Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 223-236 recibido 06-04-2025 revisado 13-05-2025 aceptado 19-05-2025

CHERISH: A Corpus for CHildren Emotion Recognition In SpeecH

CHERISH: Un Corpus para el Reconocimiento de Emociones Infantiles en el Habla

Luz Gahona Castillejos, Delia Irazú Hernández Farías, Humberto Pérez-Espinosa
Coordinación de Ciencias Computacionales
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
{luz.gahona,dirazuhf,humbertop}@inaoep.mx

Abstract: Emotion recognition in children remains significantly less explored than in adults, largely due to the limited availability of annotated data. To address this gap, we introduce CHERISH, a multimodal corpus for emotion recognition in Spanish-speaking children. Our data collection methodology includes various sources such as speech, speech transcriptions, behavioral descriptions provided by a human observer, and personality traits obtained through the Children's Personality Questionnaire (CPQ). Each of these modalities contributes key information to enhance children emotion recognition: speech provides insights into vocal expression, transcriptions reflect the semantic content of speech, behavioral descriptions offer context on body language, and personality since it influences how children express and regulate their emotions. We aim to contribute research on emotion recognition in children, which could enhance the development of more robust models that can be applied in education, healthcare, and child-assistive technologies. Additionally, we present baseline results for the emotion recognition task under two experimental conditions, establishing a foundation for future comparative studies.
Keywords: Multimodal Emotion Recognition, Emotion Analysis in Children, Spoken Spanish Corpora.

Resumen: El reconocimiento de emociones en niños es un área de estudio poco explorada en comparación con su contraparte en adultos, en parte debido a la escasez de datos disponibles. En este trabajo presentamos CHERISH, un corpus multimodal para el reconocimiento de emociones en niños de habla hispana. Nuestra metodología de obtención de datos incluye diversas fuentes como el habla, transcripciones del habla, descripción del comportamiento realizada por un observador humano y rasgos de personalidad obtenidos mediante el Cuestionario de Personalidad para Niños (CPQ por sus siglas en Inglés). Cada una de estas modalidades aporta información clave para mejorar la precisión del reconocimiento emocional: el habla proporciona información de como se expresan los mensajes vocales, las transcripciones reflejan el contenido semántico del discurso, la descripción del comportamiento ofrece contexto sobre el lenguaje corporal y la personalidad que influye en la forma en que los niños expresan y regulan sus emociones. Este corpus busca contribuir en la investigación sobre el reconocimiento de emociones en niños, lo cual podria permitir el desarrollo de modelos más robustos que puedan aplicarse en educación, salud y tecnologías de asistencia infantil. Finalmente, tambien proporcionamos una linea base de la tarea de reconocimiento de emociones en dos configuraciones.
Palabras clave: Reconocimiento Multimodal de Emociones, Análisis de Emociones en Niños, Corpus de Habla en Español.

ISSN 1135-5948 DOI 10.26342/2025-75-17

1 Introduction

Understanding and recognizing **emotional states** plays a crucial role in human interaction, influencing both communication and decision-making. Emotions are complex phenomena that involve both physiological and cognitive components, as well as behavioral aspects. According to Ortiz & Parra (Martello Ortiz and Arévalo Parra, 2017), *emotional states* are subjective responses to an experience. In the literature, there are two main models of emotions: *I. Discrete emotions theory* classifies emotions into specific categories, each with cognitive, physiological, and behavioral characteristics.

One of the most widely accepted models in emotion research is the one proposed by Ekman, comprising six basic emotions: *happiness*, *surprise*, *fear*, *sadness*, *anger*, and *disgust* (Ekman, 1992). These emotions are considered universal as they are recognized in different cultures through facial expressions and physiological responses. And *II. Multidimensional emotions theory*, which acknowledges the complexity of emotions and their dependence on cultural context, personal history, and individual differences. This model allows for a more detailed analysis through emotional spaces in two or three dimensions (Kalateh et al., 2024).

Emotion recognition is the process of interpreting human emotional states through different signs, such as speech, facial expressions, body language, and physiological signals. While humans perform this task intuitively by integrating multiple sources of information, replicating this ability computationally remains a complex challenge. Automated emotion recognition uses technologies like *machine learning*, *signal processing technologies*, and *natural language processing* to interpret emotions in contexts such as human-computer interaction, healthcare, and entertainment, among others. Depending on the number of modalities analyzed, these systems can be *unimodal*, *bimodal*, or *multimodal*. *Unimodal systems* focus on a single source of information, such as voice or text. *Bimodal systems* combine two modalities, such as facial expression and vocal intonation, while *Multimodal systems* integrate multiple sources, offering greater performance but also greater complexity in their implementation (Kalateh et al., 2024).

Although extensive research on emotion recognition has been conducted in adults, the study of this phenomenon in children remains limited. Advancing our understanding of emotion recognition in children can enable the development of systems tailored to their specific needs, thereby enhancing educational and caregiving services and yielding significant long-term benefits (Nojavanasghari et al., 2016).

Emotion recognition in children, especially in Spanish, is challenging, in particular due to data scarcity. In this paper, we propose CHERISH, a multimodal corpus for recognizing emotions in children aged *9* to *12* years. It comprises information from *audio*, *speech transcriptions*, *participant behavior descriptions* (recorded by a human observer), *personality traits* (obtained through the *CPQ* test), as well as their interest in technology topics obtained from a Science, Technology, Engineering, and Mathematics (STEM) questionnaire.

Beyond being an acronym, CHERISH[1] conveys the idea of valuing and studying children's emotions.

