





Article

Prosodic Feature Analysis for Automatic Speech Assessment and Individual Report Generation in People with Down Syndrome

Mario Corrales-Astorgano ^{*}, César González-Ferreras , David Escudero-Mancebo 
and Valentín Cardeñoso-Payo 

Research Group ECA-SIMM, Computer Science Department, University of Valladolid, 47002 Valladolid, Spain; cesargf@infor.uva.es (C.G.-F.); descuder@infor.uva.es (D.E.-M.); valentin.cardenoso@uva.es (V.C.-P)

* Correspondence: mcorrales@infor.uva.es

Abstract: Evaluating prosodic quality poses unique challenges due to the intricate nature of prosody, which encompasses multiple form–function profiles. These challenges are more pronounced when analyzing the voices of individuals with Down syndrome (DS) due to increased variability. This paper introduces a procedure for selecting informative prosodic features based on both the disparity between human-rated DS productions and their divergence from the productions of typical users, utilizing a corpus constructed through a video game. Individual reports of five speakers with DS are created by comparing the selected features of each user with recordings of individuals without intellectual disabilities. The acquired features primarily relate to the temporal domain, reducing dependence on pitch detection algorithms, which encounter difficulties when dealing with pathological voices compared to typical ones. These individual reports can be instrumental in identifying specific issues for each speaker, assisting therapists in defining tailored training sessions based on the speaker’s profile.

Keywords: Down syndrome; automatic classification; prosody



Citation: Corrales-Astorgano, M.; González-Ferreras, C.; Escudero-Mancebo, D.; Cardeñoso-Payo, V. Prosodic Feature Analysis for Automatic Speech Assessment and Individual Report Generation in People with Down Syndrome. *Appl. Sci.* **2024**, *14*, 293. <https://doi.org/10.3390/app14010293>

Academic Editor: Andrea Prati

Received: 16 November 2023

Revised: 21 December 2023

Accepted: 27 December 2023

Published: 28 December 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Prosody plays a crucial role in speech communication, as it encompasses various essential functions, including grouping linguistic units, pausing, word accentuation, and indicating sentence purposes (such as interrogative, imperative, declarative, or exclamatory), as well as conveying emotions and pragmatics [1]. In the field of automatic speech processing, prosody has been examined through various types of predictor variables, including high-performance power features and leverage features [2].

Inadequate control or improper production of prosody can lead to social stigmatization and limit individuals’ integration into society [3]. This situation may apply to individuals with intellectual disabilities in general and specifically to those with Down syndrome (DS), as they often face challenges in language control and prosodic expression, with some exceptions [4–6]. Regarding prosody, Kent and Vorperian [5] have observed disfluencies such as stuttering and cluttering, as well as difficulties in perceiving, imitating, and spontaneously producing prosodic features. Additionally, Heselwood et al. [7] have linked certain speech errors to struggles in identifying boundaries between words and sentences. Other work showed that individuals with Down syndrome (DS) may exhibit mixed patterns of speech production, potentially impacting speech intelligibility [8], using perceptual and acoustic evaluation. In a previous work, we conducted perceptual and automatic identification tests based on signal analysis, finding differences between the voices of individuals with DS and typical speakers [9].

Several learning games and software tools have been developed specifically for individuals with intellectual disabilities to enhance specific skills [10–13]. Voice therapists employ methods to address speech difficulties in individuals with specific speech problems [14]. Some of these methods have been partially implemented as software tools,

servicing as aids for therapists to work with their patients or enabling patients to engage in supplementary exercises independently [15]. In our previous work [16], we introduced a tool designed to train prosody and pragmatics in individuals with Down syndrome (DS). This tool incorporates a combination of perceptual and production activities within a graphic adventure video game, featuring an adapted interface that considers the unique characteristics of individuals with Down syndrome, such as poor short-term memory [4], attention deficits [17], information integration challenges, and language development deficits [18]. Thus far, the video game has been successfully used by real users, with the support of an adult, such as a therapist, teacher, or family member. The tool has facilitated the collection of a speech corpus consisting of individuals with Down syndrome. The main objective of the research described in this paper is to analyze the potential of these recordings for training an automatic assessment system. In the near future, this system will be incorporated into the video game, allowing users to train independently. A detailed description of the oral activities can be found in [19].

