Journal of Computer Science

Original Research Paper

VERBO: Voice Emotion Recognition dataBase in Portuguese Language

[1,2]José R. Torres Neto, [3]Geraldo P.R. Filho, [1,4]Leandro Y. Mano and [1]Jó Ueyama

[1]*Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil*
[2]*School of Eletrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada*
[3]*Institute of Computing, University of Campinas, Campinas, Brazil*
[4]*Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Groningen, Netherlands*

Article history
Received: 07-08-2018
Revised: 02-10-2018
Accepted: 02-11-2018

Corresponding Author:
José R. Torres Neto
Institute of Mathematical and Computer Sciences,
University of São Paulo, São Carlos, Brazil
Email: jrtorresneto@usp.br

Abstract: The recognition of human emotional traits based on Affective Computing is being carried out by computational systems that are able to interpret and react intelligently to the context of the user. Speech Emotion Recognition systems are capable of transforming speech signal data into information related to the feelings of individuals in specific situations. However, the emotional expression of a human being depends mainly on his origins. For this reason, emotional voice databases are peculiar to each language. In this paper, we propose a new emotional database with speech in the Portuguese language of Brazil, called Voice Emotion Recognition dataBase in Portuguese language (VERBO). The database was validated by a panel of expert judges and we achieved an agreement rate of 76% using the content validity index and substantial agreement rate of 65% using Fleiss' Kappa. In addition, an accuracy of 0.76 was achieved and it was possible to observe that the emotions anger and happiness were more easy to recognize showing 0.85 and 0.83 of f1-score, respectively, whereas the disgust and surprise emotions were the most difficult showing 0.67 and 0.68, respectively. In view of this, the main contributions to research made by this study are: (1) The establishment of a new actuated voice database; (2) support provided by voice recognition systems for the analysis of feelings and emotions; and (3) statistical validation of the database using CVI and Fleiss kappa.

Keywords: Database, Knowledge Base, Emotion Recognition, Analysis of Feelings, Speech

Introduction

Affective Computing is an interdisciplinary area that includes computer science, psychology and cognitive science, while providing a computational interface through devices capable of recognizing, interpreting, processing and simulating human affections Picard (2010); Swain *et al*. (2018); Gonçalves *et al*. (2016). Human emotional traits have long been studied in the field of psychology and there has been a significant growth in research on the recognition of emotions based on Affective Computing Lichtenstein *et al*. (2008); Picard (2010).

Studies on emotional discourse have confirmed that there is a close correlation between speech and emotion Costantini *et al*. (2014). The speech signal in human speech is a fast and easy way to understand communication and is regarded as being of great practical importance in an emotion recognition system by the voice (Speech Emotion Recognition - SER) Mano *et al*. (2016d). In addition to the syntactic and semantic clues that speech transmits, human emotional and physical states can be recognized from the voice signal processing Jurafsky and Martin (2000). SER systems are capable of transforming data from speech signals into information related to the feelings of individuals in particular situations, for example customer reactions to telemarketing services. Thus, it is feasible to make use of speech patterns for the automatic recognition of the emotional state of human beings Rázuri *et al*. (2015).

The recognition of emotions by voice has attracted the attention of researchers and become widespread in several developed countries, such as the United States,

Germany and Italy Ververidis and Kotropoulos (2003); Costantini *et al*. (2014); Meddeb *et al*. (2017); Swain *et al*. (2018). Hence, several branches of the academic and industrial world have emerged to deploy these systems in real-world scenarios (e.g., call centers, auxiliary diagnosis systems, remote education, safe direction and computer games) with the aim of providing opportunities for interaction with individuals as well as interactive TV, virtual teacher training, carrying out studies of human brain dysfunction and designing advanced systems to convert texts into speech Costantini *et al*. (2014); Jing *et al*. (2018).

The analysis of feelings by these systems is generally conducted through word processing and based on the transcription of words uttered by individuals. In some cases, in addition to the transcription of the words, there is a need to translate them into another language before carrying out the analysis, which requires an extra stage in the processing Shadiev *et al*. (2017). One of the limitations of this situation is that the reliability of these systems depends on what the individual is talking about. This is corroborated by the claims of some psychologists that it is common for people to hide and be reticent about what they really feel, which is often the case with people suffering from depression Apesoa-Varano *et al*. (2015). In addition, a simple text that does not reveal any emotion does not provide a suitable semantic representation Jurafsky and Martin (2000).

A database that is capable of representing human emotions, is an essential requirement in an emotion recognition system Costantini *et al*. (2014). The emotions expressed by the speech vary from one person to another and depend on their particular background and social origins Meddeb *et al*. (2017). For this reason, several databases of emotional voice have been designed in different languages Costantini *et al*. (2014); Meddeb *et al*. (2017); Busso *et al*. (2017). As far as we know, there is no emotional discourse database with audios available in Brazilian Portuguese, despite advances made by research in this area. With this in mind, this study is guided by the following research question: How can we create an emotional discourse database in Brazilian Portuguese? However, the creation of an emotional voice database capable of representing people's emotions is not a trivial task, because a number of factors must be taken into account, including the following: (1) What type of database is accepted by the scientific community? (2) What type of emotions can be found in the relevant literature? (3) Who will form the sample of participants during the collection of the audios? (4) What linguistic material will be used? (5) How will the audios be recorded and made available for access by the scientific community?

Thus, in addressing these questions, we intend to design a new emotional database in Brazilian Portuguese for a speech emotion recognition system, called Voice Emotion Recognition dataBase in pOrtuguese language (VERBO) and available online from website (https://sites.google.com/view/verbodatabase/). In addition, the VERBO database was validated by a panel of experts with clinical experience and evaluated by the Content Validity Index (CVI) and Fleiss' Kappa test, where it obtained a verbal agreement of 76 and 65%, respectively. The objectives of this article can be summarized as follows: (1) Establishing a new voice actuated database; (2) assisting voice recognition systems for the analysis of feelings and emotions and; (3) carrying out a statistical validation of the database using CVI and Fleiss' kappa.

