










Emotional Corpus, Feature Extraction and Emotion Classification Using the Parameterized Voice Signal

Miriam Martínez-Arroyo¹ , Vicente Bello-Ambario¹ ,
José Antonio Montero-Valverde¹ , Eduardo de la Cruz-Gómez¹ ,
Mario Hernández-Hernández² , and José Luis Hernández-Hernández³  

¹ Division of Research and Graduate Studies, TecNM/Technological Institute of Acapulco, Acapulco, Mexico

{miriam.ma, ponente.msc, jose.mv, eduardo.dg}@acapulco.tecnm.mx

² Faculty of Engineering, Autonomous University of Guerrero, Chilpancingo, Mexico
mhernandezh@uagro.mx

³ Division of Research and Graduate Studies, TecNM/Technological Institute of Chilpancingo, Chilpancingo, Mexico

joseluis.hernandez@itchilpancingo.edu.mx

Abstract. The recognition and classification of human emotions through voice analysis, it is a very interesting research area, due to the wide variety of applications: telecommunications, learning, human-computer interface, entertainment, etc. In this investigation a methodology is proposed for the recognition of emotions analyzing voice segments. The methodology is mainly based on the fast Fourier transform (FFT) and Pearson's correlation coefficients. The tone (pitch), the fundamental frequency (Fo), the strength of the voice signal (energy) and the speech rate have been identified as important indicators of the emotion in the voice. The system consist of a graphical interface that allows user interaction by means of a microphone integrated into the computer, which automatically processes the data acquired. In our environment, human beings are programmed to let our voice flow, in multiple ways to communicate and to capture through it emotional states. There are various investigations where the Berlin database is used, which is free and many researchers have used it in their research. However, the creation of an emotional corpus with Spanish phrases, was needed for testing that provide clearer results. The corpus contains 16 phrases per emotion created by 11 users (9 women and 2 men) with a total of 880 audio samples. The following basic emotions were considered: disgust, anger, happiness, fear and neutral. Results obtained indicate that the emotion recognition algorithm offers an 80% of effectiveness.

Keywords: Emotional state · Parameterization · Statistical models · Pattern recognition

1 Introduction

The most relevant research would be directed to the study of emotions through facial expressions, although in 1872 Charles Darwin himself pointed out the importance of

non-verbal aspects, vocal behavior and situations where the individual uses speech as a means of expression [2]. Technical difficulties leading to evaluation of vocal behavior and the merger with respect to its status as non-verbal behavior; are two of the reasons that have led to this mismatch with respect to facial expression [3, 4].

Studies on vocal expression of emotion can be classified into two large groups:

- **Expression Studies.** They have as basic objective determine how an emotional state is expressed or is externalized through the voice of the subject [5], its evaluation can be carried out at two levels: acoustic analysis and perceptual judgments of acoustic parameters.
- **Recognition Studies.** They try to know to what extent the receiver is able to identify, through the non-verbal aspects of the speaker's voice and the emotional state expressed by him.

One of the main problems what research on voice emotion recognition has faced [6], has been the separation of the verbal channel from the vowel, so that the linguistic content of the message does not determine making judgments on the affective state of the transmitter. It has been established that speech is an acoustic event, containing important information on the functioning of the central nervous system, and therefore carries information about the emotional state of a person.

Emotions are physically “reactions that represent modes of adaptation to certain stimuli in the individual and that affect the responses of different biological systems, including the voice” [7]. Emotions a very powerful force in human behavior, in addition to having a great influence on people's health, both physical and mental. A clear example of this fact is that a person who is excited, is able to do some action that never would do during a stable emotional state that is paralyzed by fear or sadness. However, one of the problems encountered by researchers when working with emotions, is that they present a high degree of subjectivity. This is due to the way they express themselves and it depends largely: of the speaker, culture of origin and of the environment.