In the existing literature, there have been several studies on the automated evaluation of speech quality in atypical voices [20–22]. Nevertheless, in computer-assisted pronunciation training, the goal is not just to assess but also to provide information about the underlying reasons behind judgments, whether correct or incorrect. In [8] authors investigated the influence of various components of speech production on speech intelligibility in individuals with DS. In addition, the speech disorders in DS speakers are not uniform, emphasizing the need to consider and address individual variations when developing treatment approaches [23].

In a related study [24], we examined the feasibility of evaluating the oral productions of individuals with DS in terms of quality. We achieved over 90% accuracy in identifying DS speech using an SVM classifier [9]. However, when it came to assessing the quality of the oral productions, we only achieved approximately 78.5% accuracy using the same training feature set and classifier. The evaluation focused mainly on evaluating prosodic quality, covering factors such as intonation, accent, and phrasing. The assessment of these aspects was conducted at two levels: correct or incorrect. In this paper, we delve into an analysis of the training data used in our previous studies to gain insights into the reasons behind this discrepancy in classification performance. Our findings not only shed light on the factors contributing to these performance differences but also provide valuable indications for exploring alternative approaches in future research endeavors.

In this paper, we undertake a systematic analysis of the prosodic features present in the utterances of the corpus to identify the most informative features and their corresponding values for predicting the quality of oral productions. To achieve this, we pose the following research questions:

- RQ1: Is it possible to select prosodic features by integrating information about prosodic quality and differences between types of speakers?
- RQ2: Can specific issues of a user be identified using the extracted features?

To answer these questions, we introduce a novel feature selection procedure that integrates information derived from the distinctions between correctly and incorrectly pronounced DS utterances, based on evaluations performed by human experts, and the differences between the utterances produced by DS speakers and those of typical speakers (further elaborated in Section 2.3). As a result, the significance of the selected features becomes more straightforward to convey when explaining potential limitations of the specific speakers under examination (as presented in Section 3 and discussed in Section 4).

The paper is organized as follows. The materials and methods section provides detailed information about the compiled corpus and the manual evaluation of the utterances. Additionally, it presents the procedure used for the individual analysis of various prosodic features. The results section presents the selected prosodic features and their effectiveness in modeling the quality of utterances, considering human scores as the basis for evaluation. Finally, the discussion section encompasses an examination of limitations

encountered during the study, proposes potential avenues for future research, and draws overall conclusions.

2. Materials and Methods

Figure 1 shows the scheme of the experimental procedure followed in this work. The recordings of speakers with Down syndrome are collected while users play a learning game to improve their pronunciation abilities following indications of an avatar inside the game and under the supervision of a therapist. A simplified version of the video game is used to record a mirror corpus of typical development (TD) speakers, as described in Section 2.1. The recordings of the speakers with DS are evaluated offline by a prosodic expert, as described in Section 2.2. A set of prosodic features are extracted from audio files and an original procedure for selecting the most relevant ones is applied, as described in Section 2.3. Finally, we use reference intervals and confidence intervals of the selected variables in typical speakers to build personalized reports about the particular deficits of the speakers with DS when producing the activities, as described in Section 2.5.

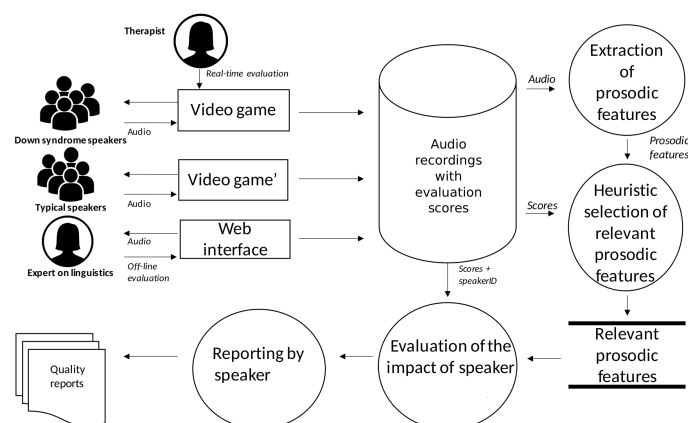


Figure 1. Experimental procedure diagram.