Section 1 provides the emotion background and how can it be measured. Section 2 describes the creation of the VERBO database. Section 3 presents the VERBO database analysis and validation. Section 4 discusses the applicability of the VERBO database. Finally, Section 5 shows the conclusion of the study.

How to Measure Human Emotions?

Human emotions are hard to identify and measure, as there are difficulties in accurately distinguishing between the emotional states of other individuals and sometimes even those of oneself. Thus, a system for quantifying emotions must be defined so that they can be classified with the aid of computing techniques. In view of this, some models have been designed to help psychologists evaluate emotional experiences Russell (1980); Scherer (2005).

Russell (1980) developed a circumplex model that treats different emotional states as two-dimensional entities in a continuous space. These can be represented by a plane where the axes refer to "pleasure" (valence) (which may be high or low) and "arousal" (which corresponds to the emotional energy levels). Studies carried out with people from different social backgrounds suggest that Ekman (1992), the various emotions that can be perceived through the circumplex model, constitute a set of 6 basic emotions (i.e., happiness, disgust, fear, anger, surprise and sadness) from which all emotional states can be derived and any individual, regardless of the social or ethnic group to which he/she belongs, is able to express these basic emotions in the same way Russell (1980).

Figure 1 is an illustration of the basic emotions and some of their derivations in the circumplex model designed by Russell (1980). The circumplex model is represented as a circle in which the basic emotions are arranged in accordance with their degree of pleasure and arousal. The pleasure axis controls how positive (high) or negative (low) the emotional feeling is, whereas the arousal axis represents the level of energy caused by the emotion Russell (1980).

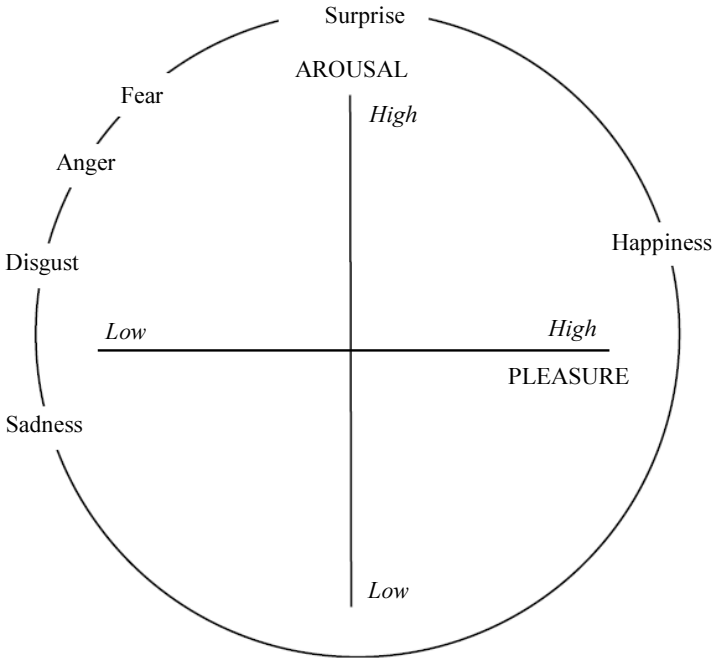

Fig. 1: Russell's circumplex model Russell (1980)

In their everyday lives, people tend to use a categorical system to identify the emotions of others because of their similarity to their past experiences. This is an intuitive process, but is related to the method employed for classifying human emotions adopted by computer systems. Thus, the spread of Affective Computing has increasingly been a driving-force behind the analysis of emotional states Picard (2010).

It is well known that each human emotion is related to the variable level of interrelated physiological changes in motor representations, between facial, body and oral expressions Mano *et al*. (2016d). Motor expressions, also known as expressive reactions, are responsible for communicating behavioral tendencies and involve changes in facial, vocal and gestural expressions that reflect the user's emotional experience Scherer (2005); Mahlke and Minge (2008); Mano *et al*. (2016b). In the context of this study, Fig. 2 shows that the voice undergoes changes in accordance with the degree of pleasure and arousal, as expressed in terms of emotional responses; this is reflected in speech features, such as speed, intensity, melody and sound Mano *et al*. (2016d); Chen *et al*. (2012). Thus, it is possible to determine different characteristics with regard to emotional states.

VERBO - Voice Emotion Recognition dataBase in Portuguese Language

This section shows how an emotional database was established in Portuguese with the aim of contributing to the advances being made in the recognition of emotions through human discourse. The database that was created consists of audios capable of representing human emotions for systems for the emotions recognition and analysis of feelings. However, the database can also be integrated with other health monitoring applications, industrial applications or academic research in Brazil. In the next subsections, the creation process of the VERBO database is presented by a sequence of stages.

Stage 1: Defining the Database Type

Several types of databases designed for emotional discourses have been created in different languages by various research groups, since different countries or regions affect the tone and rhythm of the voice in human discourse Ververidis and Kotropoulos (2003); Meddeb *et al*. (2017); Swain *et al*. (2018). Two types of database are often referred to in the literature on the emotional expression of human beings through discourse. These can be divided into the following categories Ververidis and Kotropoulos (2003); Swain *et al*. (2018):

- Speech performed corpus: This consists of audio recordings of human speech made by actors, who utter some pre-defined words, phrases or texts, that express all the emotions of each sentence
- Natural speech corpus: This consists of gatherings of people who are placed in controlled situations so that they can express their emotions naturally