Each of the modalities provides complementary information: **Voice** allows the analysis of *how* children express themselves through tone, intensity, and speech rate, among many other acoustic properties of speech. **Transcriptions** provide the semantic content, enabling the interpretation of *what* is being said. **Behavior description** offers context regarding the situation in which the child is involved, helping to identify gestures, postures, and expressions that may reflect their emotional state. **Personality Traits** assessments offer important contextual information, as personality influences how individuals express and regulate their emotions, shaping their emotional responses to various stimuli. The evaluation of **interest in STEM** subjects provides insights into the child's level of engagement with scientific and technical content, which may impact their emotional reactions in educational and problem-solving contexts.

The main contributions of the proposed work can be summarized as follows:

- *Multimodal dataset collection and annotation*: A multimodal dataset in Spanish

[1] In English, this word is associated with words like *adore* or *value*.

was collected and annotated, focusing on children's spoken emotions in a natural environment. CHERISH integrates multiple sources of information, including voice, speech transcription, behavioral descriptions, and personality traits.

- *Naturalistic data collected in an educational setting*: The data were gathered during spontaneous interactions within an educational robotics context, enabling the capture of authentic emotional responses in a real-world environment.
- *Detailed annotations with emotional labels and individual characteristics*: The corpus includes emotion labels in categorical and continuous dimensions, along with personality assessments that help to understand how personality influences emotional expression and affinity for STEM activities.
- *Focus on Spanish-speaking children*: CHERISH could contribute to the development of more accurate models for emotion recognition in Spanish-speaking children.
- *Development of a web-based annotation interface*: The tool enables both categorical and multidimensional emotion labeling, and provides functionalities to manually correct automatic speech transcriptions and segmentation errors. Additionally, it integrates observer notes recorded during direct behavioral observation.

This paper is structured as follows. Section 2 reviews previous studies on emotional speech corpora in children. In Section 3, we describe the phases involved in corpus development. The initial experiments conducted with CHERISH are presented in Section 4. Finally, Section 5 discusses the results and future research directions.

2 Related work

Emotions can be collected through three main approaches: *i)* Using stimuli such as videos or music to induce emotions, *ii)* Evoking significant emotional memories and *iii)* Constructing interaction scenarios in which subjects can talk about any topic (Pan et al., 2023). Most of the corpora available for emotion recognition in children are in English, although there are also resources in other languages such as German, French, Mandarin, Filipino, and Russian. Additionally, they differ not only in source language but also in other aspects like age range, emotions considered, and construction criteria (Matveev et al., 2022). Table 1 shows a comparison between the available datasets for children's emotion recognition and the one proposed in this work.

In the following, we will introduce the available Spanish-language corpora for emotion recognition in children. *Emo-Wisconsin* database (Pérez-Espinosa, Reyes-García, and Villaseñor-Pineda, 2011) contains recordings of speaking-spanish Mexican children between the ages of 7 and 13 interacting with an adult while playing an adaptation of the Wisconsin Card Sorting Test (a neuropsychological test (Monchi et al., 2001)), which is the most commonly used task to assess cognitive flexibility in humans.

Recordings were segmented at speaker turn level and annotated with six emotional categories: *Doubtful*, *Annoyed*, *Motivated*, *Nervous*, *Neutral*, and *Confident*, as well as three continuous emotions: *Valence*, *Activation*, and *Dominance*. On the other hand, the *IESC-Child* database (Pérez-Espinosa et al., 2020) was collected from interactions between children and robots, using a Wizard of Oz setup to induce various emotional reactions. The recordings were manually segmented and labeled with two types of emotional information: emotions (*anger*, *fear*, *happiness*, *surprise*, *fear*, *disgust*, and *neutral*) and attitudes (*confident*, *uncertain*, *apathetic*, and *enthusiastic*).

Mexican Emotional Speech Database (MESD) (Duville, Alonso-Valerdi, and Ibarra-Zarate, 2021) contains single-word speech utterances expressed by adults and children with different emotions: *anger*, *disgust*, *fear*, *happiness*, *neutrality*, and *sadness*. A portion of the MESD with children's speech was produced by six non-professional actors. .

Among the available corpora in the literature, to the best of our knowledge, only one of them made use of a robot during the development of the resource. *FAU-AIBO (Friedrich-Alexander-Universität corpus)* (Steidl, 2009) contains spontaneous speech from 51 German-speaking children aged between 10 and 13 while interacting

Corpus	Subj.	Lang.	Age	Seg.	Emo. Labels	Mod.	Type
MESD	8	ES	8–11	288	6C	A	ACT
IESC-Child	174	ES	6–11	19,793	7C, 4AC	A	I
EmoWisconsin	28	ES	7–13	3,098	6C, 3D	A	S
FAU Aibo	51	DE	10–13	13,642	5C	A	I
EmoReact	63	EN	4–14	1,102	17C	A,V	S
EmoChildRu	100	RU	3–7	20340	3C	A	S
CHERISH	35	ES	9–12	1,563	5C, 2D	A,T,SD	S

Table 1: Children's emotional speech corpora available in the literature. **Abbreviations: Subj.**: Subjects, **Lang.**: Language (ES = Spanish, DE = German, EN = English, RU = Russian), **Age**: Age range, **Seg.**: Segments, **Emo. Labels**: Emotion labels (**C** = Categorical, **AC** = Attitudinal, **D** = Dimensional), **Mod.**: Modalities (**A** = Audio, **T** = Transcription, **SD**= Scene Description, **V** = Video), **Type**: Emotion type (**S** = Spontaneous, **I** = Induced, **ACT**= Actuated).

with a Sony pet robot Aibo.

