Speech Emotion Recognition Using Minimum Extracted Features

Wisal Hashim Abdulsalam

Computer Science Department

College of Education for Pure Science Ibn-Al-Haitham, University of Baghdad, and PhD candidate at The Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers & Informatics

Baghdad, Iraq

wisal.h@ihcoedu.uobaghdad.edu.iq

Rafah Shihab Alhamdani
The Informatics Institute for Postgraduate Studies
Iraqi Commission for Computers & Informatics

Baghdad, Iraq
rafah_hamdani@yahoo.com

Mohammed Najm Abdullah
Computer Engineering Department
University of Technology
Baghdad, Iraq
mustafamna@yahoo.com

***Abstract*— Recognizing speech emotions is an important subject in pattern recognition. This work is about studying the effect of extracting the minimum possible number of features on the speech emotion recognition (SER) system. In this paper, three experiments performed to reach the best way that gives good accuracy. The first one extracting only three features: zero crossing rate (ZCR), mean, and standard deviation (SD) from emotional speech samples, the second one extracting only the first 12 Mel frequency cepstral coefficient (MFCC) features, and the last experiment applying feature fusion between the mentioned features. In all experiments, the features are classified using five types of classification techniques, which are the Random Forest (RF), k-Nearest Neighbor (k-NN), Sequential Minimal Optimization (SMO), Naïve Bayes (NB), and Decision Tree (DT). The performance of the system validated over Surrey Audio-Visual Expressed Emotion (SAVEE) dataset for seven emotions. The results of the experiments showed given good accuracy compared with the previous studies using a fusion of a few numbers of features with the RF classifier.**

Keywords— Speech emotion recognition, Minimum feature extraction, ZCR, 12 MFCC, Random forest.

I. Introduction

Emotions are essential part of human life [1]. They can be expressed through body language and speech [2]. Speech signals constitute 38% of the completely communicated emotions, hence the speech emotion recognition (SER) has attracted a great deal of carefulness in human computing [3]. It has been applied in many fields such as the medical field [4]. SER can be implemented through machine-learning methods that are composed of both speech feature extraction and classification [5]. To obtain better generalization, features should be well defined [6].

This paper is about extracting the minimum possible number of features to show if it is possible to give a good accuracy to the SER and to reduce the effort and the complexity of extracting large number of features. Different classification techniques, which are Random Forest (RF), k-Nearest Neighbor (k-NN), Sequential Minimal Optimization (SMO), Naïve Bayes (NB), and Decision Tree (DT) were applied. The best result was obtained from using RF classifier with feature fusion.

This paper is ordered as follows, Section 2 describes the literature survey, Section 3 describes the proposed system including basic preprocessing operations, feature extraction, and classification, Section 4 describes the dataset used for this work, Section 5 describes the experimental evaluation, Section 6 is for the discussion, and Section 7 gives suggestion for future work.

II. Literature survey

The acoustic features principally classified as prosody, spectral and voice quality features [7]. Features like energy, pitch, and zero crossing rate (ZCR) considered prosody features, and linear predictive coding (LPC) and Mel frequency cepstral coefficient (MFCC) considered spectral features. Therefore to increase the performance of SER system, a fusion of various types of features technique is used [8], but from literature it is not possible to determine what types of features are best for recognizing emotions [9]. Also, there has been no agreement on the classifier type which is best to classify the emotion of the speech sample [10]. Therefore, in this work, different types of features were extracted, and different types of traditional classifiers were used; to reach to which one effectively classifies the emotion of the speech sample on Surrey Audio-Visual Expressed Emotion (SAVEE) dataset. A review to some of the most recent research work that was conducted from (2013) to (2018) that used the same dataset used for this work will be discussed.