DOI: 10.3844/jcssp.2018.1420.1430

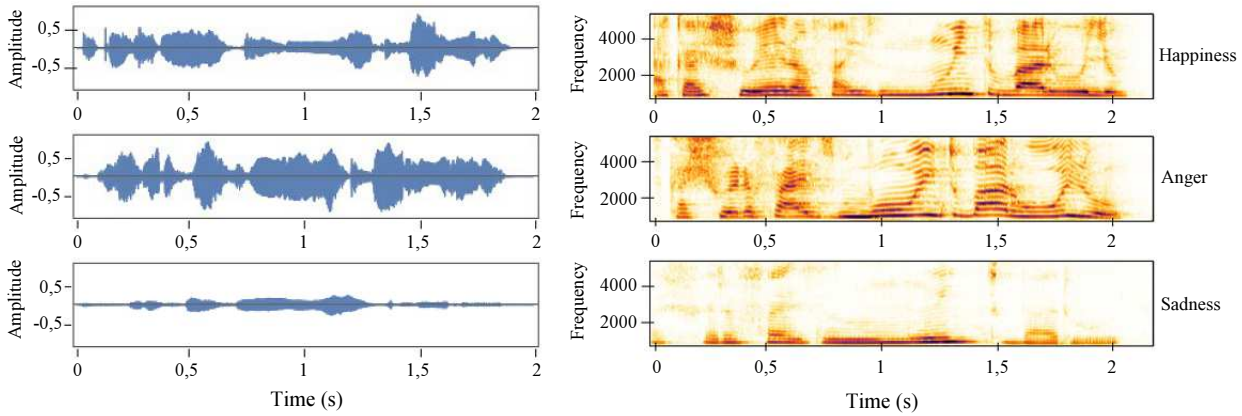

Fig. 2: Changes in the frequency and energy of the voice to reflect different emotions

The great advantage of speech performed corpus is its simplicity. It can be observed that the acting follows a script in which the actor is aware of the text to be pronounced and of the emotion to be expressed. Thus, the database can be divided into directories separated by actor and/or emotion.

On the other hand, a natural speech corpus is more complex. It consists of a wide range of different types of recordings, such as radio broadcast recordings, television audio signals, recorded conversations in offices or residences, lectures, classes, interviews, etc. There is no control over the emotional aspects of the recordings. Thus, in addition to the audio recordings, the database comes with many files containing transcriptions that show the time intervals contained in each file where the vocal expressions of emotions occur. For this reason, it is necessary to trim the audio files before extracting the required features, so that it is possible to concentrate on segments containing vocal expressions of emotions and not other types of vocalizations.

While the natural speech corpus causes several difficulties and has serious limitations, the actuated discourse database is known for its simplicity and does not require pre-processing before the required features are extracted, as well as being generally accepted by the scientific community Ververidis and Kotropoulos (2003); Meddeb *et al*. (2017); Swain *et al*. (2018). We created an actualized emotional discourse database on the basis of these differences.

Although the acted database needs fewer resources, other challenges arise from the creation of a speech corpus for performance. For instance, there is no standardization with regard to the number of sentences or structure of the linguistic material or the number of actors, which determines the total number of audios in the base and whether they will be able to represent the emotions properly. Other difficulties are obtaining the resources for the selection of the actors and actresses and finding a suitable environment for the recordings and equipment.

Stage 2: Definition of Emotions

Emotion plays an important role in communication. This factor is supported by intercultural studies carried out by Ekman (1992), who argued that individual emotions are expressed in terms of 6 basic emotions, which are happiness, disgust, fear, anger, surprise and sadness. These are the key emotions in studies carried out in the Busso *et al*. (2017); Swain *et al*. (2018). Ekman (1992) states that more complex emotions are defined by joining together more than one basic emotion, from devices for measuring emotions, such as the circumplex model Russell (1980), in which emotions can be measured through the valence and arousal axes.

In view of this, the emotions defined in the actuated emotional discourse database set out in this article, were defined on the basis of the 6 basic emotions categorized by Ekman, with the addition of the neutral zone as the 7o emotional state.

Stage 3: Choice of Actors for the Recordings

The professionals chosen as the sample were of different ages and sexes and came from different regions from Brazil. Twelve professionals were used in the creation of the VERBO database, which was phonetically balanced between 6 actresses and 6 Brazilian actors.

The actors chosen for the creation of the VERBO database are members of different theater groups in Brazil and all have e between 2 and 23 years experience of acting in a professional capacity. In addition, all of them act on a regular basis, between 2 times a week (23.1%) and 5 or more times a week (46.2%). Although academic training was not a criterion for the choice of actors, 83% of the actors have a technical or higher education diploma in Theater or the Arts.

Stage 4: Definition of Linguistic Material

The linguistic material for the recordings was based on the EMOVO Costantini *et al*. (2014) database and adapted to Brazilian Portuguese. The material consists of

14 phrases that had been validated by a professional linguist so that the audios could express all the phonemes of the Portuguese language for all the predefined emotions.

The phrases belonging to the linguistic material were divided into categories: (1) Short sentences ("s"), (2) long phrases ("l"), (3) questions ("q") and (4) nonsense phrases. Thus the content of the linguistic material was semantically neutral, i.e., not within the realm of any emotion and did not influence the emotions expressed by the actors. For example, the word "pain" is related to sadness or fear. In the VERBO database, the actor/actress is able to express her emotions in the sentences in a way that does not allow the meaning of the text to influence the acting. In addition, all the phonemes of the Portuguese language including vowels and consonants were covered.

Table 1 shows the linguistic material used in the recordings of the actors and actresses.

Stage 5: Recording of the Audios by the Actors

The VERBO database was created at the Institute of Mathematical and Computer Sciences at the University of Sao Paulo using portable digital recorders. The creation of the VERBO database was based on the Italian corpus (EMOVO) Costantini *et al*. (2014). However, the EMOVO base contains audios of only 6 professionals, while our database contains audios of 12 professionals, i.e., twice as many professionals and audios.

The actors and actresses were instructed to record the audios through basic pre-defined emotions, based on experiences already undergone by the professionals themselves. In their acting, each of the professionals expressed all 7 emotions for each sentence of the linguistic material, resulting in a total of 1167 recordings, when the recordings of all the actors and actresses are added together.