2.1. Corpus Recording

The recorded data for this study were obtained from a graphic adventure video game [16] that engages users in activities designed to help improve perception and production of prosodic cues. Users are required to perform these activities in order to progress in the game. These activities target various language functions such as questioning, expressing opinions, and social interaction. Additionally, different prosodic functions, such as chunking or prominence, are employed in different production modes such as reading, elicited speech, or free speech. All oral productions and user interactions were recorded and categorized based on the specific activity and speaker during gameplay. The corpus was compiled from various software testing and user training sessions. It contains 2151 recordings of people with Down syndrome, produced by 42 different speakers. In addition, the corpus contains recordings of the game recorded by 30 individuals without Down syndrome, totaling 1589 utterances. In order to diminish background noise during the recording phase, the performers utilized a headset equipped with an integrated microphone (specifically, the Plantronics USB headset). This headset recorded audio in MS-WAVE PCM format at a frame rate of 44,100 Hz, with 16 bits per sample, and in a mono configuration. The incorrect or highly noisy audio files were removed, but no further processing was conducted on them. Additional information can be found in [19].

Out of the 966 evaluated recordings (see Section 2.2), 605 belong to the same 5 users, and these users had more than 40 utterances each. These 5 speakers provide enough information about the potential of the proposed methodology and facilitate an individualized analysis of each one. Detailed information regarding the speakers' gender, age, and cognitive abilities can be found in Table 1. To obtain this information about the speakers' profile, a Spanish version PEPS-C test [25] was used. This test aims to evaluate prosodic compe-

tence across multiple dimensions. It assesses both the form, which includes perceptual and motor abilities, and the function, which encompasses cognitive comprehension and expressive capabilities. Additionally, it evaluates both receptive (input) and expressive (output) skills. This corpus is referred to as θ_{DS} . To compare the voices of individuals with Down syndrome to those of typical speakers, we used the subset of the corpus with recordings from typical speakers. We named this corpus θ_{TD} .

Table 1. Description of the informants analyzed in this study. The table displays the following information for each speaker: chronological age (CA), verbal mental age (VA), short-term verbal memory (STVM), and non-verbal cognitive level (NVCL). The ages are denoted in months. Furthermore, it includes the mean percentage of success in perception (MPercT) and production (MProdT) for PEPS-C tasks. The table is completed with the percentage of sentences classified as right in the offline evaluation (OLRight) and the total number of audio recordings of each speaker (#Audios).

Speaker	Gender	CA	VA	STVM	NVCL	MPercT	MProdT	OLRight	#Audios
022	f	195	84	94	17	69.79%	48.30%	72%	120
023	m	204	99	134	18	76.04%	72.10%	76%	106
024	f	178	96	78	20	73.96%	74.65%	80%	97
025	m	190	60	below 74	10	60.42%	49.76%	57%	131
026	m	223	69	below 74	13	56.25%	45.70%	51%	151

2.2. Human-Based Quality Evaluation

The samples collected in the gaming sessions have been evaluated by four different annotators at different stages of the project [19]. Two of the assessments were made during the gaming process, another was carried out by a prosody expert outside the gaming environment, and the last was made by an automatic classifier. The perceptual evaluation of audio quality is always a problematic issue due to the high degree of disagreement among annotators [26–28]. In our case, the task is particularly difficult due to speech characteristics and the fact that evaluators tend to consider the speakers' abilities (especially in online judgments). In order not to add noisy information in the models' training, we used the annotations from the prosody expert, to whom we assign a reference value, since they were involved in the game design.