Each emotion generates a set of reactions, they are automatic affective states and complexes that are unleashed in our body, which also fulfil a beneficial function of adaptation of the organism, by favoring our reaction to a stimulus. The emotions considered in this research are known as primary emotions established in Ekman's model (surprise, disgust, sadness, anger, fear and happiness) that cause a set of reactions in the human being [8–10]. A brief description of the primary emotions is provided below:

- **Surprise.** This is an emotional reaction to an unexpected situation and its main function is to make the nervous system adjust to an unforeseen change in the environment, eliminating the possible interferences that could appear.
- **Disgust.** It is characterized by a low medium tone, a wide range, the slowest speech rate with big pauses. This gesture blocks the nostrils in case of perceiving an unpleasant substance and to facilitate the need to expel it.
- **Sadness.** Sad speech exhibits a medium tone lower than normal, a narrow range and a slow speed of speech. Energy decreases, motivation is reduced, metabolism is reduced and social isolation is increased. Sadness arises after the occurrence of an emotionally

damaging situation. Responds to survival instinct, as it alerts us to a harmful situation. In addition, promotes the search for social support.

- **Anger.** It is defined as “the unpleasant and annoying impression that occurs in the mood”. This is a negative emotion that prioritizes sending blood to the hands and legs. Adrenaline levels are elevated, is associated with hostility, the fury and in the most extreme cases; with a pathological and permanent hatred.
- **Afraid.** This primary emotion is a way of warning us that a danger is approaching for which we are not prepared. Fear is awakened after a stimulus that makes us anticipate such danger. Helps the person perform some behavior to distance themselves from potential danger, therefore promotes escape behaviors, avoidance and prevents dealing with dangerous situations.
- **Happiness.** Manifests itself in an increase in the mean tone and in your range; as well as an increase in the speed of locution and in intensity. It is associated with positive feelings such as pleasure, the euphoria and the suppression of negative feelings. It is produced therefore, a sense of calm and well-being.

For an emotion to be considered primary, it must meet the following requirements:

- Have a specific neural substrate and that differentiates it from the rest. That is to say, that the neurons involved are specific ones.
- That the facial expression representing that emotion is unique and distinctive.
- That the feelings underlying the emotion, namely; the thoughts and the way we feel, they are characteristics of that emotion.
- It has an evolutionary and survival purpose.
- It has adaptive functions in a certain situation.

There are several research papers, where they indicates that some emotions are better recognized than others. Fechner [11] argues that the better recognition of negative emotions can be explained in terms of their greater survival value. In another case, in studies with adults, Scherer [3] handle the hypothesis which points out that anger is often the most widely recognized emotion, followed by sadness, indifference and happiness.

A corpus was built that contains 16 sentences per emotion created by 11 users (9 women and 2 men) with a total of 880 audio samples. Are considered the following basic emotions: disgust, anger, happiness, fear, and neutral. The algorithm of Emotion recognition gives 80% effectiveness in the results obtained.

2 Related Work

Several works have been carried out related to the identification of basic emotional states by analyzing audio segments in different languages.

In [22] Meftah et al. designed an emotional corpus in Arabic language in order to identify the following emotional states: neutral, sadness, happy, surprised, and questioning, for this the authors building and selected 16 sentences that were read by 20 male and female native speakers. In this sense, a similar work was developed by Constantini et al. [21] they built a database with Italian emotional speech from the voices of up to 6

actors who played 14 sentences simulated the next emotional states: disgust, fear, anger, joy, surprise, sadness and neutral.

A Database with the participation of actors and non-actors was developed by Banotu et al. This database provides a potential resource to analyze emotions in the speech signal. From the design of drama situations to annotation, there are 5 stages involved in the data collection process. A different annotation procedure is adopted by the mixture of context-dependent and context-independent ways. There are some utterances in the database where there is an occurrence of multiple emotions. Perceptual and acoustic analyses are carried out to understand differences in the data among actors and non-actors. Perceptual studies indicate that the actors intend to convey emotions in speech without affecting the linguistic message much. It is observed that there is more jitter in non-actors data when compared to actors [24].

Lyatsko et al. designed “EmoChildRu” which is considered as the first database containing emotional speech material from 3–7 year old Russian children. The database includes 20.340 recordings (30 h) of 100 children growing in families [23].

