ISSN: 0976-3376

Available Online at http://www.journalajst.com

ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Asian Journal of Science and Technology
Vol.07, Issue, 03, pp.2665-2669, March, 2016

RESEARCH ARTICLE

TOWARDS ENHANCED ARABIC SPEECH EMOTION RECOGNITION: COMPARISON BETWEEN THREE METHODOLOGIES

***Abdallah Al-Faham and Nada Ghneim**

Damascus University, Damascus, Syria

ARTICLE INFO

Article History:
Received 28th December, 2015
Received in revised form
07th January, 2016
Accepted 12th February, 2016
Published online 31st March, 2016

Key words:

Emotion recognition,
High Pass Filter,
PCA,
MFCC,
SVM,
RNNs.

ABSTRACT

The research on speech emotion recognition has become one of the interesting research themes in speech processing and in human-computer interaction (HCI) applications. In this paper, we present our work with the objective to recognize the Arabic user's emotional state by analyzing the speech signal. We built an Arabic Emotional Speech corpus, covering five emotions - Happiness, Anger, Sadness, Surprise - and Neutrality. For classification, we adopted Supervised Learning approach, and implemented several classification algorithms: Support Vector Machines (SVM) with the Radial Basis Function (RBF) kernel, Neural Networks (NNs) and Deep learning approach using Recurrent Neural Network (RNNs). Since we consider emotion identification, we captured information related to the frequency and spectrum of the speech's signal. We calculated Mel-Frequency Cepstral Coefficients (MFCC) after applying High Pass Filter (HPF), and normalized the length of the features by using the (PCA) Principal Component Analysis. The comparison of the different classification methods on our Arabic Speech corpus showed that SVM approach performs better than the other two methods.

INTRODUCTION

Despite the great progress in the artificial intelligence domain, the natural Human-Machine Interaction is still far from being reality. Understanding human emotions may partly help improving the naturalness of the interaction. Recently, speech emotion recognition, which aims to recognize emotion states from speech signals, has been drawing increasing attention. The general process of the emotion recognition from speech signals has two steps: extracting acoustic features, and recognizing the emotions using appropriate classification method. The speech emotion is still a very challenging task and extracting effective emotional features is an open question. In this study, we mainly focus on the Emotion recognition from Arabic speech signals, by the comparing between theperformance of different emotion classification methodologies on an Arabic Speech corpus specially built for this purpose. In the next section, we present the literature survey carried out, and then we describe our speech data corpus in Section 3. In Section 4, we explain our proposed features and classification methods in details. We show the experimental results in Section 5 and conclude the paper in Section 6

****Corresponding author: Abdallah Al-Faham***
Damascus University, Damascus, Syria

Related Works

Actually there are several techniques used for emotion detection through the speech. Emotion modeling is the main task of emotion recognition through speech. The various classifiers are used for the classification of the features of the speech. Various classifiers had been used for the classification of the speech features, such as Neural Networks (NNs), Gaussian Matrix Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM) etc. (Rawat and Mishra, 2015) proposed speech features such as, Mel Frequency cepstrum coefficients (MFCC) and evaluated these features (after applying a HPF) by using neural networks, (Meddeb *et al.,* 2014) try to achieve Intelligent Remote Control for TV program based on Emotion in Arabic speech, proposed speech features such as, Linear Prediction cepstrumcoefficients (LPCC) , Mel Frequency cepstrum coefficients (MFCC), (Devi and Nandyala, 2014; Kwon *et al.,* 2003) Energy and Pitch and used SVM for classification. (Rawat and Mishra, 2015) introduced text dependent speaker recognition with an enhancement of detecting the emotion of the speaker prior using the hybrid FFBN and GMM methods. (Meddeb *et al.,* 2014) proposed a work where the emotion had been detected by formant frequency, (MFCC), and they compared the performance of SVM, LDA, QDA and HMM classifiers based on discriminative and generative models,

concluding that Both SVM and HMM gave much accuracy thanothers. (Zhou and Huang, 2014) proposed humans emotional speeches recognition contributes much to createmore human machine interaction, also with many potential applications. I. Bisio, A. Delfino, F. Lavagetto, M. Marchese and A. Scirrone (Kwon *et al.,* 2003) proposed system which was able to recognize the emotional state of a person. Firstly, they started the registration of audio signals. The system composed of two functional blocks: Gender Recognition (GR) and Emotion Recognition (ER). The Pitch Frequency Estimation method used for the recognition of gender whereas support vector machine (SVM) used for the detection of emotion. (Meftah *et al.,* 2015) calculated rhythm metrics were used to recognize five Arabic speaker emotions (neutral, sad, happy, surprised, and angry) using an MLP-type neural network classifier, and the results of the automatic recognition system were similar to the human perceptual test.

