Automatic Estimation of Personality in Children from Spontaneous Voice in Learning Interactions

Estimación automática de perfiles de personalidad en niños a partir de voz espontánea en interacciones de aprendizaje

Gabriela Judith Alvarado-Flores, Humberto Pérez-Espinosa, Delia Irazú Hernández-Farías
Coordinación de Ciencias Computacionales
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Luis Enrique Erro 1, Tonanztinztla, México
{gabrielaj.alvarado,humbertop,dirazuhf}@inaoep.mx

Abstract: Children's voices convey subtle acoustic cues that reflect paralinguistic information such as individual personality traits, providing a promising, non-intrusive approach for studying socioemotional development in educational contexts. This study investigates the feasibility of estimating personality traits in children aged 8 up to 12 years from spontaneous speech recorded during a programming and problem-solving activity designed as an academic challenge. Personality profiles were obtained using the Children's Personality Questionnaire, a validated psychometric instrument specifically developed for children. A computational pipeline was developed comprising audio segmentation, extraction of acoustic descriptors, feature selection, and stratified cross-validation at the individual level. Several classification paradigms were evaluated, including both multiclass and binary configurations. The best Macro-F1 values were obtained in binary configurations, reaching 0.756 for **QI: Adjustment/Anxiety**, 0.799 for **QII: Introversion/Extraversion**, and 0.825 for **QIII: Calm/Excitability**. However, the **QIII** result should be interpreted with caution due to the limited representation of minority classes. The obtained findings suggest that children's voices during challenging learning interactions encode meaningful information that is predictive of personality traits.
Keywords: Children speech, Child personality, Children's Personality Questionnaire (CPQ), Machine learning.

Resumen: Las voces infantiles transmiten señales acústicas sutiles que reflejan información paralingüística, como rasgos individuales de personalidad, lo que proporciona un enfoque prometedor y no intrusivo para estudiar el desarrollo socioemocional en contextos educativos. Este estudio investiga la viabilidad de estimar rasgos de personalidad en niños de 8 a 12 años a partir del habla espontánea registrada durante una actividad de programación y resolución de problemas diseñada como un reto académico. Los rasgos de personalidad se obtuvieron mediante el *Children's Personality Questionnaire*, un instrumento psicométrico validado y desarrollado específicamente para niños. Se desarrolló un esquema computacional compuesto por segmentación de audio, extracción de descriptores acústicos, selección de características y validación cruzada estratificada a nivel individual. Se evaluaron varios paradigmas de clasificación, incluyendo configuraciones multiclase y binarias. Los mejores valores de Macro-F1 se obtuvieron en configuraciones binarias, alcanzando 0.756 para **QI: Ajuste/Ansiedad**, 0.799 para **QII: Introversión/Extraversión** y 0.825 para **QIII: Calma/Excitabilidad**. Sin embargo, el resultado de **QIII** debe interpretarse con cautela debido a la representación limitada de clases minoritarias. Estos hallazgos sugieren que las voces de los niños durante interacciones de aprendizaje y juego desafiantes codifican información significativa que permite predecir rasgos de personalidad.
Palabras clave: Habla infantil, Personalidad infantil, Children's Personality Questionnaire (CPQ), Aprendizaje automático.

1 Introduction

Identifying personality traits in children holds significant relevance across different contexts. In academic settings, a nuanced understanding of personality supports the development of personalized teaching strategies that enhance learning outcomes, increase student engagement, and strengthen collaborative and self-regulatory skills (Melzer and Schoop, 2015). For mental health professionals, early characterization of personality traits facilitates the timely identification of potential psychological conditions, including anxiety, impulsivity, and emotional sensitivity (Mahaffey et al., 2016).

Traditional methods for assessing children's personality, including questionnaires, structured interviews, and behavioral observations, remain valuable and accurate. However, they are often time-consuming and require trained personnel. Additionally, some self-assessment items may be difficult for children to interpret consistently, potentially introducing biases or reducing validity (Aishwarya and Veena, 2023). These limitations restrict scalability and may hinder broader implementation in diverse contexts.

Personality assessment in children requires instruments that are developmentally appropriate and sensitive to age-related characteristics. *Children's Personality Questionnaire (CPQ)* is a standardized psychometric instrument developed to assess personality traits in school-aged children. Grounded in Cattell's personality theory (Porter and Cattell, 2016), the CPQ is part of the 16 Personality Factors framework and evaluates 14 primary personality dimensions along with three higher-order factors. Two factors from the original model were excluded to ensure developmental appropriateness and psychometric validity for this age group, thereby focusing on traits that can be reliably and behaviorally manifested in children.

Recent advances in computational behavioral analysis suggest that expressive signals, such as speech, can serve as meaningful indicators of personality traits. Evidence from adult populations, as well as emerging studies in children (Pérez-Espinosa et al., 2022), indicates that computational paralinguistics can offer reliable and scalable alternatives for personality assessment. Prosodic, acoustic, and linguistic features in children's speech have been shown to convey cues associated with extraversion, emotionality, and self-control, even in everyday, non-clinical contexts (Zhang and Gu, 2024).

Automatic personality estimation based on speech analysis can be particularly promising in naturalistic interactive contexts, such as learning environments or play-based problem-solving tasks. This approach can capture authentic behavioral signals while reducing the burden on children and educators. Publicly available datasets comprising personality assessments with voice recordings collected during authentic learning activities remain limited (Rubio Franco, 2020). The capacity to infer personality traits from speech during interactive contexts offers significant potential for real-time adaptation of instructional strategies, enabling the personalization of feedback and responses. Moreover, such models constitute valuable tools within educational psychology, facilitating the systematic investigation of the relationships between personality traits, learning styles, and academic performance.