3 Materials and Methods

This section deals with the process of converting the original analog sound waves into digital signals for storage and reproduce them later; snapshots of analog sounds are taken and are stored. The speed with which the device takes samples is called the Sample Rate (SR) and will be expressed in Hertz or Kilo Hertz (Hz, Khz). 1 Hz will be 1 sample per second and 10 kHz is 10,000 samples per second. Each of these samples will be assigned a value corresponding to the amplitude of that instant in the original signal (Quantization).

For the case study, only the emotions of: disgust, anger, afraid, happiness, and neutral (no emotion).

Figure 1 shows the methodology used for the classification and recognition of emotions through the parameterized voice signal.

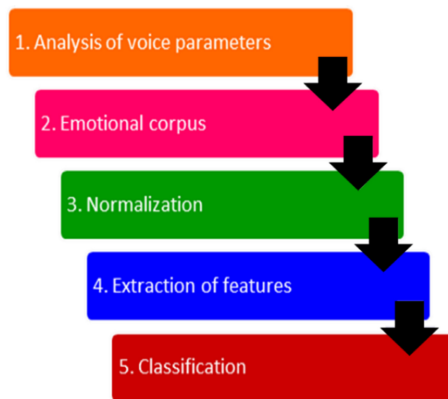


Fig. 1. Methodology implemented in recognition of emotional state.

3.1 Analysis of Voice Parameters

The voice is nothing but a sound and as such, it is characterized by a series of elements. The features that have been most recurrent in literature are the pitch, duration, voice quality, global pulse shape and vocal tract. García [12], divide his work into 13 features that have been used for emotion detection in the voice. These parameters are described in Table 1.

Table 1. Features used in Speech Emotion Recognition [12].

Features used	Description
Bandwidth	This range is measured in Hertz (Hz)
Vocal tract areas	Number of harmonics caused by non-linear airflow in the vocal tract that produces the voice signal
Spectral characteristics	Energy content of frequency bands divided by the sample length
Speech Activity Detection	This property is defined as the rhythmic profile of speech
Duration	Difference between the beginning and end time of a spoken sequence, obtaining a duration rate for emotional and neutral sentences
Energy	It is the value of the physical magnitude that expresses the greater or lesser amplitude of the sound waves
Formants	They are frequencies reinforced by resonance
Intensity	It is measured in Decibels (dB)
LPCs (<i>Linear Prediction Coefficients</i>)	A set of equivalent essential formulations for modeling a given waveform
MFCCs (<i>Mel Frequency Cepstrum Coefficients</i>)	Technique of fractioning the initial signal into a discrete set of spectral bands containing analogous information
Pitch	Represented as F0 (Fundamental Frequency)
Zero crossing rate	Represents how many times the signal changes sign passing through the abscissa axis
Speaking rate	The proportion of segmental units, syllables and pauses per unit of time produced by a speaker

Table 2 presents a summary of the relationships between emotions and the parameters of discourse. As is described, only five emotions appear. These correspond to the primary or basic emotions. It is known that there is a relationship between the prosodic information and the expression of emotions in speech; traits such as intensity, the fundamental frequency curvature and the speed of speech, are important characteristics of the discrimination of emotions in the voice [15–17].

Table 2. Comparison of emotions.

	Happiness	Anger	Disgust	Afraid	Sadness
Speed of the speaker	Slightly accelerated with increase	Slightly accelerated	Slow	Very accelerated	Paused
F ₀	Increase in mean, variability	Increase in the mean, median and variability	-----	Increase in mean F ₀ , disturbance, variability of F ₀ movement	Below the normal mean F ₀
Articulation	Normal	Tense	Normal	Accurate	Dragged
Intensity	High	High	Slow	Normal	Slow
F ₀ average	High	High	Slow	High	Slow
Spectrum	Increased high frequency energy	Raised at the midpoint	-----	Increased high frequency energy	Decrease in high frequency energy
Others	Irregular accent distribution	Speak cut off	-----	Irregularity in the sound	Rhythm with irregular pauses

3.2 Emotional Corpus

The speech corpus consists of 880 Spanish sentences recorded by TISM (Technological Institute of San Marcos) students with a range in ages between 18 and 26. These sentences express 5 different emotional states: disgust, anger, happiness, fear and neutral with a total of 16 sentences for each of them. We have chosen sentences whose semantic content does not imply any particular emotion so the classification can be made on the basis of prosodic details [15]. Figure 2 shows the graphical interface used to create the emotional corpus.