Stage 6: Audio Storage

The recorded audios by the professionals were stored in 12 folders, separated between male (m) and female (f). Fig. 3 represents the form which the audio recording file was stored. The first part of the name consists of the emotional states, which are happiness, disgust, fear, neutral, anger, surprise and sadness. Second, corresponding to the ID of the professional (f1, f2, f3, f4, f5, f6, m1, m2, m3, m4, m5, m6) who recorded them. The third part indicates the phrase that was pronounced represented by its ID, as showed in Table 1. Finally, the last part is the file extension, i.e., the audios were written in the ".wav" format. For example, in the Fig. 3, the file *ale-f1-s1.wav* matches the phrase "The workers rise early" of the actress *f1* expressing the emotion "happiness".

Table 1: Linguistic material used in the recordings

ID	Category	Phrase in Portugues	Phrase in English
l1	Long	Os bombeiros estão equipados com uma arma.	Firefighters are equipped with a gun.
l2	Long	No próximo outono, Antônio vai a Minas em quinze de outubro. on	The next fall Antônio goes to Minas October fifteenth.
l3	Long	Agora vou pôr a camiseta e sair para uma caminhada.	Now I'm going to put on my T-shirt and go for a walk.
l4	Long	Um momento depois, ele caminhou ... e tropec¸ou.	A moment later he hath walked ... and stumbled.
l5	Long	Eu queria o núumero de telefone de seu João.	I would like the telephone number of Mr. João.
ns1	Nonsense	A casa forte quer com pão.	The strong house wants with bread.
ns2	Nonsense	A Força está para cima e alho vermelho.	The force is up and red garlic.
ns3	Nonsense	O gato está rolando na pêra.	The cat is rolling in pear.
ns4	Nonsense	Salada de massa pata de carneiro amendoim.	Pasta salad leg of lamb peanut.
ns5	Nonsense	Um quarenta e três vinte e sete noventa cinco mil.	One forty-three twenty-seven ninetye five thousand.
q1	Question	Sábadoà noite, o que vai fazer?	Saturday night, what are you going to do?
q2	Question	Você vai trazer aquela coisa com você?	Will you bring that thing with you?
s1	Short	Os operários levantam cedo.	Workers get up early.
s2	Short	A cachoeira faz muito barulho.	The waterfall makes a lot of noise.

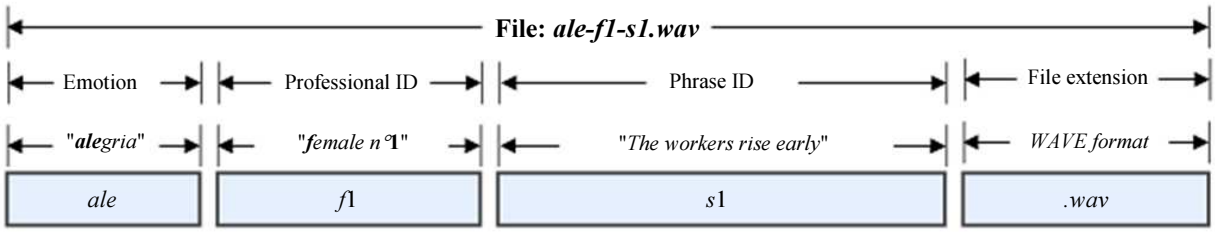

Fig. 3: Diagram representing the phrase "Os operários levantam cedo" (The workers rise early) of the actress f1 expressing the emotion "alegria" (happiness)

Analysis and Validation of the VERBO Database

In the evaluation of the VERBO database, a statistical analysis was conducted to measure how far the database audios are representative for speech emotion recognition systems. The VERBO database consists of 14 sentences of 12 professionals, 6 actors and 6 actresses, for each of the 6 basic emotions with the addition of the neutral state. This meant that 1176 audio recordings were analyzed in the evaluation.

A. Validation by Expert Judges

We conducted an audio analysis in recordings aimed at validating the reliability of the VERBO database. This type of analysis is called Content Validity and involves assessing the degree to which each characteristic of a measuring instrument is important and representative of a construct designed for the specific purpose of evaluation Haynes *et al*. (1995); Alexandre and Coluci (2011); Sjoberg *et al*. (2018). All the audios were validated by a panel of experts (judges) to ensure the validity of the evidence on the audio content. The criteria for the selection of judges were that they must be: (1) Health professionals; (2) qualified professionals in an area related to emotion; and (3) professionals with clinical experience Alexandre and Coluci (2011). Three psychologists, specialized and experienced in the clinical area, were selected to check the validity of the audio content Lynn (1986); Polit *et al*. (2007); Sjoberg *et al*. (2018). The selected professionals were considered to be capable of validating the emotional audios recorded by the actors and actresses.

The analytical procedure followed by the judges involved listening to the recorded audios and then classifying them in the following categories: Happiness, disgust, fear, anger, surprise, sadness or a neutral emotional state. Each judge evaluated and rated the 1167 audios recordings. It is worth noting that in the classification, the judges did not know what emotion the audios represented. In other words, the judges only had an idea of the emotions (categories) that the audio could represent, which were the basic emotions. Thus, the audios did not induce the judges to classify a pre-established emotion and eliminated the risk of a classification bias. After the classification of each judge, the data were analyzed on the basis of the other database assessments, such as those described in Busso *et al*. (2008); Costantini *et al*. (2014); Meddeb *et al*. (2017).

Table 2 provides the audition results given by the judges. The judges' evaluation showed, that of the 1167 audios in the database, 283 were wrongly classified by all the judges, i.e., there was no agreement between any of the judges with regard to the pre-established emotion of the audio recordings. It is worth noting that one of the judges stated that it was difficult to classify speech only by the sound of the voice, without taking into account its context, since in every day life they are interrelated. However, the audios have to be short and out of context to capture the pattern of emotion expressed by the actor in short frames Meddeb *et al*. (2017). It was also observed that when an actor uses long texts, his tone of voice tends to fluctuate and he ends up expressing more than one emotion in the same audio. Thus, through the recognition of speech patterns for each emotion, it is possible to construct a computational model that is capable of classifying whether they are short or long by dividing the audio into fragments.