The objectives of this study are threefold:

- To develop a computational model to estimate personality traits in children aged 8 up to 12 from their voice recorded during learning and play activities.
- To evaluate different machine learning settings for CPQ personality dimensions.
- To advance research on automatic personality estimation in children by analyzing a corpus of voice recordings annotated with CPQ personality profiles.

To address these objectives, we propose a computational pipeline that integrates signal preprocessing, extraction of acoustic and prosodic features, feature selection, and individual-level stratified cross-validation. This methodology provides a reproducible evaluation framework for automatic personality estimation in children.

The rest of this paper is organized as follows. Section 2 reviews previous studies on automatic personality assessment from speech. Section 3 describes the experimental design we performed. Section 4 reports the obtained results. Section 5 discusses the main insights and limitations of the study. Finally, Section 6 concludes the paper and outlines future research directions.

2 *Related work*

A wide range of computational approaches has been explored for personality recognition from speech. From those using traditional machine learning techniques (Pérez-Espinosa et al., 2022; Lukac, 2024) to deep learning approaches capable of learning high-dimensional representations directly from raw or minimally processed signals (Hammoud et al., 2024), as well as multimodal frameworks that integrate speech with facial expressions and physiological signals (Bhin and Choi, 2025; Wei et al., 2018).

Despite progress in feature engineering, classification algorithms, and evaluation protocols, research focusing on automatic personality assessment in children remains scarce. One major reason is the lack of speech corpora containing validated and developmentally appropriate personality annotations for child populations. In addition, children's speech presents characteristics that differ from adult speech, including higher acoustic variability, developmental changes in vocal production, shorter and less structured speaking turns, and age-related articulation differences (Lee, Potamianos, and Narayanan, 1999). These factors complicate feature extraction and often reduce the robustness of models originally optimized for adult speech.

Only a limited number of studies have explored automatic personality detection in children. Solera-Ureña et al. (Solera-Ureña et al., 2016) investigated personality assessment using heterogeneous speech corpora within the *OCEAN* framework (Wartberg et al., 2023). They analyzed several acoustic feature sets (including IS2012 (Schuller et al., 2012) and eGeMAPS (Eyben et al., 2016)) and knowledge-based descriptors. Their results suggested that certain acoustic–prosodic cues associated with *Extraversion* and *Agreeableness* remain relatively stable across adults and children. Pérez-Espinosa et al. (Pérez-Espinosa et al., 2022) investigated personality detection in children using the *Children's Personality Questionnaire* together with low- and high-level acoustic descriptors. The experiments were conducted on the *IESC-Child-2* corpus, which contains spontaneous speech collected in a game-based environment involving LEGO Mindstorms robots. Unlike earlier resources, this corpus incorporates a validated child personality instrument, providing a more reliable foundation for computational personality modeling.

Most existing corpora rely on semi-controlled interaction settings and provide limited evidence of how personality traits emerge during authentic learning or problem-solving activities. Speech produced in structured educational challenges can reveal behavioral patterns associated with motivation, persistence, emotional regulation, and social interaction, all of which are relevant to personality expression. However, current resources rarely capture such contexts. In summary, while previous studies demonstrate the feasibility of personality inference from speech, research focused on children remains underdeveloped due to the scarcity of suitable datasets, the limited availability of validated developmental personality labels, and the lack of recordings collected in naturalistic educational environments.

3 *Methodology*

3.1 Dataset Description

The dataset used in this study is the *CHERISH* corpus (Gahona Castillejos, Hernández Farías, and Pérez-Espinosa, 2025). It includes spontaneous emotional child speech along with psychometric personality assessments. Each participant performed an individual robotics programming task designed to induce emotions and promote spontaneous speech. During the activity, the children were asked to complete missions under a *think-aloud protocol*, by moving a robot from point A to point B while verbalizing their thoughts aloud. This procedure allowed collecting natural and spontaneous speech samples representative of problem-solving situations. Each participant was evaluated using the CPQ psychometric instrument. We focused on the three secondary-order factors: **QI (Adjustment/Anxiety):** assesses adjustment and life satisfaction, **QII (Introversion/Extraversion):** measures sociability and self-sufficiency, and **QIII (Sensitivity/Tough-mindedness):** evaluates behavioral stability and impulsivity.

The *CHERISH* corpus consists of recordings from 35 Spanish-speaking children (20 girls and 15 boys) aged 9 up to 12 years. Although the dataset presents a slight gender imbalance, grouping male and female children in this age range into a unified computational framework is methodologically sound. Unlike adult populations, prepubescent children do not exhibit sexual dimorphism in vocal tract anatomy (Vorperian et al., 2005). While subtle acoustic differences in voice quality (such as jitter and shimmer) begin to emerge at this stage (Banik, Arya, and Kant, 2015), primary parameters like the fundamental frequency remain relatively close between genders (averaging between 240 and 260 Hz) until the onset of vocal mutation (puberty), which typically begins after age 12 (Rodríguez, 2017; Nercelles and Centeno, 2020). Speech recordings were stored in WAV PCM format, with a sampling rate of 48 kHz, 16-bit depth, and a single channel. The complete corpus contains approximately 2.45 hours of audio, with an average duration of 5.4 seconds per segment. The audio data were preprocessed to remove silences and were provided as voice-active fragments of varying duration. In this work, these segments were subsequently normalized to a uniform duration of 6 seconds to facilitate the acoustic characterization.