Five annotators labeled each word as *neutral* or one of ten categorical emotions: *angry*, *bored*, *joyful*, *surprised*, *emphatic*, *helpless*, *irritated*, *motherese*, *reprimanding*, and *rest*.

EmoChildRu (Lyakso and others, 2015) First Russian emotional speech corpus for children, with 20k+ recordings from 100 children. Data were collected in 3 scenarios and annotated with 3 emotion labels (*comfort, neutral, discomfort*). Includes EEG, questionnaires, and perceptual/adult evaluations. *EmoReact* (Nojavanasghari et al., 2016): Multimodal dataset with 1102 audiovisual clips of children, labeled with 17 emotional states. Enables research on unimodal vs. multimodal recognition. Also analyzes key behavioral cues for emotion detection in children.

Current corpora for children's emotion recognition present several limitations. Most are based solely on audio recordings without incorporating multimodal information, which very often is crucial to correctly interpret emotional expression. Another drawback is how data is collected; very often, emotions are acted or induced rather than spontaneous, reducing its applicability in natural settings. Furthermore, there is a notable scarcity of Spanish-language corpora, which limits research in Spanish-speaking populations. To address these shortcomings, our multimodal corpus for children's emotion recognition in Spanish integrates diverse sources of information: audio, transcripts, behavioral observations, and personality. Besides, CHERISH also balances induced and spontaneous emotions and employs a standardized annotation method that combines categorical and dimensional labels.

3 Developing CHERISH

CHERISH was created to promote research on emotion recognition in children. Furthermore, it also allows us to analyze the relationship between emotions and the completion of a cognitive activity. CHERISH includes information about the profile of the participants, such as their most probable personality trait (assessed by the CPQ test (Porter and Cattell, 1986)), and data on their affinity for STEM fields were also collected through an in-house 15-question survey.

Emotions play a crucial role in academic performance as they influence working memory, motivation, cognitive flexibility, and self-regulation. *Positive activating emotions*, such as the *enjoyment of learning*, have a beneficial effect on performance, especially when they are task-focused rather than directed toward external rewards that may divert attention. In contrast, *negative deactivating emotions*, such as *boredom* and *hopelessness*, tend to be detrimental to academic performance (Pekrun, 2024). Children between the ages of 8 and 13 develop internal strategies for emotional self-regulation, such as redirecting attention to positive thoughts or using behavioral distractions to manage their emotions. At this stage, they acquire more advanced skills to regulate emotional states autonomously (Thompson, 1994). Therefore, this age range is crucial for studying emotions, personality, and academic development.

Data collection was carried out in an

elementary school setting; it is important to note that such an academic place is located in the suburbs of Puebla city, where most of the population lives under limited socioeconomic conditions. Participants were asked to perform a robotic programming task and were encouraged to verbally express their thoughts to describe ongoing events, facilitating the natural and spontaneous recording of emotional responses. The development of CHERISH involved two distinct phases. The first stage consisted of collecting data, while the second one focused on segmenting and labeling the audios. Figure 1 shows a schematic representation of the stages involved.

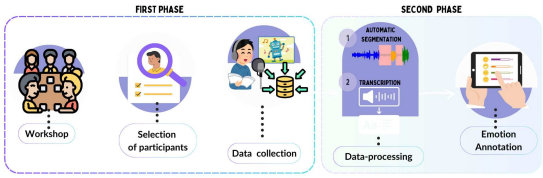

Figure 1: Diagram of CHERISH developing.

For collecting data, with the guide of pedagogical specialist, we developed a bioethics protocol describing the process to ensure that it complied with ethical standards and protected the rights and well-being of all participants. Such a protocol was submitted and approved by the ethics committee of the Popular Autonomous University of the State of Puebla (UPAEP). Afterwards, we approached an elementary school for requesting permission to apply the protocol for the students and to offer the robotics workshop. The school's authorities were informed of the benefits for the local academic population and also for the research on emotion recognition in Spanish. In the following paragraphs, we describe the different phases involved in the development of CHERISH.

Phase 1: Robotics Workshop

Three to four group sessions were held, each lasting approximately 50 minutes, for a total of 12 groups of elementary school students from grades 4^{th} up to 6^{th}. The objective of these sessions was for participants to gain confidence in using the application and to better understand robot programming. During these sessions, students became familiar with the robot programming software (in particular, students used a *LEGO Mindstorms* (LEG, 2025) device) and its components in an attempt to facilitate performing this task during data collection.

Phase 1: Participants' Selection

Students who attended all sessions of the robotics workshop were asked to provide a signed consent from their parents for participating during data collection. Those who did not have such a document were not eligible to continue. Finally, participants completed the CPQ and STEM affinity test. Before starting the data collection phase with each student, we asked again her verbal consent to participate before placing the microphone. During the activity, only the student's voice was recorded using a *BOYALINK wireless lapel* microphone [2] and the *WaveEditor for Android Audio Recorder Editor APK V 1.121* [3] application and a mobile phone. In the most comfortable manner possible.

Phase 1: Data Collection

The activity performed for collecting data required students to individually complete missions in which they had to navigate a robot from point A to point B. These points were predefined on a mission mat, where they had to avoid obstacles or follow a specific path to reach the final destination. The participant had to analyze the best strategy, program the robot, and make it move according to the given instructions.