The evaluations were conducted using a binary judgment, indicating whether the utterance should have been repeated or not. The expert followed specific decision criteria tailored to the proposed activity. The assessment depended solely on subjective judgments without employing any acoustic analysis of the sentences. Additionally, potential shortcomings in the pronunciation of individual sounds (segmental component) were not taken into account. The evaluation process involved playing audio recordings consecutively, allowing the expert to listen to each utterance multiple times before delivering their judgment. Even when dealing with speakers exhibiting significant issues in clarity, the primary criterion was whether the intonation closely resembled the anticipated one. These criteria included assessing the conformity to the expected intonation pattern, maintaining the distinction between stressed and unstressed syllables for lexical stress, differentiating accented and unaccented syllables for accentuation, and ensuring appropriate organization into prosodic groups while distinguishing between function and content words for phrasing. Based on these assessments, the corpus θ_{DS} was divided into two subsets: θ_R , denoting correct productions, and θ_W , denoting incorrect productions.

2.3. Processing and Selection of Prosodic Features

The acoustic features from each recording in the corpus were extracted using the openSmile toolkit [29]. For feature extraction, the GeMAPS feature set [30] was chosen due to its comprehensive collection of acoustic and prosodic features. This feature set was used in automatic emotion detection [31] and in our previous works [9]. GeMAPS encompasses frequency-related features, energy-related features, and temporal features. To capture

the variation along the utterance, the arithmetic mean and coefficient of variation were calculated for each feature.

Additionally, four extra temporal features were included: silence percentage, sounding percentage, silences per second, and the mean length of silences. These additional features were derived using the silences and sounding intervals identified by the Praat software (2006) [32], which utilizes an intensity threshold along with minimum silent and sounding interval durations to detect these intervals (default values were used). In total, a set of 92 features was employed, including 10 from the frequency domain, 10 from the energy domain, 11 from the temporal domain, and 61 from the spectral domain. A detailed description of these features can be found in [9].

U is the set of utterances in the corpus, with $U = \theta_{TD} \cup \theta_R \cup \theta_W$, where θ_{TD} are the utterances of typical speakers and θ_R or θ_W the utterances of speakers with Down syndrome, judged as right or wrong by the evaluator. Every $u \in U$ is characterized as $u = (f_1, f_2, f_3, \dots, f_N)$, where f_i is a feature computed in the previous stage. The distribution of values of the feature f_i in a subset s of U is referred to as $pdf(s, f_i)$, and its corresponding mean as $\mu(s, f_i)$, where s could be the samples of a single speaker or the ones of a subset θ_{TD} , θ_W , or θ_R . Algorithm 1 describes the process of selecting the most appropriate features for visualizing the particular problems of the speakers. For a feature f to be included in the set of selected features, two conditions should be fulfilled:

1. Separation: Statistical analysis using the Mann–Whitney test with a p-value threshold of less than 0.01 is applied to determine if there are significant differences between the values of f in the groups θ_R (right utterances) and θ_W (wrong utterances). This criterion ensures that clear distinctions between right and wrong utterances are observed.
2. Consistency: Let $\mu(\theta_{TD}, f)$ represent the mean value of feature f in the group θ_{TD} (typical speakers), $\mu(\theta_R, f)$ represent the mean value in the group θ_R , and $\mu(\theta_W, f)$ represent the mean value in the group θ_W . For a feature to be selected, it must satisfy the condition $|\mu(\theta_{TD}, f) - \mu(\theta_R, f)| < |\mu(\theta_{TD}, f) - \mu(\theta_W, f)|$. This criterion ensures that right utterances are closer to the typical speakers' feature value than wrong utterances.

Algorithm 1: Selection of the most appropriate features for visualizing specific problems of the speakers. The sigDiff function determines whether there are statistically significant differences between two given distributions.

```

input :U the set of utterances
output:sF the selected features
F ← {f1, f2, f3, . . . fN}
sF ← ∅;
// Separation
forall f in F do
  if sigDiff(pdf(θR, f), pdf(θW, f)) then
    | sF ← sF ∪ {f};
  end
end
// Consistency
forall f in sF do
  if |μ(θTD, f) − μ(θR, f)| ≥ |μ(θTD, f) − μ(θW, f)| then
    | sF ← sF − {f};
  end
end
return sF