As shown in Fig. 3, the audio recording was done in a closed classroom, located in the computer lab of the Technological Institute of San Marcos in the state of Guerrero, (TISM), in order to reduce noise and distractions.

There are two important factors during this process. For code development, parameters must be changed to determine what works best in the algorithm. Using a desktop program, audio is recorded with a sampling frequency of 44100 Hz and an audio rate of 16 bits. One channel (Mono) is used, resulting in a vector of thousands of data, from which non-significant data will be discriminated.

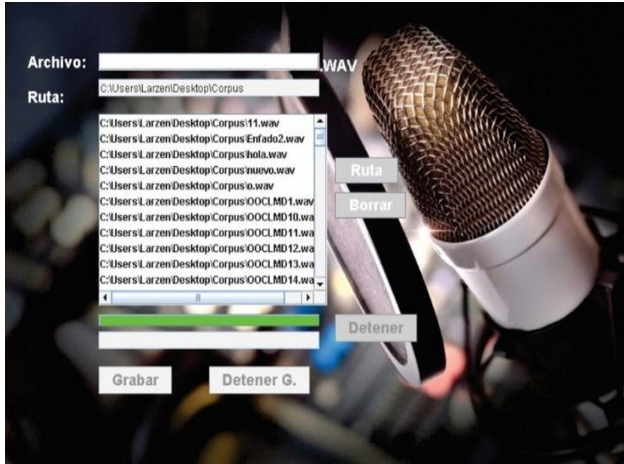


Fig. 2. Main interface of the voice recorder.



Fig. 3. TISM students in the recording process.

Table 3 includes some example sentences from the corpus used as a test to detect the corresponding emotion.

Table 3. Some example sentences from the corpus used.

No.	Phrases used
1	Times are not like before
2	What are you talking about?
3	Do you want some advice?
4	Homework is for tomorrow
5	He is the group leader
6	If true
7	I don't think so, don't be a gossip
8	You are always late
9	Can you keep quiet please?
10	If you don't like it, do it yourself
11	My mom's computer is broken
12	The school is painted blue
13	Would you live with me
14	My point of view is another
15	This activity does not correspond to me
16	There is a madman

3.3 Normalization

In general, it is understood that normalization is the operation by which a set of values of a given magnitude is transformed into others in such a way that the latter belong to a predetermined scale.

It is possible to normalize a set of values in the interval $[0, 1]$ by applying for each value the transformation shown in Eq. 1.

$$v_i = \frac{a_i - \min}{\max - \min} \quad (1)$$

Where a_i is the value to be transformed, \min and \max are the minimum and the maximum of the set of values and v_i is the normalized value.

The normalization consists of processing the acoustic signal and finding the optimal set of features that allow to perform the classification of emotions. The function algorithm that normalizes the data of a numerical vector, which it receives as a parameter, is as follows:

- Returns the maximum absolute value of the vector to transform
- Returns the number of elements of the vector to be transformed (Vector size = n)
- Returns a vector of zeros of n rows and 1 column
- A loop is made where the initial value of i is 1 and is increasing in 1 until it becomes the value of n
- The vector is divided at the position i between its absolute maximum values.

3.4 Feature Extraction

This module consists of grouping the spectral acoustic characteristics, as these describe the properties of a signal in frequency domain by means of harmonics and formants, prosodic information (volume, velocity, duration) is also extracted. The algorithm for feature extraction is the fast furrier transform (FFT) which performs the following:

- The first 60000 values of the vector are cut.
- The absolute value of the Fourier transform [18] of the recording is obtained.
- Multiply the result by the conjugate of the original vector.
- Accepts only frequencies above 150 Hz.
- Normalize the vector using the Euclidean norm.