Salankar, et al. (2013) [11] This paper aimed to understand and estimate the six basic emotions defined by Paul Ekman in addition to neutral state from speech using SAVEE dataset. They did not mention any preprocessing steps. They explored 15 basic acoustic features and fed them to the classifier. Features extracted are intensity, pitch, standard deviation (SD), jitter, shimmer, autocorrelation, noise to harmonic ratio, harmonic to noise ratio, energy entropy block, short term energy, ZCR, spectral roll-off, spectral centroid and spectral flux, and formants different features. They made different experiments by using four types of classifiers, which are Neural Network (NN), NB, Classification Tree (CT), and k-NN. The higher recognition rate they obtained was using k-NN classifier.

Farah Chenchah and Zied Lachiri (2014) [12] aimed to set up a SER system based on the wavelet packet energy and entropy features. They used SAVEE and the Interactive Emotional dyadic MOtion CAPture (IEMOCAP) datasets. They used framing only as preprocessing step. Feature extraction carried out using wavelet packet by partitioning the frequency axis analogous to the Mel, Bark and equivalent rectangular bandwidth (ERB) scale. The results showed that wavelet packet filter bank with ERB scale give

978-1-5386-9188-5/18/$31.00 ©2018 IEEE
DOI 10.1109/AiCIS.2018.00023

better accuracy for both of datasets. They obtained 78.75% using SAVEE dataset and 50.06% using IEMOCAP dataset.

Mahwish Pervaiz and Tamim Ahmed Khan (2016) [13] used feature fusion to increase accuracy of emotion recognition. They extracted pitch, energy and ZCR as prosodic features, MFCC and LPC as temporal features in the first stage, and linguistic features in the second stage. They proved that the classification mechanisms, if trained without considering age factor, do not help improving accuracy. They used three datasets for training and testing the model. These are SAVEE, Polish and a locally developed dataset of sky school Kindergarten Students' Dataset (KSD).

Fatemeh Noroozi, et al. (2017) [14] proposed SER method, where pitch, intensity, the first four formants, the first four formants bandwidths, mean autocorrelation, mean noise-to-harmonics ratio and SD, were used to recognize the emotional state. The proposed technique used RF to represent the speech signals, along with the DT approach, to classify them into different categories using SAVEE dataset.

In [15] Papakostas Michalis, et al. (2017) aimed to analyze speakers' emotions based on paralinguistic information. They compare two machine-learning approaches: a support vector machine (SVM), which trained on a set of 34 extracted features, and a convolutional neural network (CNN) that trained on a raw speech information. Windowing used as preprocessing step for both approaches, in addition to resizing the spectrogram for CNN. The datasets used were EMOVO, SAVEE, and EMOtional speech-DataBase (EMO-DB). The emotions represented from EMOVO and SAVEE datasets were the six basic emotions. Seven emotions used from EMO-DB these are disgust, anger, happy, fear, sad, boredom, and neutral.

Siddique Latif, et al., (2018) [16] exploited a transfer learning technique to improve the performance of SER systems. They used eGeMAPS feature set that contains 88 features. Evaluations on five different datasets in three different languages show that deep belief network (DBN) offer better accuracy than previous approaches. Results also suggest that using a large number of languages for training and using a small fraction of the target data in training can enhanced the accuracy.

The main contribution to this work is using a minimum number of extracted features that could give accuracy higher than that mentioned in the previous works that used a higher number of extracted features on the same used dataset.

III. THE PROPOSED WORK

The detailed architecture of the proposed SER system appears in Fig. 1 that consists of data training and data testing, with the following main steps: preprocessing, feature extraction, feature fusion and finally classification.

A. Basic preprocessing operation

- Pre-emphasis: The filter was applied to the signal for high-frequency amplification. It is useful in several ways such as balancing the spectrum because high frequencies are usually smaller than low frequencies, avoiding numerical problems during a Fourier transform, and may also improve the signal-to-noise ratio (SNR).
- Framing: it was used to divide the signal into a sequence of frames to analyze each frame independently which then represented by a single feature vector. The frame block in this work is of length 512 with 50% overlapping between frames and sample frequency of 16000 Hz.
- Windowing: spectral features are not required to be extracted from all speech signals, because the spectrum changes very quickly. Instead, they can be extracted from a small window where it is assumed that the statistical characteristics of a signal does not change within the region. This is done by using a non-zero window in certain areas. Hamming window was used for this work to eliminate edges and excess silence periods that do not contain information.