It can also be seen in Table 2 that of the 884 remaining audios considered representative of the emotion expressed, 202 obtained the agreement of at least one judge, 242 audios obtained the agreement of two judges and 440 audios obtained the agreement of all the judges. It should be stressed that in only 24.25% of the audios did the judges fail to reach an agreement, while 75.75% of the audios were agreed by at least one judge.

Figure 4 shows our analysis of the audios that were correctly labeled, i.e. the audios recordings were recognized by at least one judge. It should be noted that the judges correctly labeled the audiences as follows: 81% happiness, 69% disgust, 70% fear, 82% in the neutral state, 81% anger, 68% surprise and 72% sadness.

B. Statistical Validation using CVI and Fleiss' Kappa Test

A quantitative evaluation was carried out with the aim of validating the recorded audios in the creation of the VERBO database. In other words, the purpose of this validation was to know how many audios actually represented the emotions in which they were recorded. This evaluation allowed us to determine if the emotion was acted and recorded correctly by the actors. For this reason, it was assumed that the judges could only determine the emotion by listening to the audios and labeling them, without quantifying to what extent this audio was important or not from its use of validation instruments.

We used the Content Validity Index (CVI) for this evaluation, because it is based on expert assessment and is widely used and accepted by health researchers Polit *et al*. (2007); Sjoberg *et al*. (2018). The approach adopted in this evaluation was a scale of three specialists and 1167 audios (items). For each item (I-CVI), the CVI of the judges' response was calculated and then the average of all the items (I-CVI/Ave) was calculated Polit *et al*. (2007); Sjoberg *et al*. (2018). The average of the whole I-CVI produced a value of 0.76 of agreement when all the items were categorized.

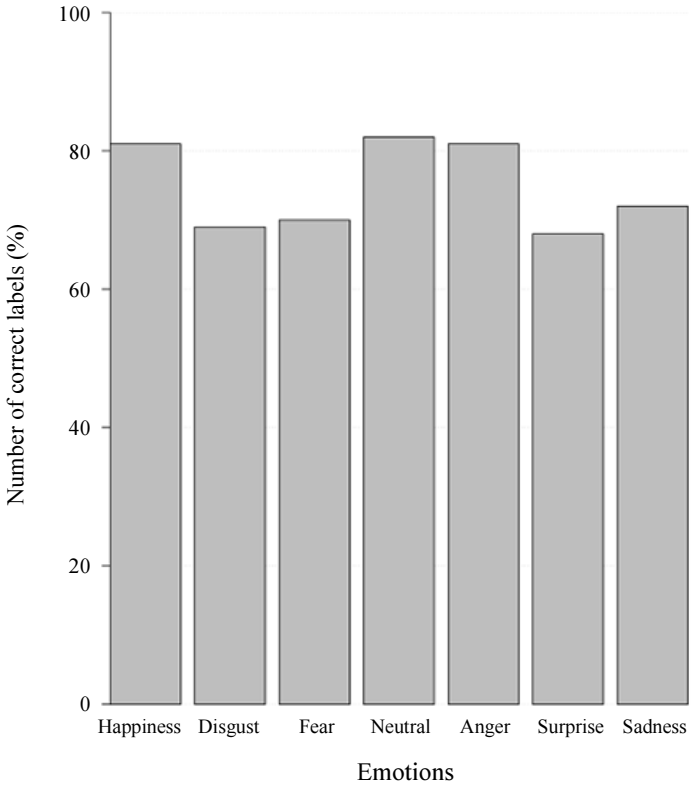

Fig. 4: Relation of audios labeled by the judges per emotion

Table 2: Measurements for agreement between the judges and the database audios (N = 1167)

Emotion	None of the judges	At least one judge	Two judges	All the judges
Happiness	16	29	28	**93**
Disgust	59	29	42	37
Fear	52	**40**	22	52
Neutral	8	19	**47**	**93**
Anger	30	27	24	86
Surprise	**64**	31	37	35
Sadness	54	27	42	44
Total	283	202	242	440

Table 3: Results of the content validity test

Parameter	Value
Number of judges	**3.00**
Categories	**7.00**
I-CVI/Ave	**0.76**
Fleiss's Kappa	**0.65**

Table 4: Evaluation the performance via accuracy, precision, recall and F1-score

	Accuracy	Precision	Recall	F1-Score
Happiness	0.76	0.81	0.85	0.83
Disgust	0.76	0.69	0.65	0.67
Fear	0.76	0.70	0.81	0.75
Neutral	0.76	0.82	0.70	0.76
Anger	0.76	0.81	0.89	0.85
Surprise	0.76	0.68	0.68	0.68
Sadness	0.76	0.72	0.71	0.71

In addition, we also used the Fleiss' Kappa Test for the emotion labeling analysis, since the choice of emotions for each item made by the specialists (three judges) represents a classification between categorical variables Fleiss *et al*. (1969). The value obtained by our analysis was 0.65, $p<0,05$, which indicates there was considerable agreement between the judges about the classified items. Table 3 summarizes the results of I-CVI and Fleiss's Kappa.

After analyzing the answers of all specialists, we performed a performance measure evaluation aim to present the relevance of all emotion performed by actors/actresses in the audio recordings. Thus, we calculated the accuracy, precision, recall and f1-score for each emotion based on the validation achieved by specialists with I-CVI and Fleiss's Kappa. In this

evaluation, we arrived at the accuracy of 0.76 of recordings correctly labeled. The accuracy was calculated from the ratio of these recordings correctly labeled to the total observations. Despite the accuracy be a great measure, only the accuracy is not to possible to ensure the recordings relevance because the false positive and false negative are almost all distinctive. Therefore, we also calculate the precision, recall and f1-score for each emotion. The results of this evaluation can be seen in the Table 4.

