Arabic Sound Emotion Recognition

Moustafa Abada
CiE Program
Zewail City
6th of October, Egypt
s-moustafa.abada@zewailcity.edu.eg

Abdelmoez Elsaadany
CiE Program
Zewail City
6th of October, Egypt
s-abdelmoez.elsaadany@zewailcity.edu.eg

Mohamed Ellebody
CiE Program
Zewail City
6th of October, Egypt
s-mohammad.ellebody@zewailcity.edu.eg

Abstract*— Emotions are the underlying hidden feelings and mentals states that people feel daily. Emotions are hard to detect, however, previous research has shown that we can use facial expressions to reveal people's emotions [1]. New research has shown that voices can also be used to recognize different emotions such as being happy, sad, surprised, angry, nature, fear and disgust [2.] Specific features are extracted from speech such as MFCC, chroma, Mel, contrast, tonnez, spectrum centroid, and zero-crossing rate. These features are used in the machine learning classification to classify emotions. Our contribution in this paper is making a comparison between different machine learning classification algorithms such as decision tree, MLP, SVM, and convolutional neural networks. Also, we are testing different combinations of the features extracted. We used both English and Arabic datasets. The English datasets used are the Toronto emotional speech set (TESS), the Berlin emotional speech set (EmoDB), and The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). For the Arabic speech emotions dataset, we used the PCA algorithm which is considered as a feature selection technique to avoid redundancy in features and select the most important features that well represent the dataset.

Keywords—Emotions, Features, Recognition, MFCC, Classification, Accuracy.

Introduction

Emotions study is a very important topic in human interactions. Hidden feelings and mentals states of people are observed using emotions study. This is a new field called “Speech Emotion Recognition System (SER)”. Many areas can benefit from this research such as E-learning, call centres, commercial applications, and computer games. For example, in E-learning or online classes, information about the underlying emotional state of students can be revealed and some improvements of the educational systems or equipment may occur. Also, we may know what subjects the students find interesting and others that the students find boring so that we develop more fun strategies to help students. SER systems also can be used in mobile services, call centres and psychological assessment. The experience of the client or the customer can be evaluated using SER Systems and further improvements can be made.

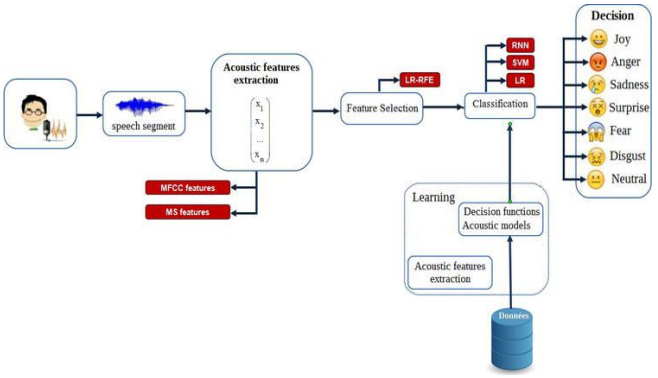

Fig. (1) SER System Flowchart

The SER system is shown above. It consists of three main modules. First, the dataset is processed and features are extracted from voices. Then, these features are fed into the training classifiers and emotions are detected. Previous research failed to make a comparison between different machine learning classification algorithms and failed to compare different combinations of the features extracted. In 2014, Gong et al only used SVM and RNN classifiers and used a Chinese dataset [3]. They failed to test additional machine learning classifiers. While Selvaraj et al, in 2016, extracted the following features: energy, pitch, linear prediction cepstral coefficient (LPCC), and Mel frequency cepstrum coefficient (MFCC) but only used Radial Basis Function and Back Propagation Network[4]. Moreover, the most important thing is that the research in Arabic speech emotion recognition is limited.

We solved these problems by using three English datasets and we made a comparison between the different machine learning classification algorithms such as MLP, SVM, Decision Tree and Neural Network. We also used different combinations of features such as “MFCC” only, “MFCC & Mel & Chroma & Contrast & Tonnetz” and “MFCC & Mel & Chroma & Contrast & Tonnetz & Spectrum centroid & Zero crossing rate”. Moreover, we decided to work on the Arabic dataset. We have used the Arabic Natural Audio Dataset (ANAD) that consists of eight long calls that are divided into small files of 1-second length and features are extracted from them. Then, we used a feature reduction and selection algorithm called “Principal Component Analysis (PCA)” to get the most important components. This led to a

reduction in the number of features from 198 features to 19 features. Then we used these features for classification using MLP and Convolutional Neural Network.