would be wrong even if we apply suitable methods. Some of the sentences that we regarded are:

- Inexhaustible treasure
  "AlqnAEpknzlAyfny"
- I traveled together with my family to a new House
  "SAfrtbSHbp EA}ltyIlYbytjdyd"
- Encountered in the old way friends
  "SAdftfyAlTryqOSdqA}yAlqdAmY"
- I saw him tired with an exuded sweaty forehead
  "rOythmtEbAFwjbynhytfS~dErqAF"

We analyzed the spectrum of a sentence such as "Inexhaustible treasure" Fig. 1. Shows the different spectrums of the emotional sentences. We notice that the spectrums of female are larger than the spectrums of male and, there is marked resemblance amongthese spectrums of to the same performer, therefore Emotion Recognition is challenging task.

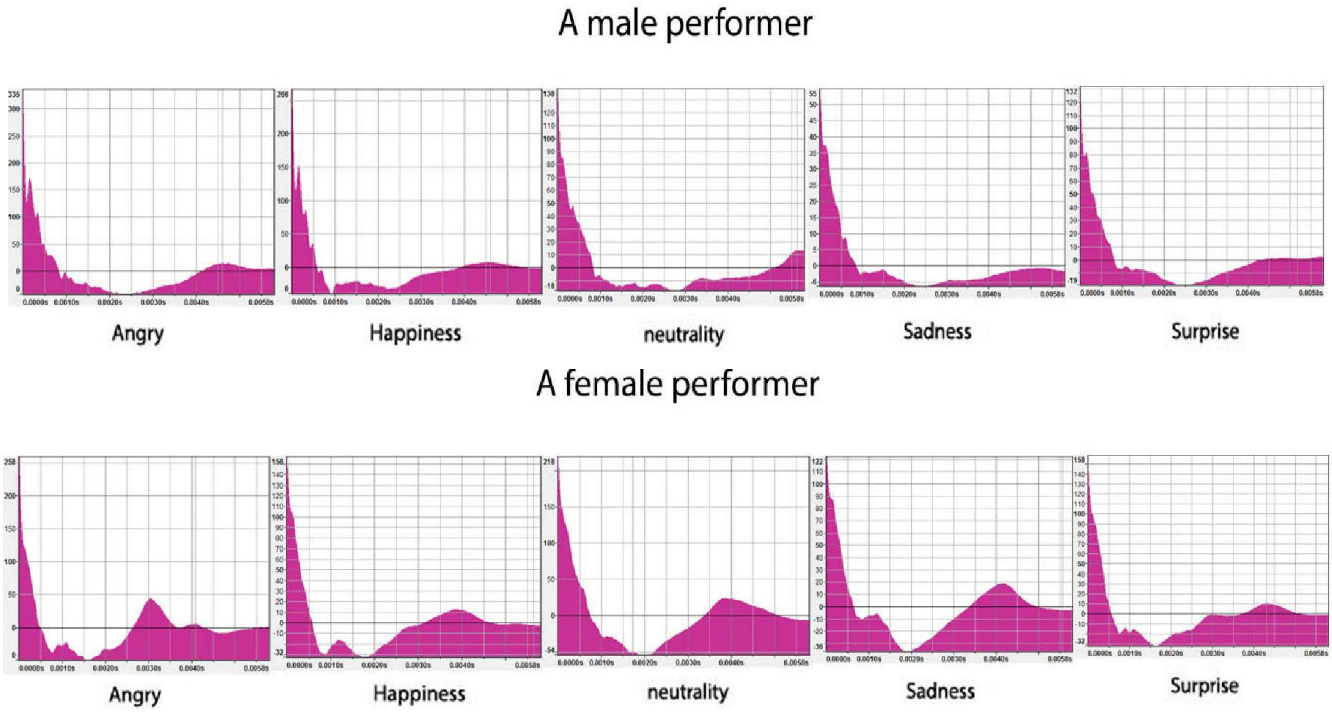

Fig. 1. The spectrums of the emotional sentence

Table 1. Confusion Matrix of experimental analysis of our speech data corpus with 96.87% mean recognition rate