In the results, it is possible to see that the anger and happiness emotions were the higher representatives in relation to other emotions presenting 0.85 and 0.83 of f1-score, respectively. Whereas the disgust and surprise emotions were the least recognized by the experts presenting 0.67 and 0.68, respectively. In addition, Table 4 also presents a variation in the precision measurement of 0.75 ± 0.07 and recall of 0.77 ± 0.12.

Applicability

Figure 5 shows how the VERBO database can be applied to smart environments with different applications. It should be noted that the VERBO database is designed to be used in systems of emotional recognition through speech. For instance, any help-desk and call center system for Portuguese speakers can benefit from VERBO database. This is because such systems may collect and analyze voice data of customers' and employees' emotions and how they react to certain situations Holman *et al*. (2002); Deery *et al*. (2002). The emotion recognition by voice can also occur using face-to-face meetings or via Skype and Hangout, to monitor the employees Jones and Graham (2015). Also, users can be persuaded by interactive systems according to their emotion (e.g., the system may recommend a movie, a game, a music, etc.) Mano *et al*. (2016b). The VERBO database can be adopted in systems to detect uni-polar and bipolar depressive disorders and aid psychologists or doctors in the clinic Huang *et al*. (2018). Additionally, the treatment of other diseases, such as mental disorders Kostoulas *et al*. (2012), stress Yogesh *et al*. (2017) and Parkinson Zhao *et al*. (2014) can also benefit from the VERBO database. Finally, it is important to emphasize that Portuguese language is the 6th most spoken language world wide Lane (2016).

In addition, there are numerous applications in which the VERBO database can be used. In education, the VERBO can aid to improve a student's experience in a learning environment as well as in virtual teacher training programs. This can take place by measuring the arousal level of the emotions while they are interacting Shernoff *et al*. (2014); Harrison (2016). In addition, the database allows making an advanced design of systems to convert texts to voice (e.g., transcribe foreign speech) or systems of language education in LIBRAS Frota *et al*. (2015); Durlak (2015); Weiss *et al*. (2017). In Ambient Assisted Living, our database can help in assisting people with disabilities to communicate and to provide interactivity with computer systems or smartphones (e.g., to recognize Gonçalves *et al*. (2016); Mano *et al*. (2016a; 2016c) and conduct decision-making process by speech Filho *et al*. (2018); Khunt and Prabu (2018)). Also, the VERBO may provide the support needed for serious game applications in the treatment of children with autism, and in engaging of teenagers with Down Syndrome, improving their communication and speech skills Bernardini *et al*. (2014); Gonzàlez-Ferreras *et al*. (2017).

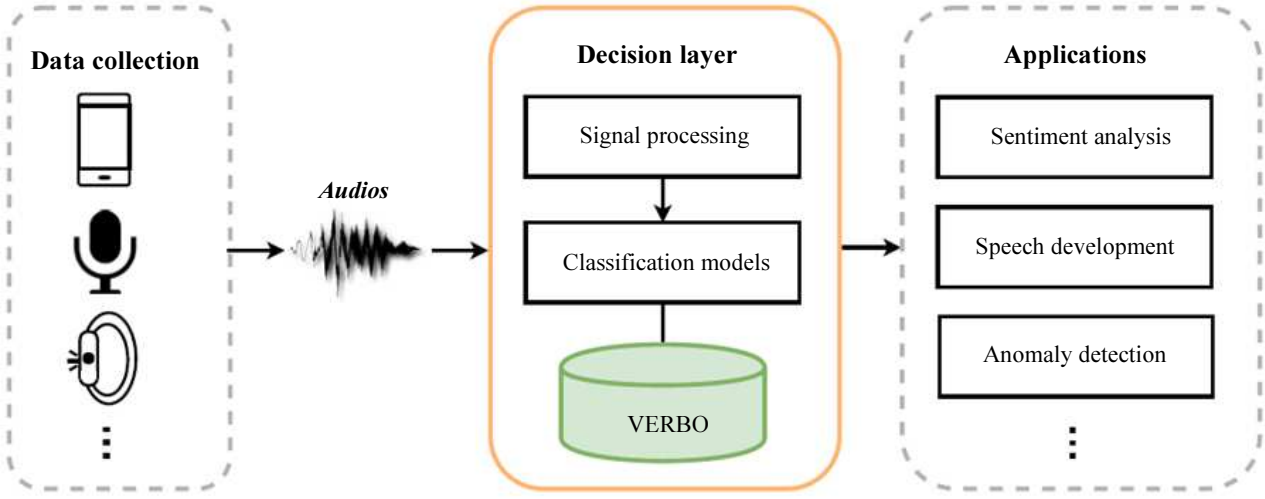

Fig. 5: VERBO applicability model for different applications

Conclusion and Future Work

The aim of this study is to contribute to the advances being made in research on the recognition of emotions through human discourse. It also seeks to answer the following question: How can we create a database of emotional discourses in Brazilian Portuguese language? Thus, an emotional database was designed with speeches in Brazilian Portuguese, called Voice Emotion Recognition dataBase in pOrtuguese language (VERBO), that is capable of representing the emotions of individuals. We ensured that the type of database was acceptable to the scientific community and included the emotions that are found in the literature, as well as having suitable participants for the collection of audios and the linguistic material.

In the database evaluation, we used the Content Validity Index (CVI) to validate the reliability of the audio recordings. A panel of specialist psychologists (judges) was used to ensure the validity of the evidence supplied by the audio content. The evaluation of this panel ensured an agreement of 76% using the Item-CVI and a considerable agreement of 65% by means of the Fleiss' Kappa Test. In the performance measure evaluation, it was observed that the emotions anger and happiness were more easy to recognize showing 0.85 and 0.83 of f1-score, respectively, whereas the disgust and surprise emotions were the most difficult showing 0.67 and 0.68, respectively.