3.2 Data Labeling Criteria for CPQ Personality Dimensions

According to the CPQ manual, each global factor is primarily defined by opposite poles: low **QI** reflects *emotional adjustment*, whereas high **QI** reflects *anxiety*; low **QII** corresponds to introversion and high **QII** to extraversion; and low **QIII** indicates calmness and sensitivity, while high **QIII** represents excitability and toughness (Porter and Cattell, 2016). It is well-documented in applied machine learning that introducing artificial cutoffs to discretize continuous psychological variables can introduce classification bias, as borderline cases might be misrepresented, and subtle variance is lost. To mitigate this risk and avoid relying on a single, potentially biased threshold, we systematically evaluated four distinct labeling schemes based on CPQ scores:

a) ***Classical Psychometric Categories.*** Specialists often rely on the *STEN* (Standard Ten) scale (Andrade, 2021) to interpret the CPQ results. We adopted the applied-psychometric convention of dividing the scale into: *low* (1.0-3.9), *medium* (4.0-6.9), and *high* (7.0-10) to preserve the interpretive meaning of the poles (Wojciechowska, Matkowski, and Pawłowski, 2022).

b) ***Quartile-based Categories.*** Inspired by research using machine learning in psychology and social sciences (Zainudin, Ng, and Khor, 2021), we generated three groups of data by splitting at the 25, 50, and 75 percentiles of the *STEN* scale. It is worth noting that the empirical distribution of CPQ scores in our dataset tends to approximate a bell-shaped curve, with most children concentrated around intermediate values and fewer cases at the extremes. By dividing the scores into quartiles, denoted as **q1**, **q2-q3**, and **q4**, we captured this natural tendency, achieving a more balanced representation of the classes and mitigating bias when training the classifiers. Figure 1 shows the distribution of CPQ dimensions and *STEN* categories when applying the quartile-based schema.

c) ***Median-based Categories.*** We used the median value as the cut-off point to generate two classes, dividing the sample into participants scoring below and above the average. This approach provides a balanced distribution, facilitating both the classification task and the interpretability of the results. Moreover, it is conceptually aligned with the CPQ manual, which characterizes each global factor in terms of opposing poles.

d) ***Fixed-threshold Categories.*** Another classic interpretation of the CPQ establishes that a *STEN* score equal to 5 represents a *neutral* point on the scale, which is then used to determine the orientation towards one pole or the other. For instance, in **QII**, scores of 5 or below correspond to *introversion*, whereas those above 5 correspond to *extraversion*. This rule-based criterion does not rely on the empirical distribution of the sample, but directly follows the conventional interpretation of the dimensions of the CPQ.

Each of the aforementioned schemes represents a different discretization criterion, enabling the comparison across different levels of granularity and assessing how boundary placement affects the estimation of children's personality traits. Tables 1 and 2 show the distribution of children according to the labeling schemes in multiclass and binary settings, respectively. Table 3 summarizes the distribution of audio segments across labeling schemes, CPQ dimensions, and class categories in the *CHERISH* corpus.

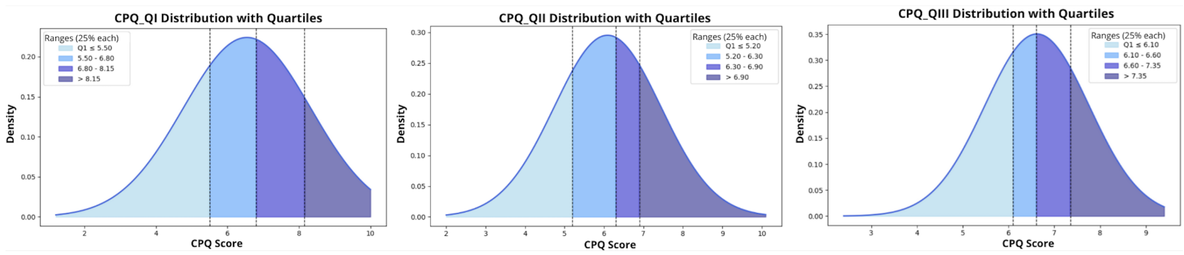

Figure 1: Distribution of CPQ scores by dimension with quartile-based cutoffs. Vertical lines indicate **QI**, median, and **QIII**.

Scheme	Dimension	Class		
		Low	Medium	High
Classical psychometric	QI	8	12	15
	QII	8	18	9
	QIII	2	19	14
Quartile-based	QI	10	16	9
	QII	10	16	9
	QIII	10	16	9

Table 1: Distribution of children by class in the multiclass schemes.

Scheme	Dimension	Low	High
Median-based	QI	19	16
	QII	19	16
	QIII	18	17
Fixed threshold (≤ 5)	QI	8	27
	QII	8	27
	QIII	2	33

Table 2: Distribution of children by class in the binary schemes.

Labeling Schema	Dimension	Low	Medium	High
Classical psychometric	**QI**	584	827	789
	QII	417	1111	672
	QIII	261	989	950
Quartile-based	**QI**	762	1076	362
	QII	501	1027	672
	QIII	563	1165	472
Median-based	**QI**	1359	–	841
	QII	1030	–	1170
	QIII	1160	–	1040
Fixed-threshold	**QI**	584	–	1616
	QII	417	–	1783
	QIII	261	–	1939

Table 3: Overall distribution of audio segments across labeling schemes, CPQ dimensions and classes. Each configuration contains a total of 2200 audio segments.