METHODOLOGY

A. Features

1. Mel Frequency Cepstral Coefficients (MFCCs)
   MFCC is a small set of features that can represent the shape of the spectral envelope of the signal.
2. Mel-energy spectrum dynamic coefficients (MEDC)
   MEDC is similar to MFCC, however, the main difference is, in MEDC, the logarithmic mean of energies is taken while in MFCC logarithmic is taken. Also, the 1st and 2nd difference are computed.
3. Spectral centroid
   The spectral centroid is a measure of how low or high the voice signal is. It is calculated by finding the centroid of the signal using the following formula:-

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

Eq. (1) Centroid Calculation

- The result of the formula is the weighted mean of the frequencies of the signal. These frequencies are obtained using a Fourier transform. In the above equation f(n) is the centre frequency at bin number n and x(n) is the value or magnitude of the signal at bin number n
- The signal regions with a higher magnitude of the spectral centroid mean that the signal is brighter there.

4. Chroma.
   It is a condensed representation of the content of the musical signal. It has an important role in preparation for high-level semantic analysis. It can be used in the analysis of voices whose pitches can be categorized into twelve categories.
5. Contrast
   The difference between different speech sounds makes each sound different from other speech sounds.
6. Zero-crossing rate
   It is a measure of the frequency content of the signal.

B. Datasets

1. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). It contains 7356 files. It has 24 professional actors (12 female, 12 male). Emotions are calm, happy, sad, angry, fearful, surprise, and disgust expressions. The song contains calm, happy, sad, angry, and fearful emotions [5].

2. Toronto emotional speech set (TESS).

A set of 200 target words were spoken by two actresses (aged 26 and 64 years) and recordings were made of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral) [6].

3. Berlin emotional speech set (EmoDB) dataset.

The EMODB database comprises seven emotions: 1) anger; 2) boredom; 3) anxiety; 4) happiness; 5) sadness; 6) disgust; and 7) neutral. The dataset consists of 535 utterances from actors who say one word per each utterance [7].

4. Arabic Natural Audio Dataset
This dataset consists of 1384 records with 7 different emotions. Each record is 1 second long and is evaluated using 18 listeners who decide which emotion appears in this record. Laughs, silence and noisy parts were eliminated from the records. The whole dataset was obtained from 8 videos between an anchor and a human outside of the studio. The features of the utterance were extracted then (Principal component analysis)PCA was applied to reduce the number of features to 10%. This means every 190 features were reduced to nineteen features. then the emotions of each utterance were detected [8]

C. Algorithms

We used more than one algorithm to compare the performance of each one on different datasets.

1. Multilayer perceptron(MLP).
   A multilayer perceptron with 300 neurons and a batch size of 256. MLP was used as it can solve problems stochastically with a good approximate solution like the one of detecting the emotion from voice.

2. Support Vector Machine(SVM).
   In this project, SVM was used to classify the emotions obtained from voice as its main use is in classification problems.

3. Decision tree.
   It was used as it can classify the input by dividing it into smaller subsets of the dataset.

D. Activation function, loss function, and optimizer.

- Relu activation function was used to eliminate the negative values as this helps in making the model work fast and better.
- Softmax is used in the output layer to turn the numeric values into probabilities with the summation of one.
- Adam optimizer was used as it combines the best properties of RMSprop and stochastic gradient descent.
- Sparse categorical cross-entropy is a loss function that was used as we have a classification problem with more than two classes. Also, each output has a unique class.

E. Train accuracy and test accuracy were calculated using:-

$$Accuracy = \frac{Correctly\ classified\ test\ samples}{Total\ number\ of\ test\ samples} \times 100$$

Eq. (2) Accuracy Calculation

Evaluation of accuracy was done on the training dataset, validation dataset, and test dataset.

RESULTS

RAVDESS Dataset

Only six emotions are tested which are: Neutral Calm Happy Sad Angry Fearful The first part compares the different classification algorithms and only MFCC features are extracted. The results are shown below:

Algorithm	Features Used	Validation Accuracy
Decision tree classifier	MFCC	64.5%
Multilayer perceptron classifier	MFCC	86%
Support vector machine classifier	MFCC	82.21%
Convolutional Neural Network	MFCC	68%

Table. (1)

The second part is the result of the comparison between different combinations of features and only an MLP classification algorithm is used:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC	86%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz	85%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz & Spectrum centroid & Zero crossing rate	73.52%

Table. (2)

TESS Dataset

This dataset contains 7 emotions and was generated by 2 females both of them generated 200 audio files for each emotion making a total number of samples in this dataset 2800 sample, the emotions found in this dataset are: Fear Pleasant surprise Sad Angry Disgust Happy Neutral.

The same procedure is done for the TESS dataset; different classification algorithms are tested and only MFCC features are extracted. The results are shown below:

Algorithm	Features Used	Validation Accuracy
Decision tree classifier	MFCC	91.43%
Multilayer perceptron classifier	MFCC	98.81%
Support vector machine classifier	MFCC	99%
Convolutional Neural Network	MFCC	98.81%

Table. (3)

Furthermore, different features combinations are tested and only MLP classification algorithm is used as shown below:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC	98.81%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz	99.76%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz & Spectrum centroid & Zero crossing rate	99.52%

Table. (4)

EmoDB Dataset

This dataset contains 7 emotions and was generated by professional speakers (five males and five females) who participated in data recording. The database contains a total of 535 utterances.