Labeled emotion	Recognized Emotion (%)				
	Happiness	Sadness	Neutrality	Anger	Surprising
Happiness	95.83	0	0.69	2.08	1.38
Sadness	0	98.61	1.38	0	0
Neutrality	0.34	2.77	96.18	0.69	0
Anger	1.38	0	0	97.22	1.38
Surprising	2.08	0	0	1.38	96. 52

Speech Data Corpus

The speech data used in the experiment contains 24 Arabic Emotional sentences recorded by 6 performers (3 male and 3 female)with every sentence recorded twice and considering 5emotions (Happiness, Anger, Sadness, Surprise). Thus, the total number of sentences is 24*2*5*6=1440 records. Since the evaluation of the corpus based on the human perceptual so we must take into account the experience of the performers, therefore if the performers are unprofessional so the result

Table 1 shows the evaluation of the corpus according to 6 evaluators.

METHODOLOGY

The general process of emotion recognition from speech signals has two steps: extracting acoustic features, and recognizing emotions using appropriate classification method

***Acoustic Features*:** In order to extract the appropriate acoustic features we applied three steps. Each step has an important effect upon the whole process. These steps are:

High Pass Filter: When speech has noise or rumble, then even humans may find difficulty to recognize the emotion, speech must be clarified, and this could be achieved that by using

