ECAI 2015 - International Conference – 7th Edition
Electronics, Computers and Artificial Intelligence
25 June -27 June, 2015, Bucharest, ROMÂNIA

Comparative Study Regarding Characteristic Features of the Human Voice

Roxana Mădălina LEXUȚAN
Research Center
Military Technical Academy
Bucharest, Romania

roxana.lexutan@mta.ro

Abstract – **Each human voice has its own characteristic features. The features that describe the human voice can be: the pitch, the energy of the signal, the mean value, the median value, the MFCC coefficients, and others. Some of the natural questions that come to mind are: which are the features that best characterize the human voice? What is the minimum number of features that can completely characterize the human voice? In order to answer these questions a database of human voices is used, along with different feature extraction algorithms. Just extracting the features is not enough. Some classification algorithms need to be compared in order to find the best one for the data that is available.**

Keywords: features, classification, human voice, pitch;

I. Introduction

The notion of affect burst was first mentioned by Scherer [1]. It refers to refers to sudden, full-blown displays of emotion in facial, vocal, and gestural components that are highly synchronized. In the field of emotion recognition based on the facial expression the first one to study them was Paul Ekman [2, 3]. As far as emotion recognition based on vocal expressions goes there was intensive research carried out in recent years. Some researches can be found in [4,5,6]. For both expression researchers have come to the conclusion that they are universal.

When considering affect bursts, they have been studied from a psychological point of view since Scherer first introduced the concept. Even if psychologically we can explain them and understand how they work, affect bursts remain some kind of a “mystery” for computer scientists. They have tried to develop algorithms for their recognition by computers. One of the main problems that computer scientists face is the lack of recordings needed for their characterization and full understanding. There are few databases in the literature that provide recordings for further study.

In this paper the problem of nonverbal expressions in the human voice is addressed. Explicitly the problem of which and how many features best characterize the human voice when one of the following three emotions are present: happiness, sadness and the neutral state. The number of features used for a certain application has a great importance. It can determine the complexity of the code, the execution time can increase with the increase of features used. Therefore determining the minimum number of features to be used for this kind of application is very important. Section II presents some feature extraction and classification methods. In section III the proposed method is presented and the results are synthesized in section IV.

II. Feature Extraction and Classification

A. Feature Extraction

A feature represents a distinctive attribute of an object. Some of the features of the human voice are: the fundamental frequency, the energy of the signal, the duration of speech, length of unvoiced segments, the ratio of voiced and unvoiced segments, the MFCC coefficients.

In order to use these features for a specific application, these need to be extracted from the signal. For each feature there are multiple algorithms found in the literature that can be used for its extraction.

For example, for pitch extraction there are three main categories of methods:

- time domain based methods
- frequency domain based methods
- time-frequency domain based methods

Each of these three categories comprises multiple algorithms. For example the frequency domain based methods can use the Fourier Transform or the Wavelet Transform, for each of the two transforms different methods for extracting the pitch can be implemented. In this paper a time-frequency domain method has been used. The method is the normalized cross-correlation.

Regarding the energy of the signal, this is computed using the well known formula

978-1-4673-6647-2/15/$31.00 ©2015 IEEE

$$E_x = \sum_{n=-\infty}^{\infty} |x[n]|^2 \quad (1)$$

As for the micro prosody features extracted in this paper, the algorithm is presented in section III.

B. Classification

The process by which an object is assigned to the class it belongs to is known as classification. Each class has its own characteristics. Based on the features of an object and on the features of each class, a classifier makes the decision if that object belongs to a class or another.

There are many known classifiers. From these we can mention neural networks, support vector machines, hidden Markov models, decision trees, naïve Bayes and k nearest neighbor. Next, a few details about neural networks, support vector machines, decision trees and k nearest neighbor classifier are presented.

A neural network consists of an interconnected group of neurons, which process the information in order to make a decision. It was inspired by the human brain network. It can have different topologies and different number of layers. They can be feed forward networks, feed-back and recurrent networks. There is no strict number of layers. The simplest network has only one layer: the input layer and the output layer.