3.3 Data Preprocessing and Feature Extraction

Before feature extraction, each recording was analyzed to verify its duration and acoustic integrity, consolidating information by child and personality class. To identify potential anomalies, three standard low-level acoustic descriptors (Root Mean Square (RMS), spectral entropy, and Zero Crossing Rate (ZCR)) were computed[1] Segments showing abnormally low energy or low spectral entropy were flagged as potentially non-informative, since such patterns typically correspond to silence or stationary background noise in speech signals (Mihalache and Burileanu, 2022). The used thresholds[2] were empirically determined based on exploratory analysis of this corpus at a sampling rate of 16 kHz. Then, all flagged recordings were manually reviewed to confirm whether they contained meaningful vocalization or only residual noise. Recordings confirmed as silence or irrelevant noise were excluded.

Although no fixed segment length has been established for children's speech analysis, several corpora and emotion-recognition studies in children employ segments of a few seconds to ensure stable prosodic representation (Milling et al., 2022; Matveev et al., 2022). Therefore, six seconds was adopted as a representative and consistent analysis window, long enough to capture prosodic and emotional variability without requiring excessive padding or truncation. Shorter clips were repeated and truncated to reach the target length, whereas longer ones were cropped from the beginning to avoid initial hesitation and recording artefacts. This normalization ensured comparable feature extraction across children and personality classes.

[1]These parameters were chosen following the general criteria of Voice Activity Detection and the ITU-T P.56 recommendations for active speech level estimation (International Telecommunication Union (ITU), 2011).

[2]Values are RMS $< .01$ or spectral entropy < 2.5.

Segment-Level Feature Extraction. Both cepstral and prosodic-spectral descriptors were extracted from each 6-second segment:

- `MFCCs`: Capture vocal tract characteristics. Because of their ability to extract useful information for a variety of linguistic and paralinguistic tasks, MFCCs are a de facto standard in Speech Emotion Recognition (Abdul and Al-Talabani, 2022).
- `Prosodic and spectral parameters`: Features such as `pitch`, `jitter`, and `shimmer` capture variations in intonation, stability, and voice quality.

Prosodic and spectral features complement MFCC coefficients. Time-domain and spectral features provide complementary information about energy and complexity, adding detail about speech rhythm and style. We employed Spectral Entropy, which reflects the complexity of the spectrum, allowing us to capture dynamic patterns of speech related to emotion (Bhangale and Kothandaraman, 2023).

3.4 CPQ score estimation as a classification task

We tackled the assessment of CPQ personality traits as a classification task within a speaker-independent framework. This configuration addresses key challenges of this, most notably the limited amount of training data and the difficulty of model adaptation, while also enabling a more realistic evaluation scenario in which the target speaker is not seen during training (Schuller et al., 2005). A critical concern when dealing with small datasets is that acoustic features might be heavily correlated with gender or individual physiological traits, which would lead the model to memorize specific subjects rather than learn the underlying personality dimensions. Therefore, a *speaker-independent* schema was implemented in a "leave-one-speaker-out" manner, using Leave-One-Group-Out (LOGO) by grouping all the data of each participant across each experiment. We evaluated six different classification algorithms[3] namely: Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Gaussian Naïve Bayes (GNB), and Multilayer Perceptron (MLP). In all cases, *GridSearchCV* was employed within the LOGO validation, using accuracy as the selection metric. For each fold, the best-performing pipeline was retrained on the training set and subsequently evaluated on the left-out child.

Preprocessing and Feature Selection

Feature selection was performed using the *SelectKBest* method with the f-ANOVA statistical test, evaluating three subset sizes of features: $k \in \{7, 10, 14\}$. When required, the *StandardScaler* was integrated into the pipeline to ensure that normalization parameters were fitted exclusively on the training data within each cross-validation fold. Furthermore, we also conducted preliminary experiments with wav2vec 2.0 (Baevski et al., 2020), a self-supervised speech representation for feature extraction.[4]

Evaluation Metrics

For each combination (*labeling scheme*, *model*, and *CPQ dimension*), the following metrics were reported: a) Aggregated accuracy (Acc) across all LOGO folds; and b) Standard classification metrics: Precision (P), Recall (R), and Macro-F1 score (F1).

Model Interpretability

We conducted a post-hoc interpretability analysis using SHAP (SHapley Additive exPlanations), which enables a transparent quantification of each feature's contribution to the model's decision-making process. Given the variety of classifiers in our pipeline, we employed *TreeExplainer* for ensemble-based models (RF) and *KernelExplainer* for kernel-based and linear models (SVM and LR), ensuring a consistent and model-agnostic evaluation of the decision boundaries.

4 Results

4.1 Classical Psychometric Scheme

Table 4 summarizes the global performance for this scheme. For each model, three separate classifiers were trained (**QI**, **QII**, and **QIII**). To obtain a single scheme-level

[3] We used the *Sklearn* implementation of these classifiers.

[4] Despite its competitive performance in the literature, we are not including results with this model because they were lower than with more traditional representations.

score per metric, the nine class-level values (three dimensions × three classes) were concatenated and averaged without weighting. Thus, the values in the table represent global, dimension-aggregated performance, rather than the performance of any specific CPQ dimension.