As shown below different classification algorithms are tested and only the MFCC feature is extracted:

Algorithm	Features Used	Validation Accuracy
Decision tree classifier	MFCC	36.7%
Multilayer perceptron classifier	MFCC	50%
Support vector machine classifier	MFCC	23%
Convolutional Neural Network	MFCC	NA

Table. (5)

Also, different features combinations and only MLP classification algorithm are used:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC	50%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz	99.76%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz & Spectrum centroid & Zero crossing rate	66.7%

Table. (6)

Confusion matrix

The confusion matrix is done for all the classification algorithms. For example, in the MLP algorithm and MFCC feature only for the RAVDESS dataset, we got an accuracy of 86% and the confusion matrix is shown below:

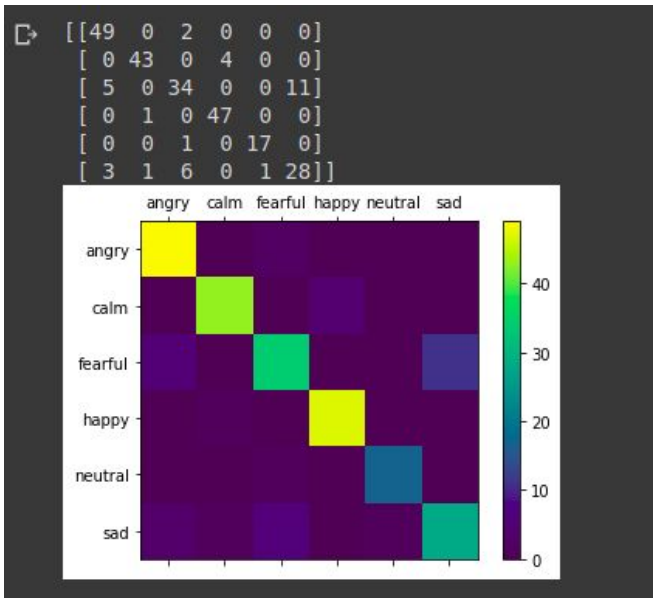

Fig. (2) Diagonal Matrix of MLP with MFCC Feature

The diagonal of the matrix is the correct predicted results. The accuracy is 86%.

The Arabic dataset accuracy without PCA:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Spectrum centroid & Zero crossing rate	79.5%
Convolutional Neural Network	MFCC & Mel & Chroma & Contrast & Spectrum centroid & Zero crossing rate	92.6%

Table. (7)

The Arabic dataset accuracy with PCA:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Spectrum centroid & Zero crossing rate	95.8%
Convolutional Neural Network	MFCC & Mel & Chroma & Contrast & Spectrum centroid & Zero crossing rate	96.5%

Table. (8)

Discussion

As shown in the previous work it's proven that emotions can be detected and recognized from speech using machine learning algorithms and voice features such as MFCC Mel Contrast Tonnetz spectrum centroid zero-crossing rate. it's proven also that both Arabic and English datasets are treated equally; both datasets are processed and features are extracted from them. Then classification is done based on these features. Moreover it's proven that the MLP classifier showed a better performance if the dataset is small while the neural network overfits. If the dataset is large, the Neural Network will be better.
Also, we showed that using a combination of all features and then doing PCA for feature reduction gives greater accuracy and performance than without PCA.
We have used real Arabicaudio files record in mobile calls and we have got a good performance.

We only used a small sample of the Englishdatasets because that's what we found online. That's why the Berlin dataset has low accuracy. However, we plan to train the whole datasets.

Acknowledgment

We would like to thank Dr MostafaElshafi and Eng. Ahmed Weal for their efforts and support.

References

[1] "REVIEW OF FACIAL EMOTION RECOGNITION SYSTEM," *International Journal of Pharmaceutical Research*, vol. 10, no. 03, 2018.

[2] 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, 2020

[3] Huang, W. Gong, W. Fu, and D. Feng, "A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM," *Mathematical Problems in Engineering*, vol. 2014, pp. 1–7, 2014.

[4] M. Selvaraj, R. Bhuvana, and S. Padmaja, "Human speech emotion recognition," *International Journal of Engineering & Technology*, vol. 8, pp. 311–323, 2016

[5] Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one. 2018 May 16;13(5):e0196391.

[6] Dupuis, Kate, and M. Kathleen Pichora-Fuller. "Toronto emotional speech set (TESS)" (2010).

[7] Burkhardt, Felix & Paeschke, Astrid & Rolfes, M. & Sendlmeier, Walter & Weiss, Benjamin. A database of German emotional speech. 9th European Conference on Speech Communication and Technology. (2005).5. 1517-1520.

[8] klaylat, Samira; Osman, Ziad; Zantout, Rached; Hamandi, Lama, "Arabic Natural Audio Dataset", Mendeley Data, V1, (2018), DOI: 10.17632/xm232yxf7t.1