Support vector machines (SVM) were first introduced by Boser, Guyon and Vapnik in 1992 [8] and represent a group of supervised learning methods, which can be applied for classification and regression. The points that lie closest to the decision surface are the most difficult to classify. These are called support vectors. Support vector machines maximize the margin between the support vectors. There are different types of SVMs: linear, non-linear, polynomial, rbf (radial basis function).

Decision trees are a supervised learning method used for classification and regression. The purpose of a decision tree is to be able to predict the value of a variable with a model created by learning simple decision rules inferred from the data. A decision tree can have a specified depth. The deeper the tree the more complex the decision rules and the model predicts a better value for the variable.

k nearest neighbor classifier is based on locating the $k > 1$ nearest neighbors in the instance space and to label the unknown instance with the same class label as the nearest neighbors instance. In order to find the neighbors a distance rule has to be set. This classifier is susceptible to noise in the training data.

III. Proposed Method

A. Database

The databases used for this project are the Montreal Affective Voices [7] and the Oxford Vocal Sounds. These have kindly been provided by the persons responsible with them.

The Montreal Affective Voices contains recordings of the following emotions: anger, disgust, fear, happiness, pain, pleasure, sadness, surprise and a neutral recording. The Oxford Vocal Sounds contains recordings of the following emotions: happiness, sadness and neutral. As the first database is made up of only 11 recordings for each emotion, the recordings from the second database were needed in order to make the simulations. Also, from the first database only the recordings of happiness, sadness and the neutral state have been selected. In total there were a number of 172 recordings, out of which 65 recordings were for the neutral state, 58 for happiness and 49 for sadness. For each emotion the recordings have been divided in gender: male, female and baby. From these recordings the features have been extracted and a classifier was trained.

B. Features used

There are three main types of features that characterize the human voice. These are:

- spectral features
- energy features
- micro prosody features

The spectral features are: the MFCC coefficients, the fundamental frequency, the minimum value, the maximum value, the first and third quartile, the mean value, the standard deviation, the range at the turn level, the slope in the voiced segments, regression coefficient and its mean square error, formants and their bandwidth, difference between third and second formant, difference between second and first formant – minimum, maximum, mean, standard deviation, range.

The energy features include: the minimum value, the maximum value, the mean, the standard deviation, the range, the slope, the regression coefficient and its mean square error.

From the micro prosody features group the following features can be mentioned: the jitter, the shimmer, noise to harmonic ratio (NHR) and harmonic to noise ratio (HNR).

Out of these features the following have been chosen for research: the minimum value, the maximum value, the range, the mean, the median, the standard deviation, the variance, the regression coefficient of the pitch, the minimum value, the maximum value, the range, the mean, the median, the standard deviation , the variance, and the regression coefficient of the energy and the ratio of the voiced/unvoiced segments of the signal.

C. Algorithm

The data used for the training and testing phases of the algorithm is made up of a matrix containing all of the features extracted for each individual recording. The matrix has as many rows as there are recordings in the database and as many columns as there are features that were extracted.

First, for each recording used the feature vector was computed and a label has been given to it. A label of "0" was given for the neutral recordings, a label of "1" was given for the happy recordings and a label of "2" was given for the sad recordings. In order to compute the feature vector, each recording was read, it was divided into frames of 10 ms and the pitch was extracted. Next, the pitch and the energy features were calculated. The frames in which the pitch was equal to zero was considered an unvoiced segment, so that the ratio of voiced/unvoiced segments was determined.

Two types of tests were made. In the first one all the recordings were used, while in the second one an equal number of recordings per emotion have been used. The results presented in section IV are for the case in which an equal number of recordings have been used for each emotion.