Classifier	P	R	F1	Acc
RF	0.569	0.528	0.525	0.629
LR	**0.583**	**0.633**	**0.596**	0.619
SVM	0.543	0.547	0.529	**0.667**
KNN	0.455	0.406	0.457	0.552
GNB	0.485	0.485	0.474	0.533
MLP	0.475	0.546	0.461	0.495

Table 4: Global macro-averages in the Classical Psychometric scheme.

Table 5 presents the dimension-level macro-F1 for each classifier. RF achieved the best performance in **QI**, whereas LR obtained the highest F1 in both **QII** and **QIII**. MLP consistently yielded the lowest performance across the three dimensions. At the dimensional level, macro-F1 provided a reliable evaluation of classifier performance. RF achieved the highest F1 in **QI**, while LR obtained the best results in both **QII** and **QIII**. This confirms that macro-F1 is more informative than accuracy under class imbalance.

Classifier	QI	QII	QIII
RF	**0.648**	0.469	0.467
LR	0.581	**0.531**	**0.677**
SVM	0.614	0.427	0.545
KNN	0.480	0.501	0.389
GNB	0.587	0.483	0.352
MLP	0.443	0.490	0.450

Table 5: Results in terms of F1-score for each dimension for the classical psychometric scheme.

Overall, RF was the strongest model for **QI**, while LR provided the most stable performance in **QII** and **QIII**. The classical psychometric scheme remains strongly affected by class imbalance, especially in **QII** and **QIII**, where minority groups have only a few instances. However, several classes achieved consistent precision-recall patterns, indicating that the acoustic markers of the child's expression carry meaningful personality information. These findings motivated the exploration of quartile- and median-based schemes to improve class balance and interpretability across CPQ dimensions.

4.2 Quartile-based Scheme

Table 6 summarizes the global results, while Table 7 reports values for each CPQ dimension under the quartile-based configuration. Overall, LR, SVM, and KNN showed the most consistent behavior across dimensions, with performance levels that remained stable despite the added difficulty of multi-class classification.

Classifier	P	R	F1	Acc
RF	0.447	0.521	0.527	0.571
LR	0.598	**0.612**	**0.592**	**0.590**
SVM	**0.607**	0.611	0.586	**0.590**
KNN	0.601	0.578	0.568	0.571
GNB	0.432	0.461	0.430	0.438
MLP	0.508	0.520	0.498	0.514

Table 6: Global macro-averages in the quartile-based scheme.

Model	QI	QII	QIII
RF	0.586	0.508	0.519
LR	**0.630**	0.545	**0.601**
SVM	0.627	0.591	0.541
KNN	0.616	0.586	0.500
GNB	0.542	0.326	0.423
MLP	0.533	**0.620**	0.342

Table 7: Results in terms of F1-score for each dimension for the quartile-based scheme.

For **QI**, LR achieved the highest F1, closely followed by SVM and KNN. For **QII**, the highest F1 was obtained by MLP, followed by SVM and KNN. For **QIII**, LR achieved the best F1, outperforming SVM and RF. Overall, the quartile-based scheme yielded moderate but stable class separability across the three CPQ dimensions. While accuracy favored majority classes, F1 highlighted the models' ability to capture minority and intermediate profiles, revealing that LR, SVM, and MLP provided the most balanced discrimination under the three-level quartile structure.

4.3 Median-based Scheme

Table 8 summarizes the global macro-average results, and Table 9 reports the dimension-level F1 scores. Overall, SVM obtained the strongest performance across the median-based configuration, achieving the highest macro-F1 in **QI** and **QIII**, while both LR and SVM reached the best F1 in **QII**.

For **QI**, SVM obtained the highest F1, outperforming LR and KNN. For **QII**, both

Classifier	P	R	F1	Acc
RF	0.684	0.683	0.683	0.686
LR	0.713	0.713	0.712	0.714
SVM	**0.755**	**0.753**	**0.751**	**0.752**
KNN	0.696	0.697	0.693	0.695
GNB	0.653	0.632	0.626	0.628
MLP	0.644	0.635	0.628	0.628

Table 8: Global macro-averages in the median-based scheme.

Classifier	QI	QII	QIII
RF	0.681	0.742	0.627
LR	0.712	**0.799**	0.627
SVM	**0.742**	**0.799**	**0.712**
KNN	0.714	0.708	0.656
GNB	0.712	0.597	0.570
MLP	0.650	0.656	0.570

Table 9: Results in terms of F1 by classifier and dimension for the median-based scheme.

LR and SVM achieved the highest F1. For **QIII**, SVM again achieved the best F1. Overall, both SVM and LR demonstrated strong and stable performance under this scheme. Results suggest that median-based binarization provides a balanced compromise between interpretability and class separability, enabling MFCC-based and prosodic features to distinguish opposite behavioral poles with reliable accuracy.

4.4 Fixed-threshold Scheme

Table 10 summarizes the global macro-average metrics, and Table 11 reports the macro-F1 values for each CPQ dimension. Under this configuration, the best-performing models were LR for **QI** and **QII**, and RF for **QIII**.

Classifier	P	R	F1	Acc
RF	**0.728**	0.606	0.624	0.828
LR	0.687	0.662	**0.668**	0.668
SVM	0.647	0.700	0.648	**0.857**
KNN	0.550	0.556	0.539	0.840
GNB	0.563	0.586	0.571	0.780
MLP	0.654	**0.747**	0.665	0.809

Table 10: Global macro-averages in the fixed-threshold scheme.

At the dimensional level, macro-F1 provided the clearest view of classifier behavior under this configuration. The obtained results emphasized minority-class behavior and showed that prosodic and spectral cues remain reliable predictors of children's CPQ profiles even under strict cut-points.