As an attempt to encourage speech, participants were prompted to think aloud while solving the task. In some cases, their peers used cue cards to ask how they were progressing, further promoting verbalization of their process. This activity was designed to elicit spontaneous emotional expressions in the students. Participants were provided with written instructions along with a small version of the mission board. At the end of the activity or the allotted time, the student was congratulated and allowed to leave.

Direct observation was used to highlight relevant participant actions (e.g., *"she looked nervous"*, *"she kept her gaze on the floor"*, *"she smiled while reading"*) or gestures that could not be identified from the audio alone. This approach allowed registering of non-verbal behaviors in real-time, ensuring a cor-

[2] https://www.boyamic.com/product/boyalink

[3] https://waveeditor-for-android-audio-recorder-editor.softonic.com/android

respondence between audio data and behavioral annotations by synchronizing them with the recording time. The notes made by the observer later will serve as additional support when recording emotions. The observer was positioned near the participant but maintained a respectful distance, as shown in Figure 2, to ensure the participant did not feel uncomfortable.

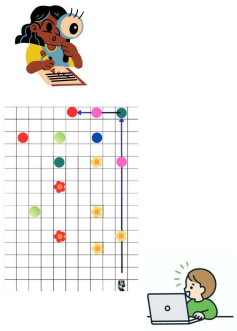

Figure 2: Scene layout during the activity performance. It includes the main elements involved: the child, the robot programming activity, and the human observer.

Phase 2: Data-Processing

At a first phase, recordings in which the participant did not speak or whose speech was incomprehensible were manually removed. Then, automatic segmentation was performed to divide the recordings into fixed-length segments and detect speech and silence periods. Each audio file was divided into 1-minute segments for analysis. Within each segment, periods of 2 seconds of silence were identified[4] and then used to extract speech segments.

Once segmentation was completed, automatic transcription of the audio segments was performed using the **Whisper** model in its **large** version (Radford et al., 2022).
Direct observation notes often contain affective information, which is very useful while annotating emotions, but for automatic processing purposes, the presence of emotional words could bias the processing of textual information. Besides, considering that such descriptions of what is happening during the activity could be automatically generated through video-to-text or image-to-text models, which are very likely to exclude descriptions with detailed affective information. Having these in mind, we decided to take advantage of the capabilities of large-language models for generating more "objective" descriptions, avoiding including emotional terms. Observer notes were processed using **GPT-4o-mini**[5] with the following prompt:

> *"The following text describes an activity in which a child must complete a series of challenges with a robot. Transform this text into a scene description without adding additional information or including words related to emotions or feelings."*

Attempting to assess this approach, we carried out a pilot test using a camera for recording a student[6] while performing a similar robot programming activity as in the protocol. This data will not be included in the dataset. We used a pre-trained video-to-text model (Li et al., 2024) to generate a scene description. The model was fed with the same instruction provided to the annotators through the following prompt:

> *"Describe the child's behavior and reactions during the video."*

The model produced the following response:

> *"The child is seated in a black chair with a white interior, and the room has a neutral color palette with a red door and a white wall. The child's hair is dark and styled, and they are wearing a white shirt. The child's facial expressions and body language suggest that they are engaged in a conversation or activity."*

As it can be noticed, the model focused more on describing the environment rather than the child's behavior, which is the type of information needed for better leveraging the video modality to generate notes as in an indirect observation setting. To verify whether clearer instruction could help redirect the model's focus, we provided a more specific prompt:

> *"Describe the child's behavior and reactions during the video, don't focus on the environment, just on the child's behavior."*

However, the response remained similar:

> *"The child is seated in a room with a beige wall and a red door. The child is wearing a*

[4] Using the `detect_silence` function from the `pydub` library.

[5] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[6] We asked for authorization of the child's tutor.

white shirt and has short dark hair. The room has a neutral color palette with a red door and a white table. The child is looking at the camera with a neutral expression."

This highlights some of the model's limitations, as it primarily focuses on the child's general appearance instead of offering a meaningful behavioral description, despite explicit instructions.

Phase 2: Emotion Annotation

An interface was developed for labeling purposes using **Label Studio** (Tkachenko et al., 2020-2022). This tool allowed annotators to select between the five emotion labels, as well as to rate the difficulty of annotation, activation, and valence. The interface displayed each audio segment along with its **Whisper**-generated transcription and the scene description provided by the human observer. The observer's description of the scene could serve as a support tool for the annotation. One person per participant provided this description; examples of these observations can be found in Appendix A. The annotators were instructed to rely primarily on the voice and, if it was difficult to label, to use the description of the actions as a reference. Additional details about the annotators and their training can be found in Appendix B.

Figure 3 shows the interface used for annotation purposes.

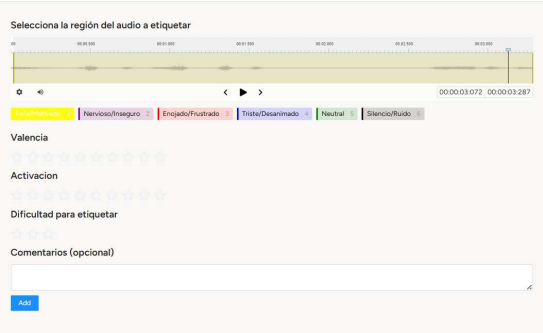

Figure 3: Label Studio interface.