The Euclidean norm (also called vector magnitude, Euclidean length, or 2-Norm) of a vector v with the elements of N is defined by Eq. 2.

$$\|v\| = \sqrt{\sum_{k=1}^N |v_k|^2} \quad (2)$$

FFT is the usual abbreviation for Fast Fourier Transform, and is an efficient algorithm that allows to compute the discrete Fourier transform and its inverse given vectors of length N by Eq. 3.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi k\omega t \frac{n}{N}} \quad (3)$$

The FFTs of each section are obtained, containing 5 vectors for each emotion with the objective of generating a surface in which the frequencies and their variation in time can be observed. The FFTs of each section are averaged to obtain a pattern of the pronounced phrase.

3.5 Classification

Pearson's correlation coefficient is defined as an index that can be used to measure the degree of relationship of two variables as long as both are quantitative and continuous. Pearson's correlation coefficient is an easily performed index. In the first instance, its absolute values range from 0 to 1. If we have two variables X and Y , then we define Pearson's correlation coefficient between these two variables as $r_{x,y}$ as shown in Eq. 4.

$$r_{x,y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \quad (4)$$

3.6 Description of the System that was Developed

The development of the project was carried out in a period of 13. Figure 4, shows the general structure of the proposed system, starting with the capture of voices to have the emotional corpus. For the emotional corpus design, Students from the San Marcos Technological Institute in the state of Guerrero, Mexico, were invited to collaborate.

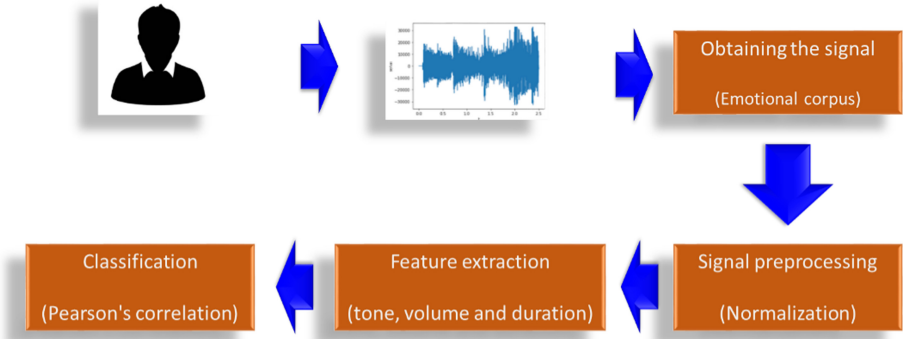


Fig. 4. Block diagram of the general structure of the proposed system.

4 Results

A Mexican emotional corpus has been created for the testing of the speech emotion recognition algorithm using a statistical method as a classifier. It is worth mentioning that more features are expected to be added to the vector to increase the efficiency of the recognizer using Machine Learning techniques.

Summary of results: In the processing stage, the audio signal was processed as shown in Fig. 5.

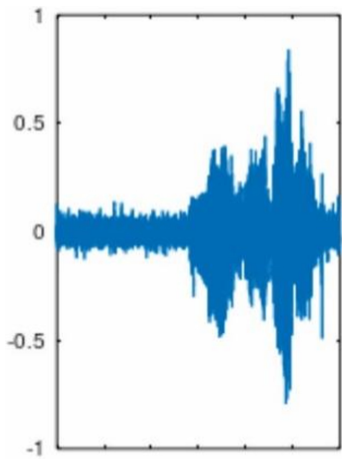


Fig. 5. The phrase: “You will live with me” recorded by TISM students.

In the feature extraction stage, we were able to obtain the frequency spectrum containing a vector with patterns necessary to detect the 5 emotions shown in Fig. 6.