Classifier	QI	QII	QIII
RF	0.533	0.514	**0.825**
LR	**0.756**	**0.763**	0.485
SVM	0.703	0.756	0.485
KNN	0.698	0.435	0.485
GNB	0.602	0.626	0.485
MLP	0.720	0.514	0.761

Table 11: Results in terms of F1 by classifier and dimension in the fixed threshold-based scheme.

4.5 SHAP Analysis

Table 12 details the features selected through the f-ANOVA process ($k \in \{7, 10, 14\}$) and their ranking by predictive impact across the 12 experimental scenarios. Biological *sex* emerged as the primary predictor in nearly all models, reinforcing its role as a fundamental baseline for interpreting acoustic variance in children's speech. Beyond *sex*, the spectral envelope, represented by MFCCs, accounted for the majority of the predictive weight. Specifically, MFCC 1, 3, and 10 were consistent markers for all personality dimensions, suggesting that these coefficients capture stable resonance patterns related to socio-emotional traits. In binary schemes, a higher reliance on prosodic descriptors like *pitch* and *jitter* was observed, particularly for the **QI** (*Anxiety*) and **QII** (*Extraversion*) dimensions, indicating that voice stability and fundamental frequency are key for distinguishing between polar personality profiles. These results seem to confirm that the models effectively leverage a compact prosodic signature to estimate personality profiles with high precision.

Although changes in the vocal tract are typical of puberty, this study aligns with research showing gender differences in personality from early childhood (Slobodskaya and Kornienko, 2021). Girls often score higher in *agreeableness* and *conscientiousness* (Slobodskaya and Kornienko, 2021), matching CPQ patterns. *Sex* may reflect both acoustic variability and shortcut effects. Beyond demographics, SHAP links model decisions to known acoustic properties. According to (Tracey et al., 2023), MFCC 1 reflects overall energy, while MFCC 3 contrasts mid-range versus low-plus-high frequencies. Higher-order MFCC 10 capture finer spectral variations, although their physical interpretation blurs as the order increases. The relevance of MFCC 1, 3, and 10 suggests personality dif-

Dimension	Classifier	Ordered Features (from highest to lowest SHAP impact)
***Classical Psychometric* (Multiclass)**		
QI	RF ($k = 7$)	Sex, MFCC 8, MFCC 4, MFCC 3, Jitter, MFCC 11, Pitch
QII	SVM ($k = 14$)	Sex, MFCC 3, MFCC 10, MFCC 1, MFCC 2, Jitter, Entropy, MFCC 5, MFCC 11, Shimmer, MFCC 12, Pitch, RMS, ZCR
QIII	SVM ($k = 14$)	Sex, MFCC 3, MFCC 11, MFCC 1, MFCC 4, Jitter, Shimmer, MFCC 2, MFCC 6, MFCC 12, RMS, MFCC 10, MFCC 7, MFCC 8
***Quartile-based* (Multiclass)**		
QI	LR ($k = 14$)	MFCC 4, Sex, RMS, MFCC 10, MFCC 1, Age, ZCR, Entropy, MFCC 12, MFCC 11, MFCC 3, Shimmer, Jitter, Pitch
QII	SVM ($k = 10$)	MFCC 1, Sex, MFCC 12, MFCC 10, Shimmer, ZCR, Entropy, Pitch, MFCC 11, Jitter
QIII	LR ($k = 7$)	Sex, MFCC 11, MFCC 1, MFCC 4, MFCC 2, Jitter, Shimmer
***Median-based* (Binary)**		
QI	SVM ($k = 14$)	Sex, MFCC 4, MFCC 5, RMS, MFCC 9, Shimmer, MFCC 1, MFCC 10, MFCC 2, MFCC 8, MFCC 11, MFCC 12, MFCC 7, Jitter
QII	LR($k = 14$)	MFCC 1, RMS, MFCC 10, MFCC 3, Entropy, MFCC 5, ZCR, Jitter, MFCC 11, MFCC 4, MFCC 12, Shimmer, Pitch, Sex
QIII	SVM ($k = 7$)	MFCC 11, Age, MFCC 8, Sex, MFCC 1, MFCC 13, Jitter
***Fixed-threshold* (Binary)**		
QI	LR ($k = 7$)	Jitter, MFCC 1, Sex, MFCC 10, Entropy, MFCC 3, MFCC 4
QII	LR ($k = 14$)	MFCC 2, Pitch, MFCC 13, MFCC 1, Sex, MFCC 12, Entropy, Age, MFCC 5, RMS, ZCR, MFCC 10, Shimmer, Jitter
QIII	RF ($k = 7$)	Sex, MFCC 11, MFCC 10, MFCC 1, MFCC 3, Pitch, MFCC 7

Table 12: Complete feature ranking by SHAP importance for all labeling schemes and CPQ dimensions.

ferences are encoded across the spectral envelope, from global energy to fine resonance. Furthermore, *jitter* and *pitch* importance in **QI** confirms that phonatory stability and frequency variability aid discrimination. Overall, the model leverages meaningful biomarkers rather than arbitrary patterns.

4.6 Analyzing the role of *Sex* as a key variable in our proposal

To assess the influence of the variable *Sex* in the proposed approach, an ablation test removing this variable was performed. Table 4.6 shows the best-performing Macro-F1 value obtained for each dimension level, considering all features (*All*) and when *Sex* is removed (*W/O Sex*). In most cases, the classification rates are almost the same in both configurations, indicating that predictive performance did not depend exclusively on the *Sex* variable. Indeed, there are some cases where the performance is higher when such a variable is not available.