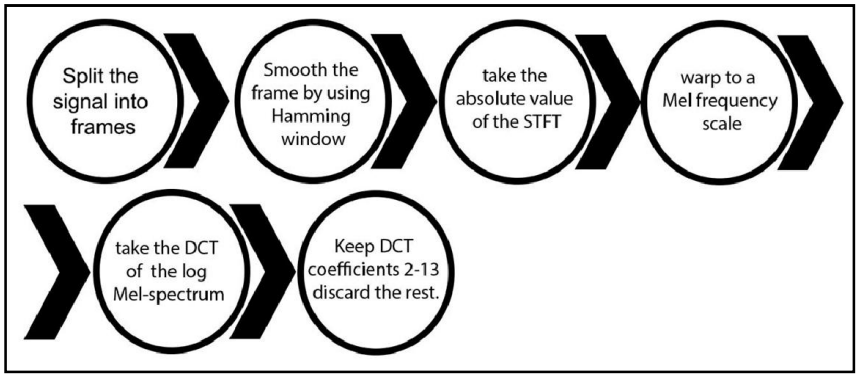

Fig. 2. Feature Extraction using MFCC

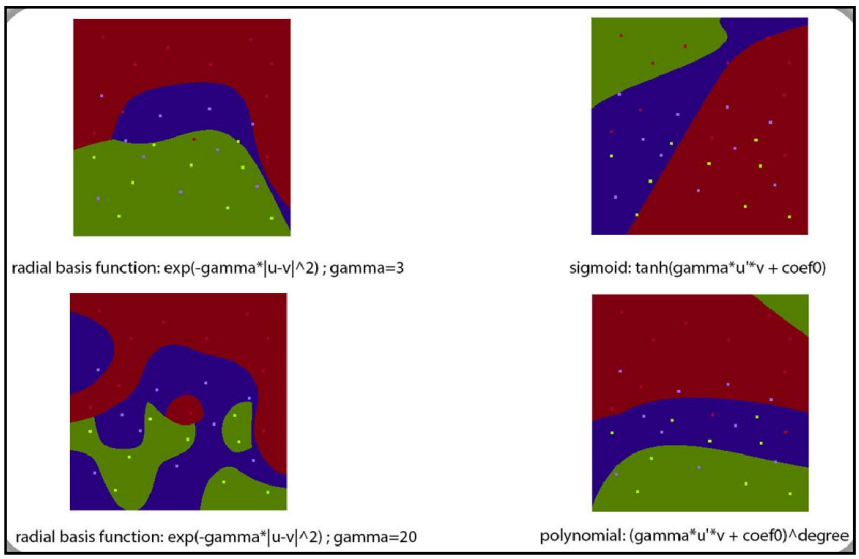

Fig. 3. Classification in various kernels

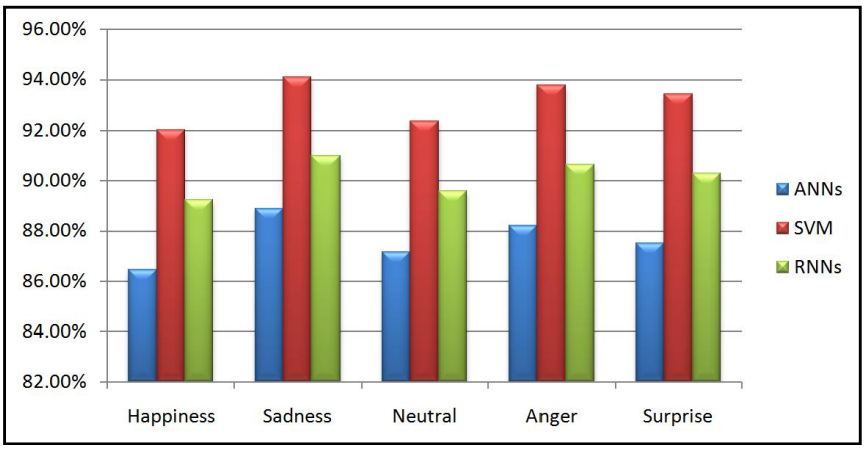

Fig. 4. Comparison of different approaches' results

Table 2. Confusion matrix of the ANNs classifier with 87.63% mean recognition rate

Labeled emotion	Recognized Emotion (%)				
	Happiness	Sadness	Neutrality	Anger	Surprising
Happiness	86.45	0.34	2.77	6.94	3.47
Sadness	0.69	88.88	7.29	1.73	1.38
Neutrality	1.73	7.29	87.15	2.77	1.04
Anger	3.81	0.69	1.04	88.19	6.25
Surprising	5.55	1.73	1.38	3.81	87.50

steps we follow to calculate MFCC is shown in Fig. 2. From each frame, we get 13 DCT coefficients, but the number of frames is various depending on the length of the speech signal. Therefore, we applied Principal component analysis **(**PCA) method to normalize the number of the data and reduce feature dimensionality.

Artificial Neural Networks (ANNs): Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. We use NNs with 50 neurons in hidden layer with the Gradient Back propagation as the training function.

Table 3. Confusion matrix of the SVM classifier with 93.12% mean recognition rate

Labeled emotion	Recognized Emotion (%)				
	Happiness	Sadness	Neutrality	Anger	Surprising
Happiness	92.01	0.34	1.38	3.47	2.77
Sadness	0	94.09	4.86	1.04	0
Neutrality	0.34	4.51	92.36	1.73	1.04
Anger	2.43	0	0.69	93.75	3.12
Surprising	3.47	0	1.04	2.08	93.40

Table 4. Confusion matrix of the RNNs classifier with 90.13% mean recognition rate

Labeled emotion	Recognized Emotion (%)				
	Happiness	Sadness	Neutrality	Anger	Surprising
Happiness	89.23	0.69	1.73	3.81	4.51
Sadness	0	90.97	5.90	2.43	0.69
Neutrality	0.69	6.59	89.58	2.08	1.04
Anger	3.12	0	1.04	90.62	5.20
Surprising	4.16	0.34	1.73	3.47	90.27

Principal component analysis (PCA): Every sample is represented by a vector of data which is extracted by using MFCC. To reduce the length of vectors of the samples we applied PCA method, where the data is considered as a group of points in a space with N dimensions (N expresses the length of data). Each point has N axises but these axises don't usually perpendicular in pairs because there is a correlation among these features. So, the method tries to get features whose axises are independent and perpendicular in pairs, and the vector of those features will be representative of asample. PCA reduces the number of correlate features by combining them, and the resultant features represent the eigenvalues of the sample, so the accuracy of the vector depends on the covariance between its features.