There are 8 combinations of features that are compared in order to find the one that best characterizes the human voice. The 8 combinations are:

1. pitch: minimum value, maximum value, range, mean value, median value, standard deviation, regression coefficient; energy: minimum value, maximum value, range, mean value, median value, standard deviation, regression coefficient; ratio of voiced/unvoiced segments – total of 15 features
2. pitch: mean value, standard deviation, maximum value, minimum value, range, median value; energy: mean value, standard deviation, maximum value, minimum value, range, median value; ratio of voiced/unvoiced segments – total of 13 features
3. pitch: mean value, standard deviation, maximum value, minimum value, range; energy: mean value, standard deviation, maximum value, minimum value, range; ratio of voiced/unvoiced segments – total of 11 features
4. pitch: mean value, variance, maximum value, minimum value, range; energy: mean value, variance, maximum value, minimum value, range; ratio of voiced/unvoiced segments – total of 11 features
5. pitch: mean value, variance, maximum value, minimum value, range, median value; energy: mean value, variance, maximum value, minimum value, range, median value; ratio of voiced/unvoiced segments – total of 13 features
6. pitch: mean value, standard deviation, maximum value, minimum value, range; energy: mean value, standard deviation, maximum value, minimum value, range; ratio of voiced/unvoiced segments – total of 11 features
7. pitch: mean value, standard deviation, maximum value, minimum value, range, regression coefficient; energy: mean value, standard deviation, maximum value, minimum value, range, regression coefficient – total of 12 features
8. pitch: mean value, standard deviation, maximum value, minimum value, range, regression coefficient; energy: mean value, standard deviation, maximum value, minimum value, range, regression coefficient; ratio of voiced/unvoiced segments – total of 13 features.

After the 8 feature vectors were computed the training phase of the algorithm began. As a cross-validation algorithm was implemented, the data was divided in four quarters. For the training part 3 quarters of the data have been used, while the remaining quarter was used for testing. The classifiers tested were linear support vector machines, decision trees and k nearest neighbor.

Some details regarding the implementation of the algorithm are presented next.

D. Implementation

The entire code was written using Python programming language.

For pitch extraction the Snack Sound Toolkit has been used. The toolkit was designed and developed by the Speech, Music and Hearing part of the School of Computer Science and Communication, Royal School of Technology in Norway.

The classification algorithms used for training and testing are the ones implemented in the sklearn Python module and the OpenCV2 module implemented in Python.

The feature vector which had the greatest recognition rate was implemented on Nao robot. When a recording was provided to it, Nao was able to recognize the emotion present in the recording.

IV. Results

As previously stated two Python modules were compared. From these two, the sklearn module provided the greater recognition rates.

The following table summarizes the recognition rates obtained using decision trees and linear support vector machines from the sklearn module.

TABLE I. Recognition Rates using Decition Trees and Linear SVM

Feature vector	Classifier	
	Decision tree	Linear Support Vector Machine

1	0.7948	0.6956
2	0.8461	0.7173
3	0.7692	0.6956
4	0.7948	-
5	0.8461	-
6	0.7179	0.6739
7	0.7692	0.6956
8	0.7435	0.6956

As can be seen in table 1, the best recognition rate were obtained when using the 2nd and the 5th feature vectors. The difference between them is that in the 2nd vector the standard deviation was used, while in the 5th the variance. There is a well known relationship between the two parameters. The standard deviation represents the square root of the variance. This relationship explains why for these two feature vectors the same recognition rate was obtained. Also, from the same table it can be stated that the number of features and their order has an influence on the recognition rate. The recognition rate obtained using the decision tree is greater with almost 10% than the one obtained using the linear support vector machine.

For a graphical comparison each feature vector a confusion matrix was also computed. Figure 1 represents the confusion matrix for the 2nd feature vector, the one with the highest recognition rate.