In the classification stage we used the differences between the vector to be classified and the feature vectors stored in the database by means of Pearson’s correlation, detecting the differences by means of the error coefficient. Table 4 shows the success in detecting

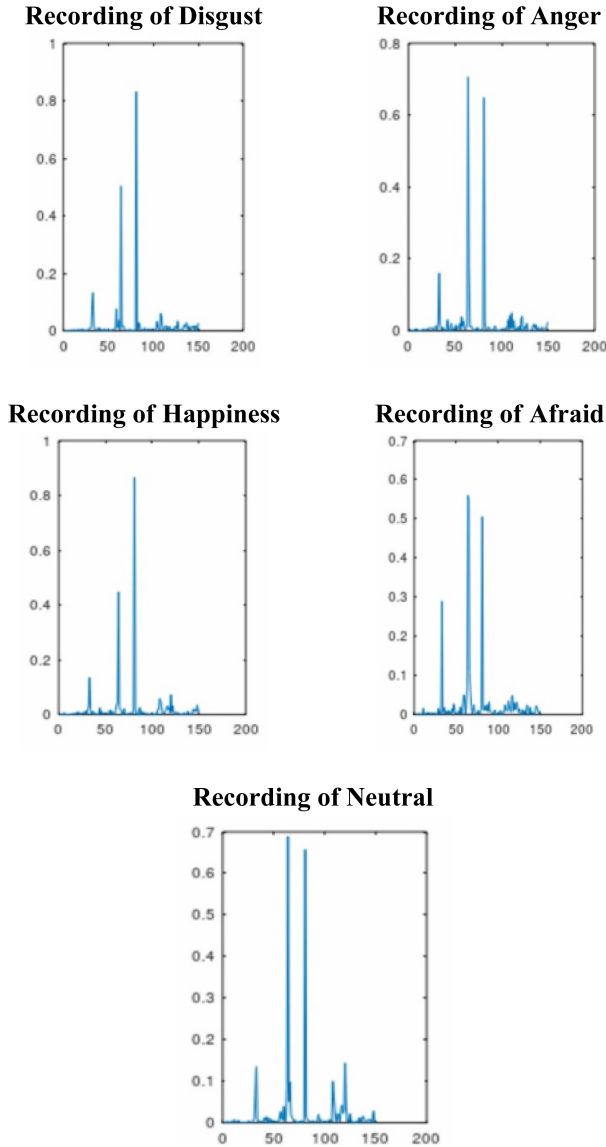


Fig. 6. Pattern of the sentence pronounced in each emotion.

the emotion “Disgust” by detecting the error coefficient which is the closest to 0 and thus pointing out the most significant similarity in the vector of features with emotion to be recognized.

Table 5 shows the confusion matrix of the algorithm used in this work where it can be observed that the neutral emotion has greater confusion than the other emotions, it is also worth mentioning that disgust and anger are emotions that are clearly identified with greater accuracy by this classifier.

Table 4. Recognition of “Disgust” using the sample correlation method.

PEARSON CORRELATION	0.15327
Error coefficient disgust	0.018317
Error coefficient anger	0.021492
Error coefficient happiness	0.022185
Error coefficient fear	0.020861
Error coefficient neutral	0.052955
Identified emotion:	DISGUST

Table 5. Confusion matrix for the classification algorithm using Pearson’s correlation.

O B S E R V A T I O N S		PREDICTION					Totals
		Disgust	Anger	Happiness	Afraid	Neutral	
	Disgust	143	8	9	5	11	176
	Anger	23	134	12	2	5	176
	Happiness	25	13	116	4	18	176
	Afraid	18	14	9	115	20	176
	Neutral	39	34	21	2	80	176
	Totals	248	203	167	128	134	880

5 Conclusions

A research of the acoustic parameters for the recognition of emotional states in the voice was carried out in the area of Intelligent Systems. It was achieved to have an algorithm capable of recognizing 80% of the sentences with emotion acted by TISM students. The results show the need for more parameters in the feature extraction step. It was necessary to creation a corpus due to the lack of standardization in the elicitation of emotions and the non-existence of norms that guarantee reproducibility. Although the initial results are satisfactory, we considered that, it is essential to use more classification methods and artificial learning techniques in order to have a higher classification efficiency.