Schema	Features	QI	QII	QIII
Classical Psychometric	*All*	0.648	0.531	0.677
	W/O Sex	0.553	0.626	0.664
Quartile-based	*All*	0.630	0.620	0.601
	W/O Sex	0.653	0.599	0.538
Median-based	*All*	0.742	0.799	0.712
	W/O Sex	0.714	0.885	0.686
Fixed-threshold	*All*	0.756	0.763	0.825
	W/O Sex	0.757	0.757	0.826

Table 13: Comparison of the best dimension-level Macro-F1 values with all features and after excluding Sex.

5 *Discussion*

This study introduced a speech-based computational framework for estimating children's personality traits through a detailed characterization of paralinguistic information. In addition, we systematically explored multiple strategies for transforming the continuous CPQ scores into discrete categories, enabling the formulation of the task as a multi-class classification problem across different CPQ-based labeling schemes. Below, we summarize the key findings and methodological implications.

Transforming continuous personality scores into discrete categories inherently risks introducing boundary biases, where artificial cutoffs force acoustically similar speech profiles into different classes. By evaluating four distinct discretization criteria, this study mitigated such biases and revealed how the choice of threshold directly impacts classifier performance. Binary schemes, particularly the fixed-threshold and median-based approaches, outperformed multiclass setups in both accuracy and F1. This effect occurs because binary discretizations minimize the boundary overlap between intermediate profiles, maximizing the acoustic separability between opposite poles of each CPQ dimension. Conversely, multiclass setups suffered more from border effects and class imbalance, yielding lower average F1 scores.

Evaluating multiple ways of discretizing CPQ scores enables the determination of which configuration is more suitable, depending on the needs of professionals who may use an automatic personality estimation system. For instance, binary schemes offer higher precision and clearer separation between opposite poles of each CPQ dimension, making them appropriate for rapid screening tools or educational scenarios where a stable and easily interpretable decision is required. In contrast, three-level schemes allow a more nuanced representation of intermediate profiles, even if this comes at the cost of slightly reduced accuracy, which may be preferable in clinical or research contexts that benefit from capturing subtle variations in socio-emotional characteristics. Overall, these findings suggest that the choice of labeling scheme is functional: binary schemes may be preferable for screening or rapid decision-support scenarios, whereas multi-level schemes may be more appropriate when intermediate personality profiles are relevant. However, in educational or child-evaluation contexts, model outputs should be used only as decision-support information and complemented with other psychometric instruments and professional judgment to reduce the risk of bias or incorrect personalization.

Across CPQ dimensions, clearer acoustic separability emerged in **QII** and **QIII**, particularly under binary schemes. However, in **QIII** under the fixed-threshold configuration, the strong dominance of the *tough-minded* class indicates that performance should be interpreted with caution. Because the **QIII** minority class contained very few children, results may be strongly influenced by the majority class and should not be interpreted as balanced performance. The corresponding macro-F1 provides a more realistic estimate of minority-class behavior and is therefore more informative than accuracy in this imbalanced setting. This discrepancy between accuracy and macro-F1 highlights the importance of complementary metrics when evaluating personality classification under imbalanced distributions.

6 Conclusions

In this paper, we introduced an approach for predicting CPQ profiles in children through speech. The obtained results demonstrated the feasibility of inferring children's personality dimensions directly from their speech in educational contexts. A rigorous subject-independent validation scheme was applied to ensure a robust evaluation and prevent any form of information leakage. Four labeling schemes and six classifiers were systematically compared to assess their robustness and generalization capacity. Across the evaluated models, LR and SVM consistently offered the best trade-off between accuracy, interpretability, and generalization. Furthermore, the results confirmed that a compact subset of features (7–14), including MFCCs, pitch, jitter, shimmer, RMS, ZCR, and spectral entropy, was sufficient for effective prediction, supporting the existence of a concise prosodic footprint underlying children's personality expression. Future work will focus on expanding and diversifying the corpus to improve representativeness and allow evaluating self-supervised speech representations, and testing cross-corpus generalization with other corpora in the literature using CPQ-based personality assessment.

Acknowledgments

This research was supported by the Secretariat of Science, Humanities, Technology, and Innovation (SECIHTI) through the Scholarship No. 4024878 awarded to the first author. The authors thank the institutions involved in both the ethical approval and data collection procedures that made the CHERISH corpus possible.

References

Abdul, Z. K. and A. K. Al-Talabani. 2022. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158.

Aishwarya, N. and M. N. Veena. 2023. Ai based automatic personality assessment using speech analysis. *International Advanced Research Journal in Science, Engineering and Technology*, 10(7):733–736.

Andrade, C. 2021. Z scores, standard scores, and composite test scores explained. *Indian Journal of Psychological Medicine*, 43(6):555–557.

Baevski, A., Y. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Banik, A., S. Arya, and A. Kant. 2015. Vocal parameters in children between 4 to 12 years of age: An attempt to establish a prototype database. *International Journal of Scientific and Research Publications (IJSRP)*, 5(11), November.

Bhangale, K. and M. Kothandaraman. 2023. Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. *Electronics*, 12(4):839.

Bhin, H. and J. Choi. 2025. Multimodal personality recognition using self-attention-based fusion of audio, visual, and text features. *Electronics*, 14(14):2837.