To facilitate the labeling process, a correspondence was established between certain basic emotions and related feelings; for example, *happiness* was associated with the feeling of *motivation*, and *anger* with *frustration*. Annotators classified each segment into one of the following categories:

- **Happy/Motivated**: Higher voice, more energy, and varied intonation. It reflects enthusiasm, satisfaction, or determination.
- **Nervous/Insecure**: Characterized by an unstable tone, frequent pauses, and possible hesitations. It may express anxiety or lack of confidence.
- **Angry/Frustrated**: More intense voice, with a higher tone and abrupt pronunciation. It indicates annoyance, anger, or impatience.
- **Sad/Discouraged**: Distinguished by a lower tone, slow rhythm, and reduced vocal energy. It expresses discouragement, melancholy, or resignation.
- **Neutral**: Stable voice tone, consistent rhythm, and no marked emotional variations. It does not convey any specific emotion.

The aforementioned emotions were chosen for their relevance in learning contexts and their impact on children's performance and interaction. **Happy/Motivated** and **Sad/Discouraged** represent opposite states that directly influence engagement in activities. **Nervous/Insecure** and **Angry/Frustrated** are common when children face challenges, reflecting difficulties in emotional regulation. Finally, **Neutral** was included as a baseline reference, allowing differentiation between emotional states and those without evident affective load. These emotions cover a relevant spectrum of responses without introducing redundant or hard-to-distinguish categories in children's speech.

For segments labeled with an emotion, annotators also provided ratings for three additional aspects: **activation**, indicating the strength or energy level of the emotion; **valence**, reflecting how positive or negative the emotion is perceived by people (Valenza, Lanata, and Scilingo, 2011) (rated on a scale from **1 to 10**); and **labeling difficulty**, capturing how challenging it was to annotate the audio segment (rated on a scale from **1 to 3**).

Apart from annotating emotions, we add an option to highlight **Silence/Noise** for referring to moments without speech or external sound interferences that do not contain relevant verbal information.

The annotation process involved five annotators. Each segment was labeled by three annotators. For a label to be considered valid, at least *two annotators* had to assign the same category. In case of disagreement among all three, a fourth annotator reviewed the segment and determined the final label.

Additionally, the interface allowed for corrections to the transcriptions generated by **Whisper**, as the model occasionally produces hallucinations. For example, when the audio was unclear, it often inserted words like *"gracias"* (*"thanks"*) that was identified

as not present in the audio segment; another common error were unnecessarily repeated fragments, such as *"100 centímetros, 100 centímetros, 100 centímetros..."* (*"100 centimeters, 100 centimeters, 100 centimeters..."*, when in the audio the sentence occurs only once. Therefore, it was important to review and adjust the transcriptions to ensure their accuracy and to analyze the contexts in which the model tended to get confused. The inter-rater reliability was assessed using Krippendorff's Alpha (Krippendorff, 2011), resulting in a coefficient of 0.44.

3.1 Overview of CHERISH

CHERISH consisted of audio recordings of 31 participants during the robotics programming activity. For each participant, we have a set of segments annotated with a given categorical emotion as well as its respective transcription (both the automatic transcription and the one corrected by manual annotation were included). Information regarding *Participant Information*, *Recording Duration*, and *Audio Specifications* of CHERISH are shown in Table 2.

Participant Information	
# Participants	35
Gender Distribution	20 girls, 15 boys
Average Age	9.8 years
Recording Duration	
Total Duration	2.45 hours
Average Duration	5.4 seconds
Shortest Recording	0.05 seconds
Longest Recording	1 minute
Audio Specifications	
Format	WAV PCM
Sampling Rate	48,000 Hz
Channels	Mono
Bit Depth	16-bit

Table 2: Specifications and metadata of the CHERISH corpus.

After manual annotation concerning the presence of categorical emotions, there are 1563 segments distributed as shown in Table 3.

With respect to the CPQ applied to the participants, the results were interpreted according to the *second-order factors*. These values are obtained through factor analysis as follows, *QI evaluates anxiety*: a low score reflects good adjustment and life satisfaction, while a high score may indicate significant anxiety and possible maladaptation. *QII*

Label	**Percentage (%)**
Happy/Motivated	11.6
Neutral	31.9
Nervous/Insecure	46.9
Sad/Discouraged	3.6
Angry/Frustrated	6.0
Total	100.0

Table 3: Percentage distribution of emotion labels in CHERISH (Total = 1563).

differentiates between introversion (reserved, self-sufficient) and extraversion (sociable, uninhibited), traits that are advantageous depending on the context. *QIII* measures a range from calmness (sensitive, cautious) to excitability/hardness (active, impulsive, independent). Data distribution on the CPQ results are shown in Table 8 in the Appendix C. It is important to mention that each participant is associated with different factors.

Figure 4 illustrates the different modalities available in CHERISH as well as complementary information that may be valuable for automatic classification tasks, such as exploring the connection between personality and emotional responses. Note that video-based annotations in CHERISH are derived from *direct observation*. Currently, this version of the corpus does not include *indirect observation* modality due to the prioritization of spontaneous expressions of emotion. This design choice ensures that emotional responses are captured in context rather than through external interpretations.

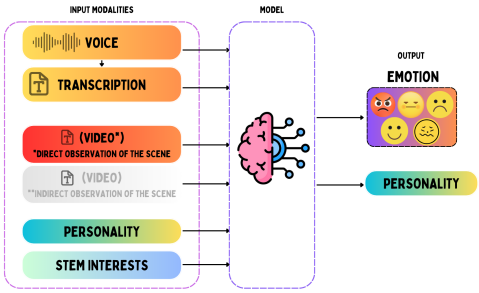

Figure 4: Modalities used for emotion recognition: voice, transcription, observer-described scene, interest in STEM, and personality.