```

2.4. Automatic Classification

In order to demonstrate the validity of the feature selection procedure, we conducted some experiments to check the classification power of the selected features to predict prosodic quality of the 605 selected recordings. We used the Weka machine learning toolkit [33]. To compare their performance, we employed three distinct classifiers that have shown success in similar classification tasks in our previous work: the C4.5 decision tree (DT), the multilayer perceptron (MLP), and the support vector machine (SVM). Given the size of the dataset, we used the default hyperparameters provided by the Weka software Version 3.8.6. The DT employed hyperparameters with a confidence threshold for pruning set at 0.25 and a minimum number of instances per leaf set at 2. For the SVM, a normalized polynomial kernel with an exponent of 2 was utilized. The complexity constant (C) was configured to 1, the tolerance parameter was established at 0.001, and the epsilon for round-off error was defined as 1.0×10^{-12} . Regarding the MLP architecture, it featured a single hidden layer with a number of neurons determined by the sum of input features and output classes divided by 2. The learning rate was set to 0.3, the momentum rate to 0.2, and the training proceeded for 500 epochs. To construct the training and testing sets, we implemented the leave-one-speaker-out cross-validation technique. The idea is to leave out one entire speaker's data as the test set in each iteration while training the model on the remaining data. This process is repeated for each speaker in the dataset. These classifiers used the label defined by the expert for the assessment of speech quality. In addition to verifying the effectiveness of the selected features in classifying the quality of the recordings, the use of automatic classifiers is justified by the need to integrate an automatic feedback system for players and therapists into the video game. This enables autonomous gaming and the potential correction of interventions after receiving such feedback.

In this study, we also used three established feature selection methods in order to compare the classifier performance with our proposed method. The number of features selected is set to the same number obtained with the proposed method. Firstly, we employed information gain with respect to the class for ranking the features and selected the features with higher scores. Additionally, we utilized forward feature selection (FFS) with correlation-based feature subset selection [34] to identify the most relevant features. This method assesses the worth of a subset of features that exhibits high correlation with the class while maintaining low intercorrelation among the selected features. Furthermore, we employed forward feature selection (FFS) to select the most relevant features using the accuracy of a DT classifier as the evaluation metric.

2.5. Generation of Personal Reports

A report with information about the individual performance of each of the speakers is generated. This report permits a comparison of the values of the relevant prosodic features of the DS speakers with respect to the variability of the same features in typical speakers. These individualized reports can be useful for identifying specific issues with a speaker's use of prosody and enable the adaptation of training activities to the speaker's prosodic skills. The reports are represented by radar charts drawn using the Python library matplotlib [35]. These charts have been used in other works to compare the speech of typical speakers and speakers with Parkinson's [36]. These individualized reports can be useful for identifying specific issues with a speaker's use of prosody. They enable the adaptation of training activities to the speaker's prosodic skills.

From each selected feature, two different intervals have been calculated: reference interval, defined as the interval in which 95% of the values of a reference population (TD) fall, and the confidence interval (95%), which is a range of plausible values for the population mean. The reference intervals were calculated using the `refLimits` function (using the Cook's outlier detection method) from R `referenceIntervals` library [37] and the confidence intervals were computed using the Python library `scipy` [38]. These charts show the reference and the confidence intervals computed with all the recordings of TD speakers and the confidence interval of each speaker with DS for each of the selected

features and type of evaluation. The minimum and maximum values of all the intervals were calculated to define the scale of the charts, adding a 0.1 margin to the lower limit to facilitate the visualization.

3. Results

Table 2 shows the 95% confidence intervals of the mean values for the selected features, categorized by groups. Out of the ninety-two input features analyzed, only seven features meet the established criteria using the offline evaluation data. In previous studies, using different criteria, 21 features were selected in [24] and 27 features were selected in [9].

Table 2. List of automatically chosen frequency, energy, and temporal features. All of these features exhibit statistically significant differences (as determined by the Mann–Whitney test with a p -value less than 0.01) when comparing correct and incorrect productions of people with Down syndrome. Moreover, in the selected features, the difference between the mean of productions of TD speakers and the mean of right productions is lower than the difference between the mean of productions of TD speakers and the mean of wrong productions. The interpretation of these features is detailed in [9]. Within the cells, we display the 95% confidence interval of the mean value, with the units specified in [29].