In future studies, we seek to explore a speech emotion recognition system with the aid of the VERBO database by applying it to real-world scenarios. Thus, it will be possible to explore which features best represent the emotions through the speech of individuals, e.g., Mel-frequency cepstral coefficients or prosody.

Acknowledgement

The authors would like to thank the supporting organizations for funding this work: The Agency Coordenac¸ão de Aperfeiçoamento de Pessoal de Nível Superior - CAPES and the grants #2016/25865-2, #2016/14267-7, #2017/21054-2 and #2017/23655-3 from Foundation for Research Support of the State of Sao Paulo - FAPESP. In addition, the authors thank the University of São Paulo and the University of Ottawa.

Author's Contributions

José R. Torres Neto: The author designed the research proposal, organized the study, designed of the proposed work plan, collected the data, recorded the audios, analyzed the data, reviewed the various published articles in the field, contributed to the hypothesis, writing of the manuscript and final approval.

Geraldo P. Rocha Filho: The author contributed to the data analysis, editing of the manuscript, article review and final approval.

Leandro Y. Mano: The author contributed to text, hypothesis construction, editing of the manuscript, article review and final approval.

Jó Ueyama: The author contributed as the research guide, technical corrections, reviewing it critically and final approval.

Ethics

This work is original and not published elsewhere. The authors confirm that they have read and approved the manuscript and there is no conflict of interest. Also, the authors confirm that there are no ethical issues involved.

References

Alexandre, N.M.C. and M.Z.O. Coluci, 2011. Validade de conteúdo nos processos de construc¸ão e adaptaç ão de instrumentos de medidas. Ciência Saúde Coletiva, 16: 3061-3068. DOI: 10.1590/S1413-81232011000800006

Apesoa-Varano, E.C., J.C. Barker and L. Hinton, 2015. Shards of sorrow: Older men's accounts of their depression experience. Soc. Sci. Med., 124: 1-8. DOI: 10.1016/j.socscimed.2014.10.054

Bernardini, S., K. Porayska-Pomsta and T.J. Smith, 2014. Echoes: An intelligent serious game for fostering social communication in children with autism. Inform. Sci., 264: 41-60. DOI: 10.1016/j.ins.2013.10.027

Busso, C., M. Bulut, C.C. Lee, A. Kazemzadeh and E. Mower *et al*., 2008. IEMOCAP: Interactive Emotional dyadic Motion Capture database. Lang. Resources Evaluat., 42: 335-335. DOI: 10.1007/s10579-008-9076-6

Busso, C., S. Parthasarathy, A. Burmania, M. AbdelWahab and N. Sadoughi *et al*., 2017. Mspimprov: An acted corpus of dyadic interactions to study emotion perception. IEEE Trans. Affective Comput., 8: 67-80.

Chen, L., X. Mao, Y. Xue and L.L. Cheng, 2012. Speech emotion recognition: Features and classification models. Digital Signal Process., 22: 1154-1160. DOI: 10.1016/j.dsp.2012.05.007

Costantini, G., I. Iaderola, A. Paoloni and M. Todisco, 2014. EMOVO corpus: An Italian emotional speech database. Proceedings of the 9th International Conference on Language Resources and Evaluation, May 26-31, European Language Resources Association (ELRA), Reykjavik, Iceland, pp: 3501-3504.

Deery, S., R. Iverson and J. Walsh, 2002. Work relationships in telephone call centres: Understanding emotional exhaustion and employee withdrawal. J. Manage. Stud., 39: 471-496. DOI: 10.1111/1467-6486.00300

Durlak, J.A., 2015. Handbook of Social and Emotional Learning: Research and Practice. Guilford Publications.

Ekman, P., 1992. Are there basic emotions? Psychol. Rev., 99: 550-553. DOI: 10.1037/0033-295X.99.3.550

Filho, G.P.R., L.A. Villas, H. Freitas, A. Valejo and D.L. Guidoni *et al*., 2018. Residi: Towards a smarter smart home system for decision-making using wireless sensors and actuators. Comput. Netw., 135: 54-69. DOI: 10.1016/j.comnet.2018.02.009

Fleiss, J.L., J. Cohen and B.S. Everitt, 1969. Large sample standard errors of kappa and weighted kappa. Psychol. Bull., 72: 323-327. DOI: 10.1037/h0028106

Frota, S., P. Oliveira, M. Cruz and M. Vigário, 2015. P-tobi: Tools for the transcription of Portuguese prosody.

Gonçalves, V.P., E.P. Costa, A. Valejo, G.P. Rocha Filho and T. Johnson *et al*., 2016. Enhancing intelligence in multimodal emotion assessments. Applied Intell., 46: 470-486. DOI: 10.1007/s10489-016-0842-7

Gonzàlez-Ferreras, C., D. Escudero-Mancebo, M. Corrales-Astorgano, L. Aguilar-Cuevas and V. Flores-Lucas, 2017. Engaging adolescents with down syndrome in an educational video game. Int. J. Human–Comput. Interact., 33: 693-712.

Harrison, G.A., 2016. An investigation of the influence of emotional intelligence on Successful doctoral students in an online program at one University. Lamar University-Beaumont.

Haynes, S.N., D.C.S. Richard and E.S. Kubany, 1995. Content validity in psychological assessment: A functional approach to concepts and methods. Psychol. Assess., 7: 238-247. DOI: 10.1037/1040-3590.7.3.238

Holman, D., C. Chissick and P. Totterdell, 2002. The effects of performance monitoring on emotional labor and well-being in call centers. Motivat. Emot., 26: 57-81. DOI: 10.1023/A:1015194108376

Huang, K., C. Wu, M. Su and Y. Kuo, 2018. Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model. IEEE Trans. Affective Comput. DOI: 10.1109/TAFFC.2018.2803178

Jing, S., X. Mao and L. Chen, 2018. Prominence features: Effective emotional features for speech emotion recognition. Digital Signal Process., 72: 216-231. DOI: 10.1016/j.dsp.2017.10.016

Jones, N.B. and C.M. Graham, 2015. Virtual teams in business and distance education: Reflections from an MBA class. J. Bus. Economic Policy, 2: 49-59.