Finally, it is worth mentioning that the notes regarding the *direct observation* for each participant are also included. The corpus is available under request[7] in an attempt to promote the research on children emotion recognition in Spanish. analysis of the rela-

[7]It can be requested in the emails previously given in this paper

tionship between personality and emotions is attached in Appendix D

4 *Preliminar Experiments*

Aiming to provide a baseline for this task, we propose a set of experiments. Emotion recognition was defined as a Multi-class classification problem of *Categorical emotions*, where the aim is to identify the emotional category[8] associated with a given sample.

Figure 5 illustrates the preliminary workflow followed during the experimental settings. The input consists of three modalities: audio recordings, transcriptions, and scene descriptions. Each of these inputs undergoes a specific pre-processing phase tailored to its modality, followed by feature extraction. Subsequently, the extracted features are passed through a classification model trained to identify emotional states. Finally, the predictions are evaluated using standard metrics to assess the models' performance.

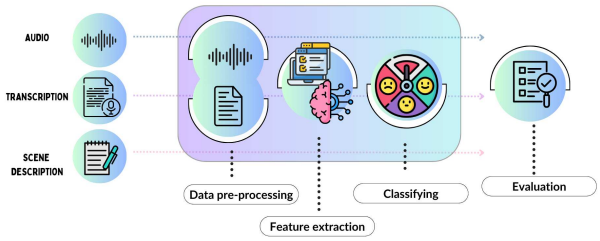

Figure 5: Diagram of the emotion recognition process.

For experimental purposes, we assessed the emotion recognition for each of the modalities we have on its own. For the *textual* modality (in both cases for the transcriptions and the scene descriptions), three text representations were exploited: traditional bag-of-words with a binary (**bin**) and Term Frequency – Inverse Document Frequency (**tf-idf**) weighting schemas, and by using the **[CLS]** vector obtained from a BERT-based model[9]. On the other hand, for the *audio* modality, we use **IS09**[10], **eGE**MAPS[11], and wav2vec (**w2v**)[12]. For classification purposes, a Support Vector Machine (SVM) was exploited, and as an evaluation metric, we used the F1 score. Table 4 shows the distribution of training and test data used for experiments. It should be noted that there was no overlap between participants in the training and test sets, which ensured that the evaluated models were exposed to entirely new data during the testing phase.

Label	Train (%)	Test (%)
Happy/Motivated	8.0	3.6
Neutral	21.3	10.6
Nervous/Insecure	31.9	15.0
Sad/Discouraged	2.6	1.1
Angry/Frustrated	4.7	1.3
Total	68.5	31.6

Table 4: Percentage distribution of emotion labels in the training and test sets relative to the full dataset (N = 1563).

Table 5 presents the F1-scores for each emotion across different modalities and feature extraction techniques. The best performance was obtained for *Nervous/Insecure*, most likely due to its more frequent representation in the dataset (Table 4), which facilitates model learning. In contrast, *Sad/Discouraged* showed near-zero scores, reflecting difficulties in detecting this class, likely caused by both data imbalance and confusion with *Nervous/Insecure* and *Neutral*, as observed in model predictions.

Happy/Motivated showed moderate results, particularly in transcriptions using **[CLS]** embeddings and in audio with **w2v**, suggesting these features could better capture expressive cues. *Neutral* was more accurately detected through scene descriptions and audio, possibly due to its more stable patterns. *Angry/Frustrated* remained low in all modalities, probably due to its limited presence in the data.

Preliminar Analisys on the Emotional Changes

It is important to highlight that the temporal sequence of the recordings was maintained. This allows us to observe the emotional variations of the participants throughout the entire session. Each challenge session lasted approximately 7 minutes. We decided to perform a pilot analysis with the idea of having first insights on the emotional changes captured during the robotic programming activity.

[8] The categories described in Section 3 were considered.

[9] In particular, we used the roberta-base-bne model(Fandiño et al., 2022).

[10] The openSMILE configuration file for IS09 emotion challenge (Eyben, Wöllmer, and Schuller, 2010) was exploited.

[11] We used the openSMILE configuration eGeMAPSv01b(Eyben, Wöllmer, and Schuller, 2010)

[12] https://huggingface.co/facebook/wav2vec2-large-xlsr-53-spanish

Emotion	Transcriptions			Scene Description			Audio		
	bin	**tf-idf**	**[CLS]**	**bin**	**tf-idf**	**[CLS]**	**IS09**	**eGE**	**w2v**
Happy/Motivated	0.12	0.24	0.29	0.14	0.16	0.07	0.12	0.06	0.26
Neutral	0.11	0.13	0.21	0.36	0.36	0.24	0.26	0.35	0.41
Nervous/Insecure	0.50	0.51	0.52	0.42	0.47	0.35	0.41	0.63	0.55
Sad/Discouraged	0.00	0.00	0.00	0.00	0.00	0.03	0.12	0.00	0.00
Angry/Frustrated	0.11	0.11	0.14	0.04	0.05	0.03	0.08	0.11	0.03

Table 5: Table summary of feature extraction across different modalities (F1 score).

Figure 6 shows the emotional changes experienced by a female participant of 11 years old who is in the last grade of elementary school. She did not complete the activity during the lapse time allowed. It is possible to observed a noticeable tendency towards negative emotions, particularly on **Nervous/Insecure** and **Angry/Frustrated**, which could be strongly related with the final result obtained, i.e., not completing the programming robot task.