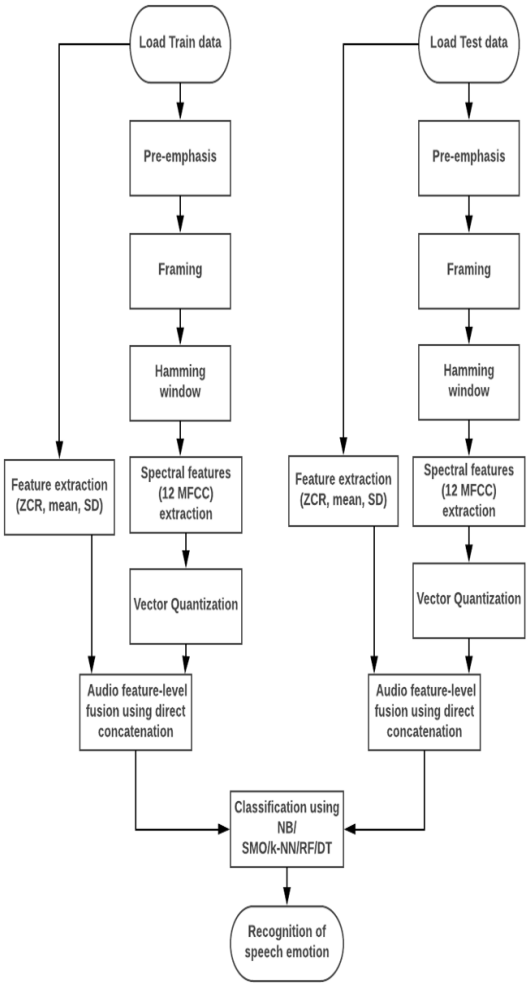

Figure 1. The detailed structure of the proposed SER.

B. Feature extraction

In the current study, MFCC features were extracted, as spectral features from each frame, and the ZCR, as prosodic feature from the original signal to find pause intervals. In addition to the SD to represent, the distribution of parameters and the mean that extracted because needed it to compute the SD. The last three features extracted from the original signal, and each one of them returns only a single value giving three features, while 12 MFCC extracted from each frame, and return two-dimensional values. Therefore, vector quantization (VQ) used to convert them to a unidimensional and then fused them with ZCR, mean and SD features in a single feature vector using direct concatenation method to enhance the effect of SER. Finally, the result will be a vector of 15 features for each frame.

C. Classification

The extracted features fed into five machine-learning models to choose the most effective one. In this work, experiments made on RF, k-NN (with k=7 according to the seven emotion classes exist in the used dataset), SMO, NB, and DT. The best result obtained using RF.

IV. Speech emotional dataset

For this work, SAVEE dataset used. It is an audio-visual emotional dataset from four English male subjects in seven emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral). The dataset consists of 120 utterances per subject, which gave 480 sentences in total. The dataset is available for research purposes under request free of charge [17].

V. Experimental evaluation

Dataset was divided into training and testing set, with 75% of the whole data samples were used for training and 25% for testing. A pre-model trained with SAVEE dataset using five types of classifiers to find a reasonable good pre-model using ten-fold cross-validation. Test speech samples classified using a classifier and the information provided by the training model. Three experiments were done: the first one extracting only three features; ZCR, mean, and SD that were entered into five types of classifiers. The second experiment was done by extracting the first 12 MFCC only, and the last experiment was done by applying feature fusion to all extracted features, which are then classified using the same five classification techniques. The accuracy obtained from applying these three experiments appears in TABLE I.

TABLE I. The accuracy obtained from applying three experiments.