Eyben, F., K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Gahona Castillejos, L., D. I. Hernández Farías, and H. Pérez-Espinosa. 2025. Cherish: A corpus for children emotion recognition in speech. *Procesamiento del Lenguaje Natural*, 75(0):223–236.

Hammoud, M., M. N. Getahun, A. Baldycheva, and A. Somov. 2024. Machine learning-based infant crying interpretation. *Frontiers in Artificial Intelligence*, 7:1337356.

International Telecommunication Union (ITU). 2011. Objective measurement of active speech level, dec. Recommendation ITU-T P.56 (12/2011), Series P: Terminals and Subjective and Objective Assessment Methods, Objective Measuring Apparatus, Fifth Edition.

Lee, S., A. Potamianos, and S. Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Lukac, M. 2024. Speech-based personality prediction using deep learning with acoustic and linguistic embeddings. *Scientific Reports*, 14(1):30149.

Mahaffey, B. L., D. Watson, L. A. Clark, and R. Kotov. 2016. Clinical and personality traits in emotional disorders: Evidence of a common framework. *Journal of Abnormal Psychology*, 125(6):758–767.

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.

Melzer, P. and M. Schoop. 2015. Personality-based versus task-based factors as indicators for personalised learning environments. In *UK Academy for Information Systems Conference Proceedings 2015*, page 21.

Mihalache, S. and D. Burileanu. 2022. Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection. *Sensors*, 22(3):1228.

Milling, M., A. Baird, K. D. Bartl-Pokorny, S. Liu, A. M. Alcorn, J. Shen, T. Tavassoli, E. Ainger, E. Pellicano, M. Pantic, N. Cummins, and B. W. Schuller. 2022. Evaluating the impact of voice activity detection on speech emotion recognition for autistic children. *Frontiers in Computer Science*, 4:837269.

Nercelles, L. and D. Centeno. 2020. Muda vocal: el proceso de cambio de la voz durante la pubertad. *Revista Mexicana de Pediatría*, 87(4):153–157.

Porter, R. B. and R. B. Cattell. 2016. *CPQ: Cuestionario de Personalidad para Niños. Manual.* TEA Ediciones, S.A.U., Madrid, 9 edition.

Pérez-Espinosa, H., B. Gutiérrez-Serafín, J. Martínez-Miranda, and I. E. Espinosa-Curiel. 2022. Automatic children's personality assessment from emotional speech. *Expert Systems with Applications*, 187:115885.

Rodríguez, A. 2017. Datos normativos de la voz. Serie creación n° 38, Facultad de Ciencias de la Salud, Escuela de Fonoaudiología, Centro de Investigación en Educación Superior CIES - USS.

Rubio Franco, V. J. 2020. Uso de la inteligencia artificial para la estimación de la personalidad: La conducta expresiva como forma de hacer predicciones sobre estados y características internas del individuo. *Reports de inteligencia económica y relaciones internacionales*, (5):1–11.

Schuller, B., S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll. 2005. Speaker independent speech emotion recognition by ensemble classification. In *2005 IEEE International Conference on Multimedia and Expo*, pages 864–867.

Schuller, B., S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss. 2012. The INTERSPEECH 2012 speaker trait challenge. In *13th Annual Conference of the International Speech Communication Association 2012 (INTERSPEECH 2012)*, volume 1, pages 254–257. Curran.

Slobodskaya, H. R. and O. S. Kornienko. 2021. Age and gender differences in personality traits from early childhood through adolescence. *Journal of Personality*, 89(5):933–950.

Solera-Ureña, R., H. Moniz, F. Batista, R. Fernández Astudillo, J. Campos, A. Paiva, and I. Trancoso. 2016. Acoustic-prosodic automatic personality trait assessment for adults and children. In *Proceedings of IberSPEECH 2016*, pages 192–201.

Tracey, B., D. Volfson, J. Glass, R. Haulcy, M. Kostrzebski, J. Adams, T. Kangarloo, A. Brodtmann, E. R. Dorsey, and A. Vogel. 2023. Towards interpretable speech biomarkers: exploring mfccs. *Scientific Reports*, 13(1):22787.

Vorperian, H. K., R. D. Kent, M. J. Lindstrom, C. M. Kalina, L. R. Gentry, and B. S. Yandell. 2005. Development of vocal tract length during early childhood: A magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 117(1):338–350.

Wartberg, L., K. Potzel, C. Spindler, and R. Kammerl. 2023. The big five personality domains and their facets: Empirical relations to problematic use of video games, social media and alcohol. *Behavioral Sciences*, 13(6):444.

Wei, X.-S., C.-L. Zhang, H. Zhang, and J. Wu. 2018. Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, 9(3):303–315.

Wojciechowska, I., R. Matkowski, and T. Pawłowski. 2022. Type d personality and big five personality traits and the risk of breast cancer: A case-control study. *Frontiers in Psychiatry*, 13:723795.

Zainudin, N. S., K.-H. Ng, and K.-C. Khor. 2021. Identifying the important demographic and financial factors related to the mortality rate of covid-19 with data mining techniques. In A. Mohamed, B. W. Yap, J. M. Zain, and M. W. Berry, editors, *Soft Computing in Data Science*, volume 1428 of *Communications in Computer and Information Science*, pages 241–253, Singapore. Springer Singapore.

Zhang, Y. and Y. Gu. 2024. Individual differences in multimodal child-directed language: Unraveling individual style, empathy and the big five personality traits. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, pages 5966–5974, Rotterdam, Netherlands. University of California.