$$\frac{\frac{1}{m}\sum_{i=1}^{m} \|x(i) - x(i)approx\|^2}{\frac{1}{m}\sum_{i=1}^{m} \|x(i)\|^2} \leq 0.01$$

The previous equation gives features with 99% of covariance, where the numerator represents the total error of the projection (which reflects the correlation between the mergedaxises) and the denominator expresses the diversity of the data. We use the shortened formula of the equation. Where Sii is the matrix of eigenvalues.

$$1 - \frac{\sum_1^k sii}{\sum_1^m sii} \leq 0.01$$

Classification

Three classification methods were used in this experiment to study each method's accuracy, and compare the results on the same corpus. We used Neural Networks, Support Vector Machines, and Recurrent Networks.

Support Vector Machine (SVM): Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. SVMs can efficiently perform a non-linear classification using what is called the kernels, implicitly mapping their inputs into high-dimensional feature spaces. We train with the Radial Basis Function (RBF) kernel. The advantage of using RBF kernel is that it restricts training data to lie in specified boundaries. The RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. The RBF kernel has less numerical difficulties than polynomial kernel. The Fig. 3 shows the classification in various kernels.

Recurrent Neural Network (RNNs): Unlike feed forward Neural Networks, RNNs can use their internal memory to process arbitrary sequences of inputs. In this experiment, we used RNN with 10 hidden, 2 delays with Levenberg Marquardt back propagation as the training function

RESULTS

Table 2 shows that best percentage of using ANNs approach reach up to 87.63% mean recognition, with most confusion between Sadness and Neutrality. Table 3 shows that using SVM approach has better results which reach up to 93.12% mean recognition rate. Moreover we notice that sadness emotion doesn't contain any wrong prediction as anger or surprising feeling, also in happiness emotion there is no sadness predicate. Fig 4 presents a comparison between different emotions recognition rate in each approach. The *Sadness* emotion has the highest recognition rate, and *Happiness* emotion has the lowest recognition rate.

Conclusion

In this study we have proposed three techniques for Arabic Speech Emotion classification (ANN, SVM, RNN). The used features were extracted using MFCC, after filtering the speech with HPF and finally normalizing the length of features' vector using PCA method. The results show that in comparison with other classifiers, the SVM approach has a better recognition rates than the ANN and RNN. In our future work, our aim is to evaluate different variations of PCA method (such as multi linear PCA, and Kernel Principal Component Analysis (KPCA)) on the recognition rates. Moreover, more classification approaches can be applied in order to evaluate their accuracy.

REFERENCES

Bisio, I., A. Delfino, F. Lavagetto, M. Marchese and A. Scirrone, "Gender Driven Speech Recognition Through Speech Signals for Ambient Intelligent Applications," *IEEE Transactions on Emerging topics in Computing*, vol. 1, 2013.

Devi, J. S., and S. P. Nandyala, "Automatic Speech Emotion and Speaker Recognition based on Hybrid GMM and FFBNN," *International Journal on Computational Sciences & Applications (IJCSA),* vol. 4, 2014.

Kwon, O.W., K. Chan, J. Hao and T.-W. Lee, "Emotion Recognition by Speech Signals," Eurospeech - Geneva, 2003.

Meddeb, M., H. Karray and Adel, M. Alimi, "Intelligent Remote Control for TV Program based on Emotion in Arabic Speech," *International Journal of Scientific Research & Engineering Technology (IJSET)*, vol. 1, pp. 2277-1581, 2014.

Meftah, A., Y. A. Alotaibi and S.-A. Selouani, "Arabic Speaker Emotion Classification using Rhythm Metric and Neural Networks," Conference: 23rd European Signal Processing Conference (EUSIPCO), At Nice, France, pp. 1426 - 1430, 2015.

Rawat, A. and P. K. Mishra, "Emotion Recognition through Speech Using Neural Network," *International Journal of Advanced Research in Computer Science and Software Engineering*, pp. 422-428, 2015.

Yu, L., K. Zhou and Y. Huang, "A Comparative Study on Support Vector Machines Classifiers for Emotional Speech Recognition," Immune Computation (IC), vol. 2, 2014.

****