Parameters	ZCR, mean, SD	MFCC	Fusion Method
NB	65.66%	70.44%	71.86%
SMO	62.38%	74.49%	77.75%
DT	68.71%	78.55%	79.18%
k-NN	66.37%	82.88%	83.58%
RF	67.74%	86.05%	87.18%

As shown in the table above, the accuracy obtained from using feature fusion was the best method, and RF was the best-used classifier.

The calculated root mean square error (RMSE) decreased as the accuracy of the proposed system increased, as shown in TABLE II.

TABLE II. RMSE of the three experiments.

Parameters	NB	SMO	DT	k-NN	RF
ZCR, mean, SD	0.317	0.327	0.328	0.316	0.306
MFCC	0.329	0.316	0.303	0.242	0.231
Feature fusion	0.321	0.313	0.303	0.239	0.222

The details of applying the first experiment of extracting only ZCR, mean, and SD and then feeding them to the classifiers are shown in TABLE III.

TABLE III. The performance criteria for seven emotions using different classifiers in SAVEE dataset using three features only.

Parameters	Average Specificity	Average F-measure	Average Recall	Average Precision
NB	11.23	40.84	42.57	41.12
SMO	14.83	28.66	39.6	28.04
DT	9.75	46.52	47.19	47.08
k-NN	10.81	41.93	43.56	41.94
RF	9.73	44.54	45.21	44.13

The details of using the first 12 MFCC features only (spectral features) and then feeding them to the classifiers are presented in TABLE IV.

TABLE IV. The performance criteria for seven emotions using different classifiers in SAVEE dataset using MFCC features only.

Parameters	Average Specificity	Average F-measure	Average Recall	Average Precision
NB	7.96	47.92	48.84	49.76
SMO	8.75	56.33	57.75	63.49
DT	6.91	63.82	64.02	63.94
k-NN	5.51	71.07	71.28	72.47
RF	4.46	76.41	76.56	76.85

Finally, the details of applying fusion to the previous three features with the first 12 MFCC and then feeding them to the classifiers are shown in TABLE V.

TABLE V. The performance criteria for seven emotions using different classifiers in SAVEE dataset using features fusion.

Parameters	Average Specificity	Average F-measure	Average Recall	Average Precision
NB	7.74	50.51	51.48	53.72
SMO	7.18	61.02	62.7	62.66
DT	6.31	64.55	64.68	65.01
k-NN	5.11	71.99	72.27	72.99
RF	4.17	78.05	78.54	78.75

TABLE VI shows a comparison of the proposed SER system with some of the recent published studies

(mentioned in the literature survey) for the years (2013-2018) that used SAVEE dataset.

TABLE VI. COMPARISON WITH THE PREVIOUS STUDIES THAT USED SAVEE DATASET.

Related work, Issuing Year	No. of emotions used	Recognition Accuracy (%)
Nilima Salankar Fulmare, et al. (2013) [11]	7-emotions	74.39
Farah Chenchah and Zied Lachiri (2014) [12]	4-emotions	78.75
Mahwish Pervaiz and Tamim Ahmed Khan (2016) [13]	5-emotions	83.4
Fatemeh Noroozi, et al. (2017) [14]	6-emotions	66.28
Papakostas, M., et al. (2017) [15]	4-emotions	30
Siddique Latif, et al. (2018) [16]	7-emotions	56.76
Current study using 12 MFCC only	7-emotions	**86.05**
Current study using feature fusion		**87.18**