	Typical Speakers	DS Right Productions	DS Wrong Productions
F0 domain			
f1 jitterLocal_sma3nz_stddevNorm	(1.15, 1.18)	(1.32, 1.43)	(1.52, 1.66)
Energy domain			
e1 loudness_sma3_percentile20.0	(0.85, 0.89)	(0.71, 0.78)	(0.63, 0.71)
Temporal domain			
d1 loudnessPeaksPerSec	(5.74, 5.84)	(4.10, 4.32)	(3.77, 4.06)
d2 StddevVoicedSegmentLengthSec	(0.18, 0.19)	(0.18, 0.23)	(0.25, 0.33)
d3 silencePercentage	(0.11, 0.13)	(0.08, 0.11)	(0.21, 0.27)
d4 silencesPerSecond	(0.37, 0.41)	(0.27, 0.35)	(0.52, 0.63)
d5 silencesMean	(0.20, 0.22)	(0.13, 0.18)	(0.31, 0.41)

Regarding the selected features, jitterLocal_sma3nz_stddevNorm represents the coefficient of variation of deviations in consecutive F0 period lengths. In the energy domain, loudness_sma3_percentile20.0 indicates the 20th percentile estimate of perceived signal intensity from an auditory spectrum. Finally, in the temporal domain, loudnessPeaksPerSec refers to the number of loudness peaks per second, StddevVoicedSegmentLengthSec represents the standard deviation of continuously voiced regions, silencePercentage indicates the duration percentage of unvoiced regions, silencesPerSecond represents the number of silences per second, and silencesMean denotes the mean length of unvoiced regions.

The confidence interval values in Table 2 reveal that a less stable F0 contour (higher f1 feature) is associated with an abnormal utterance pronunciation at a perceptual level. A weaker intensity (lower e1 feature) is penalized more. Utterances belonging to wrong groups exhibit slower speech (lower d1 feature), more speed changes (higher d2 feature), increased frequency of inner pauses (higher d3 and d4 features), and longer pauses (higher d5 feature). Furthermore, Table 2 allows a comparison of feature values between typical and DS speakers. It is relevant to note that there is a separation in the intervals between typical and right productions for all features, except for d2, where the confidence intervals overlap between right and typical utterances.

The proposed feature selection procedure successfully pinpointed the seven most informative variables out of the ninety-two analyzed ones, enabling the prediction of prosodic quality with satisfactory outcomes. Three other established feature selection methods have also been used for comparison. As shown in Table 3, the feature set obtained with the proposed method, comprising seven features, yields superior classification accuracy compared to

using all ninety-two features across the three examined classification methods. This indicates that these seven features encapsulate essential aspects for effectively classifying samples. Compared to other feature sets of equal size (seven features) acquired through conventional feature selection methods like information gain, correlation-based FFS, or DT classifier FFS techniques, the outcomes are similar. Although the features extracted using the proposed method yield slightly lower accuracy in the case of DT and SVM, they demonstrate higher accuracy in the case of MLP compared to feature sets obtained through traditional methods. Regarding the F1 score, the optimal result is achieved with our proposed feature selection method and an MLP classifier. As a result, we believe that employing these features, carefully chosen to balance the discriminative capabilities between typical users and users with Down syndrome, while also considering the ability to differentiate between correct and incorrect pronunciation among users with Down syndrome, achieves this dual objective without substantial loss in representation capacity for classification tasks, compared to other techniques that exclusively consider the data of users with Down syndrome.

Table 3. Results of prosodic quality classification using different sets of features and classifiers. The accuracy (Acc) and F1 score (F1) are reported for decision trees (DT), support vector machines (SVM), and multilayer perceptron (MLP). The number of features in each set is shown in brackets.