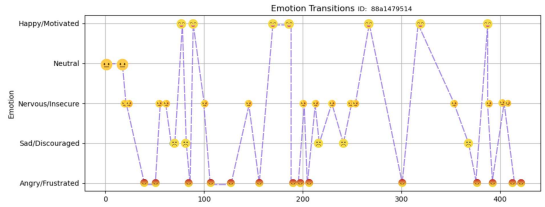

Figure 6: Graph of emotion variation throughout a test.

We have the intuition that the emotional changes through the activity performing will allow us to identify interesting and useful patterns for helping the participants to improve their academic performance.

5 Conclusions

In this paper, we introduce CHERISH a Spanish multimodal corpus comprising information of emotions, personality, and affinity for STEM fields of children between the ages of 9 and 13. CHERISH represents an advance in collecting children's emotional data in real-life contexts, combining multiple modalities such as audio, transcription, and scene descriptions in Spanish. This multimodal approach allows for a more comprehensive analysis of emotional responses. Although the corpus does not include indirect observation by third parties or automated coding tools, this opens an opportunity for future research integrating these complementary sources. Conducting a robotics workshop in a natural educational setting allowed for capturing children's spontaneous emotions, which is essential for the development of more realistic emotion recognition models. At this stage, data is annotated at the segment level with information about emotions in both categorical and dimensional models. Baseline experiments were performed using each of the modalities in CHERISH. We also analyzed the emotional changes experimented by the participants during the activity.

As future work, we intend to expand the CHERISH corpus by adding new recording sessions that feature a wider range of emotional expressions and a more balanced representation of the different classes. We are also interested in further investigating how to combine the information of the different modalities with the intention of improving the performance of automatic emotion recognition in children. Additionally, we will investigate the potential of utilizing the multimodal data collected to predict childhood personality traits. This could pave the way for new interdisciplinary research opportunities between psychology and artificial intelligence.

Acknowledgment

The first author gratefully acknowledges the support of Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) through the research grant [CVU No. 1286801], which made this study possible. We also thank the Lazaro Cardenas Elementary School, where the data collection took place. Thanks are due to PhD. Martha Leticia Gaeta González for her guidance in the design of the data collection protocol.

References

2025. LEGO MINDSTORMS — Inventa un robot. Accessed: March 29, 2025.

Duville, M. M., L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate. 2021. Mexican Emotional Speech Database Based on Semantic, Frequency, Familiarity, Concreteness, and Cultural Shaping of Affective Prosody. *Data*, 6(12):130.

Ekman, P. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Eyben, F., M. Wöllmer, and B. Schuller. 2010. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the ACM Multimedia Conference (MM)*, pages 1459–1462, Florence, Italy. ACM.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68.

Kalateh, S., L. A. Estrada-Jimenez, S. Nikghadam-Hojjati, and J. Barata. 2024. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access*, 12:103976–104019.

Krippendorff, K. 2011. Computing krippendorff's alpha-reliability.

Li, F., R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li. 2024. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models.

Lyakso, E. et al. 2015. EmoChildRu: Emotional Child Russian Speech Corpus. In *Speech and Computer. SPECOM 2015*, volume 9319 of *Lecture Notes in Computer Science*. Springer, Cham.

Martello Ortiz, O. M. and J. M. Arévalo Parra. 2017. Funcionamiento cognitivo y estados emocionales de un grupo de niños y adolescentes con bajo rendimiento académico. *Neuropsicología Latinoamericana*, 9(3), Dec.

Matveev, Y., A. Matveev, O. Frolova, E. Lyakso, and N. Ruban. 2022. Automatic speech emotion recognition of younger school age children. *Mathematics*, 10(14):2373.

Monchi, O., M. Petrides, V. Petre, K. Worsley, and A. Dagher. 2001. Wisconsin Card Sorting Revisited: Distinct Neural Circuits Participating in Different Stages of the Task Identified by Event-Related Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, 21(19):7733–7741.

Nojavanasghari, B., T. Baltrušaitis, C. E. Hughes, and L.-P. Morency. 2016. EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 137–144, New York, NY, USA. Association for Computing Machinery.

Pan, B., K. Hirota, Z. Jia, and Y. Dai. 2023. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, 561:126866, 10.

Pekrun, R. 2024. Control-Value Theory: From Achievement Emotion to a General Theory of Human Emotions. *Educational Psychology Review*, 36:83.

Pérez-Espinosa, H., J. Martínez-Miranda, I. Espinosa-Curiel, J. Rodríguez-Jacobo, L. Villaseñor-Pineda, and H. Avila-George. 2020. IESC-Child: An Interactive Emotional Children's Speech Corpus. *Computer Speech & Language*, 59:55–74.

Pérez-Espinosa, H., C. A. Reyes-García, and L. Villaseñor-Pineda. 2011. EmoWisconsin: An Emotional Children Speech Database in Mexican Spanish. In *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*, pages 62–71. Springer.

Porter, R. and R. Cattell. 1986. Cpq. *Cuestionario de personalidad para niños. Manual (3a edición) TEA ediciones.*

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Steidl, S. 2009. *Automatic Classification of Emotion Related User States in Spontaneous Children's Speech*. Ph.D. thesis, University of Erlangen-Nurnberg, Berlin, Germany.

Thompson, R. A. 1994. Emotion regulation: A theme in search of definition. *Monographs of the society for research in child development*, pages 25–52.