VI. DISCUSSION

In the current work, ZCR, mean, SD, and MFCC features were extracted from SAVEE dataset which was used to train and test the study model. A minimum number of features were extracted and tested in different ways to show if it give a good recognition rate. The first experiment of extracting only three features gave a reasonable accuracy; but not what we aspire to. The second experiment of extracting the first 12 MFCC features gave accuracy higher than that mentioned in the recent previous studies (Table 6), but the best accuracy obtained from applying the third experiment that used feature fusion using concatenation method (Table 6). Five types of classification algorithms used, including SMO, k-NN, DT, NB, and RF, with the best accuracy achieved by using the RF algorithm (Table 1). Experiments also show that RMSE decreased as accuracy increased (Table 2). This work achieves better accuracy using less number of extracted features in comparison with recent previous studies that used the same dataset. This success is related to choosing the best-tested combination of prosodic features, ZCR, mean and SD (data comparing the choice of best combination from ZCR, mean, SD, energy and pitch are not presented in this paper) and MFCC, to feed the best-tested classifier, RF.

VII. FUTURE WORK

An algorithm for audio-based emotion recognition to seven emotions was proposed in this paper using features fusion to a minimum number of extracted features in python language for the implementation. Future work should attempt to test other types of features, other types of classifiers, combine the technique presented in this study with other modalities such as video modality, and including working with other datasets.

REFERENCES

[1] Fong, B. and J. Westerink, Affective computing in consumer electronics. IEEE transactions on affective computing, 2012. 3(2): p. 129-131.

[2] Vinciarelli, A., M. Pantic, and H. Bourlard, Social signal processing: Survey of an emerging domain. Image and vision computing, 2009. 27(12): p. 1743-1759.

[3] El Ayadi, M., M.S. Kamel, and F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 2011. 44(3): p. 572-587.

[4] Huang, Z., et al. Speech emotion recognition using CNN. in Proceedings of the 22nd ACM international conference on Multimedia. 2014. ACM.

[5] Park, J.-S., J.-H. Kim, and Y.-H. Oh, Feature vector classification based speech emotion recognition for service robots. IEEE Transactions on Consumer Electronics, 2009. 55(3).

[6] A Dictionary of Physics. 7 ed. 2015: Oxford University Press.

[7] Zhibing, X., Audiovisual Emotion Recognition Using Entropy-estimation-based Multimodal Information Fusion. 2015, Ryerson University.

[8] Luengo, I., E. Navas, and I. Hernáez, Feature analysis and evaluation for automatic emotion identification in speech. IEEE Transactions on Multimedia, 2010. 12(6): p. 490-501.

[9] Kuchibhotla, S., et al., A comparative analysis of classifiers in emotion recognition through acoustic features. International Journal of Speech Technology, 2014. 17(4): p. 401-408.

[10] Kuchibhotla, S., H.D. Vankayalapati, and K.R. Anne, An optimal two stage feature selection for speech emotion recognition using acoustic features. International Journal of Speech Technology, 2016. 19(4): p. 657-667.

[11] Fulmare, N.S., P. Chakrabarti, and D. Yadav, Understanding and estimation of emotional expression using acoustic analysis of natural speech. International Journal on Natural Language Computing (IJNLC), 2013. 2(4): p. 37-46.

[12] Farah Chenchaha, Z.L., Speech emotion recognition in acted and spontaneous context, in 6th International conference on Intelligent Human Computer Interaction, IHCI 2014. 2014, ELSEVIER. p. 139 – 145.

[13] Pervaiz, M. and T.A. Khan, Emotion Recognition from Speech using Prosodic and Linguistic Features. Emotion, 2016. 7(8): p. 7.

[14] Noroozi, F., et al., Vocal-based emotion recognition using random forests and decision tree. International Journal of Speech Technology, 2017. 20(2): p. 239-246.

[15] Papakostas, M., et al., Recognizing Emotional States Using Speech Information, in GeNeDis 2016. 2017, Springer. p. 155-164.

[16] Siddique Latif, R.R., Shahzad Younis, Junaid Qadir, Julien Epps, Transfer Learning for Improving Speech Emotion Classification Accuracy. arXiv:1801.06353v3 [cs.CV] 2018.

[17] Jackson, P. and S. Haq, Surrey audio-visual expressed emotion (savee) database. University of Surrey: Guildford, UK, 2014.