	DT		SVM		MLP	
	Acc	F1	Acc	F1	Acc	F1
All Features (92)	63.5%	0.628	67.3%	0.617	62.8%	0.609
Proposed Method (7)	64.6%	0.603	68.9%	0.650	69.4%	0.681
Information Gain (7)	69.1%	0.653	71.6%	0.661	68.3%	0.650
Correlation FFS (7)	69.3%	0.653	70.9%	0.648	68.8%	0.665
Classifier DT FFS (7)	67.5%	0.662	67.5%	0.625	66.3%	0.641

Figure 2 presents the report individualized by speaker (the radar chart) of the potential anomalous use of the selected prosodic features. The radar charts present reference and confidence intervals of the TD speakers. The polygons in blue represent the upper and lower limits of the reference intervals and the black polygons correspond to the confidence intervals. The charts also present the confidence interval of the five speakers with DS for each type of evaluation (right or wrong). These confidence intervals are represented by polygons in red. When the red polygon coincides with the black one, it means that the features of the DS speaker are in the same range of the TD speakers. The figure shows that the differences between right DS recordings and recordings of TD speakers are lower than the differences between wrong recordings in speakers with DS and TD speakers.

Speaker 022 has higher values in f1 feature, lower values in e1 feature, and similar values in all temporal features except in d1, than TD speakers when recordings are evaluated as right. When recordings are evaluated as wrong, this speaker presents higher values in all features except e1 and d3, with more variability than TD speakers. In this speaker, the number of inner pauses and their length (d3, d4, and d5 features) and the frequency variation (f1) seem to be the most informative features to evaluate their recordings as right or wrong. Speaker 023 has similar values in all features, except f1, d1, and d2, to TD speakers when recordings are evaluated as right. When recordings are evaluated as wrong, this speaker presents higher values in all features, except in d1 and d5 features, than TD speakers. The wrong recordings of speaker 023 show high variability of f0 contour (f1), more speed changes (d2), and more inner pauses (d3 and d4) than the right recordings.

Speaker 024 presents lower values in d1, d3, d4, and d5 when recordings are evaluated as right. This speaker has higher variability in all features, except e1 and d1, than TD speakers in wrong recordings. This speaker has higher variability in all features in wrong recordings than in right recordings. Speaker 025 has similar values in all features, except d1 and d2, to TD speakers in right recordings and higher values in all features, except d1 and e1, than TD speakers in wrong recordings. The number of inner pauses and the pauses' length are higher in wrong recordings than in right recordings.



Figure 2. Radar charts of the reference (RI) and confidence (CI) intervals of the TD speakers and confidence interval of the five speakers with DS for each type of evaluation (right or wrong). The polygons in blue represent the upper and lower limits of the reference intervals, the polygons in black correspond to the confidence intervals of the features extracted from the audio recordings of TD speakers, and the polygon in red corresponds to values of the confidence intervals obtained from the speaker with DS.

Finally, speaker 026 has higher values in e1 and f1 than TD speakers in right recordings and higher values in f1, d2, d3, d4, and d5 features than TD speakers in wrong recordings. As well as speakers 022 and 024, the number of inner pauses and the pauses' length are higher in wrong recordings than in right recordings.

4. Discussion

The temporal domain's acoustic features appear to be more informative in assessing oral turns regardless of the speaker. Out of the seven selected features listed in Table 2, four variables specifically pertain to temporal aspects. This outcome is promising because computing such features, unlike those in the F0 or spectral domain, demonstrates greater resilience against adverse conditions that may arise with users affected by DS. Generally, pitch detection algorithms encounter more challenges when dealing with pathological voices compared to typical voices [39]. Moreover, features related to the temporal domain can be readily associated with disfluent speech, such as stuttering or cluttering, which, while not universal, are common issues among this population [40–42].

The presentation of individual reports about oral production proficiency of the speakers has proven to be necessary to be accurate in the judgments concerning the particular deficits of the speakers. We have shown in Figure 2 that presenting the personal results permits us to make a diagnosis that takes into account the important inter-speaker differences. Presenting results that take into account the different prosodic features is especially important in this context, as the subjective evaluation could be dependent on many aspects.