1. References

1. Darwin, C.: The Expression of the Emotions in Man and Animals. University of Chicago Press (2015)
2. Carrera, M.J.M., Fernández, A.J.: El reconocimiento de emociones a través de la voz. *Estud. Psicol.* **9**(33–34), 31–52 (1988)

3. Scherer, K.: Speech and emotionnal states. *Speech Evaluation in psychiatry*. J. Darby. New York, Grune and Stratton (1981)
4. Scherer, K.R.: Methods of research on vocal communication: Paradigms and parameters. *Handbook of methods in nonverbal behavior research*, pp. 136–198 (1982)
5. Brown, B.L., Bradshaw, J.M.: Towards a social psychology of voice variations. In: *Recent Advances in Language, Communication, and Social Psychology*, pp. 144–181. Routledge, London (2018)
6. Zhao, M., Adib, F., Katabi, D.: Emotion recognition using wireless signals. In: *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pp. 95–108, October 2016
7. Levenson, R.W.: Human emotion. a functional view. In: Ekman, P., Davidson, R.J. (eds.). *The Nature of Emotions: Fundamental Questions*, pp. 123–126. Oxford University Press, New York (1994)
8. Kim, E.H., Hyun, K.H., Kim, S.H., Kwak, Y.K.: Speech emotion recognition using eigen-fft in clean and noisy environments. In: *The 16th IEEE International Symposium on Robot and Human interactive Communication. RO-MAN 2007*, pp. 689–694. IEEE (2007)
9. Solís, V.J.F.: *Modelo de procesamiento de voz para la clasificación de estados*. PhD thesis, Instituto Politécnico Nacional. Centro de Investigación en Computación (2011)
10. Plutchik, R.: The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* **89**(4), 344–350 (2001)
11. Fechner, E.H.: *Children’s understanding of the nonverbal communication of emotion in the visual, vocal and gestural* (Doctoral dissertation, ProQuest Information & Learning) (1978)
12. García, M.J.V., Cobeta, I., Martín, G., Alonso-Navarro, H., Jimenez-Jimenez, F.J.: Acoustic analysis of voice in Huntington’s disease patients. *J. Voice* **25**(2), 208–217 (2011)
13. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. *Speech Commun.* **41**(4), 603–623 (2003)
14. Montero, J.M.: *Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano* (Doctoral dissertation, Tesis Doctoral, ETSI Telecomunicación, UPD) (2003)
15. Ortego Resa, C.: *Detección de emociones en voz espontánea* (Bachelor’s thesis) (2009)
16. Cowie, R., et al.: Emotion recognition in human computer interaction. *IEEE Signal Process. Mag.* **18**(1), 32–80 (2001)
17. Hasrul, M.N., Hariharan, M., Yaacob, S.: Human affective (Emotion) behaviour analysis using speech signals: a review. In: *International Conference on Biomedical Engineering (ICoBE)*, vol. 27, p. 28, February 2012
18. Nussbaumer, H.J.: The fast Fourier transform. In: *Fast Fourier Transform and Convolution Algorithms*, pp. 80–111. Springer, Heidelberg (1981). <https://doi.org/10.1007/978-3-642-81897-4>
19. Darwin, C.: 1965. *The Expression of the Emotions in Man and Animals*. John Marry, London (1872)
20. Ortego Resa, C. et al.: *Detección de emociones en voz espontánea*. B.S. thesis (2009)
21. Costantini, G., Iaderola, I., Paoloni, A., Todisco, M.: EMOVO corpus: an Italian emotional speech database. In: *International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 3501–3504. European Language Resources Association (ELRA) (2014)
22. Mefteh, A., Alotaibi, Y., Selouani, S.A.: Designing, building, and analyzing an Arabic speech emotional corpus. In: *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, p. 22, May 2014

23. Lyakso, E., et al.: EmoChildRu: emotional child Russian speech corpus. In: International Conference on Speech and Computer, pp. 144–152. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23132-7_18
24. Banothu, R., Botsa, K., Paidi, G., Suryakanth, V.: IIIT-H TEMD semi-natural emotional speech database from professional actors and non-actors. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 1538–1545. European Language Resources Association (2020)