Jurafsky, D. and J.H. Martin, 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Khunt, A.R. and P. Prabu, 2018. An empirical analysis of android permission system based on user activities. J. Comput. Sci., 14: 324-333. 10.3844/jcssp.2018.324.333

Kostoulas, T., I. Mporas, O. Kocsis, T. Ganchev and N. Katsaounos *et al*., 2012. Affective speech interface in serious games for supporting therapy of mental disorders. Expert Syst. Applic., 39: 11072-11079. DOI: 10.1016/j.eswa.2012.03.067

Lane, J., 2016. The 10 most spoken languages in the world. https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world."

Lichtenstein, A., A. Oehme, S. Kupschick and T. Jürgensohn, 2008. Comparing Two Emotion Models for Deriving Affective States from Physiological Data. In: Affect and Emotion in Human-Computer Interaction, Peter, C. and R. Beale (Eds.), Springer, Berlin, Heidelberg, ISBN13: 978-3-540-85099-1, pp: 35-50.

Lynn, M., 1986. Determination and quantification of content validity. Nurs. Res., 35: 382-386. DOI: 10.1097/00006199-198611000-00017

Mahlke, S. and M. Minge, 2008. Consideration of Multiple Components of Emotions in Human-Technology Interaction. In: Affect and Emotion in Human-Computer Interaction, Peter, C. and R. Beale (Eds.), Springer, ISBN-13: 978-3-540-85099-1, pp: 51-62.

Mano, L., M. Funes, T. Volpato and J. Neto, 2016a. Explorando tecnologias de iot no contexto de health smart home: Uma abordagem para detecc¸ ão de quedas em pessoas idosas. J. Adv. Theoretical Applied Inform., 2: 46-57. DOI: 10.26729/jadi.v2i1.1667

Mano, L.Y., B.S. Faiçal, L.H. Nakamura, P.H. Gomes and G.L. Libralon *et al*., 2016b. Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition. Comput. Commun., 89-90: 178-190. DOI: 10.1016/j.comcom.2016.03.010

Mano, L.Y., L.O. Sawada and J. Ueyama, 2016c. Abordagem para a interaçao afetiva: Um estudo de caso com player de musica. J. Adv. Theoretical Applied Inform., 2: 47-54. DOI: 10.26729/jadi.v2i2.2108

Mano, L.Y., E. Vasconcelos and J. Ueyama, 2016d. Identifying emotions in speech patterns: Adopted approach and obtained results. IEEE Latin Am. Trans., 14: 4775-4780. DOI: 10.1109/TLA.2016.7817010

Meddeb, M., H. Karray and A.M. Alimi, 2017. Building and analysing emotion corpus of the Arabic speech. Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition, Apr. 3-5, IEEE Xplore Press, Nancy, France, pp: 134-139. DOI: 10.1109/ASAR.2017.8067775

Picard, R.W., 2010. Affective computing: From laughter to IEEE. IEEE Trans. Affective Comput., 1: 11-17. DOI: 10.1109/T-AFFC.2010.10

Polit, D.F., C.T. Beck and S.V. Owen, 2007. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res. Nurs. Health, 30: 459-467. DOI: 10.1002/nur.20199

Rázuri, J.G., D. Sundgren, R. Rahmani, A. Larsson and A.M. Cardenas *et al*., 2015. Speech emotion recognition in emotional feedback for human-robot interaction. The Science and Information Organization.

Russell, J., 1980. A circumplex model of affect. J. Personality Soc. Psychol., 39: 1161-1178. DOI: 10.1037/h0077714

Scherer, K.R., 2005. What are emotions? and how can they be measured? Soc. Sci. Inform., 44: 695-729. DOI: 10.1177/0539018405058216

Shadiev, R., B.L. Reynolds, Y.M. Huang, N. Shadiev and W. Wang *et al*., 2017. Applying speech-to-text recognition and computer-aided translation for supporting multi-lingual communications in cross-cultural learning project. Proceedings of the IEEE 17th International Conference on Advanced Learning Technologies, Jul. 3-7, IEEE Xplore Press, Timisoara, Romania, pp: 182-183. DOI: 10.1109/ICALT.2017.20

Shernoff, D.J., M. Csikszentmihalyi, B. Schneider and E.S. Shernoff, 2014. Student engagement in high school classrooms from the perspective of flow theory. In: Applications of Flow in Human Development and Education, Springer, pp: 475-494.

Sjoberg, H., U. Aasa, M. Rosengren and L. Berglund, 2018. Content validity index and reliability of a new protocol for evaluation of lifting technique in the powerlifting squat and deadlift. J. Strength Condit. Res.

Swain, M., A. Routray and P. Kabisatpathy, 2018. Databases, features and classifiers for speech emotion recognition: A review. Int. J. Speech Technol., 21: 93-120. DOI: 10.1007/s10772-018-9491-z

Ververidis, D. and C. Kotropoulos, 2003. A review of emotional speech databases.

Weiss, R.J., J. Chorowski, N. Jaitly, Y. Wu and Z. Chen, 2017. Sequence-to-sequence models can directly transcribe foreign speech. CoRR, abs/1703.08581.

Yogesh, C.K., M. Hariharan, R. Ngadiran, A.H. Adom and S. Yaacob *et al*., 2017. A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. Expert Syst. Applic., 69: 149-158. DOI: 10.1016/j.eswa.2016.10.035

Zhao, S., F. Rudzicz, L.G. Carvalho, C. Marquez-Chin and S. Livingstone, 2014. Automatic detection of expressed emotion in parkinson's disease. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 4-9, IEEE Xplore Press, Florence, Italy, pp: 4813-4817. DOI: 10.1109/ICASSP.2014.6854516