هشام فاروق**[2]، نبيلة محمد حسن *[3]، عرفة صبري حسن***[4]

*قسم العلوم الاساسية، كلية الحاسبات والمعلومات، جامعة الفيوم، الفيوم، مصر
[1]era00@fayoum.edu.eg
[3]nmh00@fayoum.edu.eg
**قسم هندسة الرياضيات والفيزيقا, كلية الهندسة، جامعة القاهرة، جيزة، مصر
[2]mhesham@eng.cu.edu.eg
***قسم الفيزياء، كلية العلوم، جامعة الفيوم، الفيوم، مصر
[4]ash02@fayoum.edu.eg

ملخص

يقدم هذا العمل نظام جديد للتعرف التلقائي على العاطفة عن طريق التعرف التلقائي على الكلام (ASR) باستخدام الشبكات العصبية العميقة (DFFNN) للكلام العربي. نقدم نتائج للتعرف على العواطف الثلاثة: السعادة والغضب والمفاجأة. يتم استخدام مجموعة بيانات الصوت العربي الطبيعي (ANAD). يتم استخراج خمسة وعشرين ميزة (LLDs) من الإشارات الصوتية. يتم فحص مجموعة مختلفة من الميزات المستخرجة. كما تم فحص تأثير استخدام تقنية تحليل المكونات الرئيسية (PCA) لتقليل الأبعاد. لمرحلة التصنيف ، يتم استخدام DFFNN. أيضًا ، تتم التعامل مع مشكلة عدم تساوي العينات في كل مجموعة من البيانات باستخدام تقنية (B-SMOTE). اتضح من النتائج أن أفضل دقة يتم الحصول عليها عند تطبيق PCA على الميزات المستخرجة هي 98.56٪. أيضًا ، تبلغ الدقة 98.33٪ عند استخدام جميع الميزات المستخرجة. لا تختلف هذه النتيجة كثيرًا عن دقة استخدام PCA. تليها دقة استخدام MFCC و LSF وهي 97.79٪. ويلاحظ أن الدقة 95.63٪ عند استخدام خصائص LSF مما يدل على أنها من السمات السائدة. أظهرت النتائج المتحصل عليها تحسنا مقارنة بالدراسات السابقة.

الكلمات المفتاحية

التعرف التلقائي علي العاطفة من الكلام العربي، مجموعة بيانات الصوت العربي الطبيعي (ANAD)، DFFNN، B-SMOTE، PCA