Tkachenko, M., M. Malyuk, A. Holmanyuk, and N. Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Valenza, G., A. Lanata, and E. P. Scilingo. 2011. The role of nonlinear dynamics in affective valence and arousal recognition. *IEEE transactions on affective computing*, 3(2):237–249.

A Examples of direct observation

The Table 6 shows some examples of direct observations, remembering that the instruction given was to write down the participant's behavior.

Observation in Spanish	Observation in English
"Realiza correcciones; sigue pensando en voz alta"	Makes corrections; continues thinking out loud
"Se le nota seguro, aunque titubea un poco al explicar"	Appears confident, although hesitates slightly when explaining
"Se pone un poco nervioso cuando le piden que explique qué está haciendo"	Gets a bit nervous when asked to explain what he is doing
"Mira confundido la pantalla tras fallar el recorrido"	Looks confused at the screen after the failed run
"Voltea los ojos hacia arriba mientras sigue haciendo cuentas"	Rolls eyes upward while continuing to do calculations
"Realiza la misma cuenta de distancia muchas veces"	Repeats the same distance calculation multiple times
"Realiza un intento; sonríe y grita un poquito cuando no sale en el primer intento; vuelve a una expresión neutra"	Makes an attempt; smiles and lets out a small shout when it doesn't work on the first try; returns to a neutral expression

Table 6: Examples of direct observations during the activity.

B Statistics of annotators

The Table 7 presents information about the annotators, whose ages range from 22 to 45 years.

Annotator	Academic Background	Specialization	Gender
A1	PhD in Computer Science	Biosignals	Male
A2	PhD in Computer Science	Natural Language Processing	Female
A3	Medical Student	Medicine	Female
A4	Medical Student	Medicine	Female
A5	Biomedical Engineering Student	Biomedical Sciences	Male

Table 7: Annotators' academic profiles, specializations, gender, and age.

C Distribution of personality

QI	Participants
High Anxiety	14
Average Anxiety	13
Low Anxiety	8
QII	**Participants**
Average	18
Extraversion	9
Introversion	8
QIII	**Participants**
Average	19
Excitability/Hardness	14
Calm	2
Total(per trait)	35

Table 8: Distribution of personality traits based on QI (Anxiety), QII (Introversion/Extraversion), and QIII (Calm/Excitability) among participants. Each dimension includes 35 participants.

D Analysis of emotional expression

Girls tend to express more negative or insecure emotions, whereas boys show a higher prevalence of positive or reactive emotions. Negative emotions appear to be more frequent in slightly older children, and when analyzing the psychological Profile, we can say that: **Anxiety** is strongly associated with emotions such as *Sadness* and *Frustration*. **Introversion** and **Excitability/Hardness** also show a tendency toward negative emotions, while **Extraversion** and **Calmness** are linked to positive or neutral emotions.

The Figures 7 and 8 show the percentage relationship between age, personality traits, and emotions.

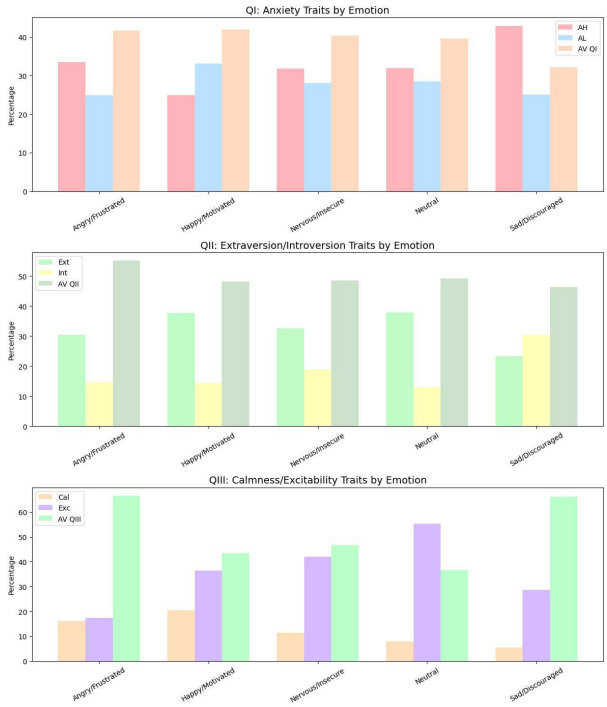

Figure 7: Personality abbreviations: AV (Average), Anxiety High (AH), Anxiety Low (AL), Extraversion (Ext), Introversion (Int), Calm (Cal), Excitability (Exc).

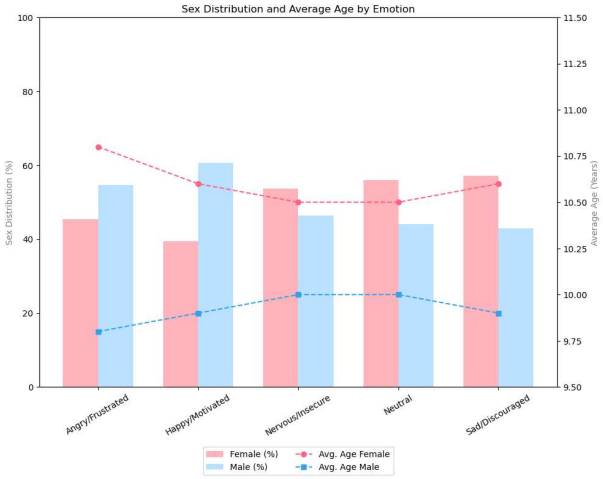

Figure 8: Sex distribution and average age by emotion.