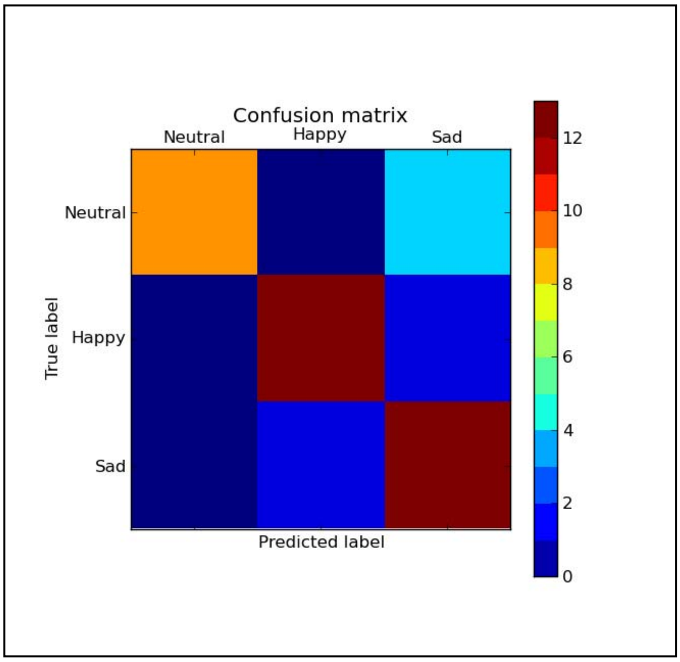

Figure 1. Confusion matrix for the 2nd feature vector

CONCLUSION

In conclusion, the proposed algorithm provides a recognition rate of approximately 85%. When it was tested using the Nao robot, it was capable of recognizing the emotion of the speaker in the recording.

There are three main things that need to be improved in the future. One is represented by the number of recordings in the database. As it is now, the database contains too few recordings to be conclusive in a real life application. New databases need to be found for this purpose.

The second problem concerns the features used. For now the application uses a maximum of 15 features. More tests need to be made with other features as well. The best combination that characterizes the two databases is represented by the 2nd feature vector.

Lastly, other classifiers should be tested as well. The three classifiers used for this project might not be the best ones for the given application. For a future research the decision tree classifier should be compared with other classification algorithms.

ACKNOWLEDGMENT

The project was developed and implemented at École Nationale Supérieure de Techniques Avancées in Palaiseau, France under the guidance of Prof. Adriana Tapus.

I would like to thank the AFCEA Romania organization for funding this research project.

REFERENCES

[1] K. R. Scherer, “Affect Bursts” in Emotions (S. H. M. van Goozen, N. E. van de Poll & J. A. Sergeant, eds.), p. 161-193, 1994. Hillsdale, NJ: Lawrence Erlbaum.

[2] Paul Ekman, W. Friesen, P. Sorenson, “Pan-cultural Elements in Facial Displays of Emotions”, Science, 1969.

[3] Paul Ekman, W. Friesen, J.C. Hager, “Facial Action Coding System – The Manual”, Research Nexus Division of Network Information Research Corporation, 2002

[4] Horia Nicolai Teodorescu, Monica Feraru, “The Statistics of Nonlinear Parameters for the Normal and Emotional Voice”, 1st International Conference on Electronics, Computers and Artificial Intelligenge, ECAI, 1-2 July, 2005, Pitesti, Romania.

[5] Horia Nicolai Teodorescu, Pistol Laura, “Emotional States Recognition in Speech in Romanian – Differences Between Male and Female Speakers” 3rd International Conference on Electronics, Computers and Artificial Intelligence, ECAI, 2-5 July, 2009, Pitesti, Romania.

[6] Monica Feraru, Marius Zbancioc, “Speech Emotion Recognition for SROL Database using Weighted kNN Algorithm”, 5th International Conference on Electronics, Computers and Artificial Intelligence, ECAI, 27-29 June, 2013.

[7] P. Belin, S. Fillion-Bilodeau, F. Gosselin, “The 'Montreal Affective Voices': a validated set of nonlinguistic emotional vocal expressions for research on auditory affective processing”, Behavior Research Methods, 2008, 40(2), 531-539.

[8] B. Boser, I. Guyon, V. Vapnik, “A Training Algorithm for Optimal Margin Classifiers”, 5th Annual Workshop on Computational Learning Theory, pages 144-152, 1992