The presented reports permit us to shed light on the particular features that could most influence therapists' decisions when they score the quality of DS utterances. We already highlighted in previous works the high variability among DS individuals [24]. The five speakers exhibit different patterns with clear differences among them; the shapes of the red areas in the graphics are different among them and contrast significantly with the one of the typical speakers. As expected, the graphics corresponding to utterances labeled as wrong are more distant to the mean TD pattern and closer to the borders of the reference interval.

All the speakers have lower values of the feature d1 (loudnessPeaksPerSec) than TD speakers, related with the speed of speech. They are clearly slower when compared to the typical values, close to the limit of the reference interval in all the cases except 024. Speakers 022 and 024 show lower values for the feature loudness_sma3_percentile20.0 (e1). We already pointed out that low speed and abnormal control of loudness are characteristic clues of these types of speakers, probably due to muscular hypotonia [9].

Abnormal higher values for jitter (f1) are observed in the cases of speakers 022 and 023 and in a lower degree in the case of 026. The fact that the differences are higher in the cases of utterances labeled as wrong by the therapist seems to indicate that these speakers exhibit a particular problem with the control of this feature. As the jitter has to do with changes in the dynamics of the F0 contour, these speakers could be evidencing an unsatisfactory control of pitch.

The graphs allow us to find a problematic use of speech rhythm in the case of speaker 023. The StddevVoicedSegmentLengthSec feature (d2) varies much more than typically. This speaker shows a problem of cluttered speech with unstable duration of the segments, providing a higher standard deviation of voiced segments lengths than in a typical voice. This behavior is also observed in speakers 025 and 026 but to a lesser degree.

The two speakers that were expected to have the worst results are 025 and 026 due to their lower CA and VA (see Table 1). Apart from the extreme values of d1 that has been already mentioned (they speak extremely slow), speaker 026 has the highest values for d3, d4, and d5 features, evidencing that this speaker interrupts himself very frequently, more when the utterances are labeled as wrong. In the case of speaker 025, wrong utterances are marked when there are more pauses as well (higher d5, d4, d3, and d2 values). Additionally the significant lower values for these features in the case of right utterances, in conjunction with the low control of the peaks of energy and high variability of voiced segments' length, evidence a pattern of abnormal articulation.

The information provided by prosodic features about specific individual speech anomalies could serve to select personalized training activities. For example, with a glance at the results in Figure 2, a therapist could decide to assign exercises related with the intonation to speaker 023. Although, in this study, we have focused on five speakers with Down syndrome, this type of analysis could be generalized to other populations because the recordings used as reference belong to individuals without any speech-related issues, and the extracted characteristics do not depend on a specific speaker profile. The use of the video game for recording new speech samples and the recording environment could be thought of as a restricted profile, but in an alternative one, the fundamental idea of using recordings of typical speakers as a reference for evaluating the quality of utterances with special characteristics remains.

As future work, we intend to conduct a comprehensive analysis of the interaction between features to identify dependencies within the features of the dataset. Also proposed as future work is the exploration of new features or combinations of existing ones that may be associated with specific activities or types of activities. This endeavor aims to enhance the results while also compiling a larger corpus to enable the testing of more sophisticated models or alternative machine learning techniques. The task at hand is challenging due to the diverse range of activities offered by the video game. In the future, we also intend to utilize the utterances from the corpus to construct an unsupervised classification of activities. This classification will consider the various acoustic prosodic features alongside human judgments of quality.

5. Conclusions

The paper presents a feature selection procedure that utilizes human-based evaluations and empirical observations to differentiate between utterances of individuals with Down syndrome and those with typical development. The human-based evaluations were conducted by a prosody expert, considering various aspects of prosody quality such as intonation, stress, accent, or organization into prosodic groups. Perceptual evaluation remains a significant challenge, but we have used it solely as a reference for quality. We are currently working on increasing the number of evaluators to enhance the evaluation process.

Acknowledging the limitations of the right/wrong evaluation, the results obtained can still provide valuable information to therapists about the prosodic quality of recordings from individuals with Down syndrome, using only seven prosodic features without compromising much on classification power. This procedure identifies the most significant features in each domain: temporal, frequency, and energy-related. The reduced number of features allows the creation of individual reports for each speaker, facilitating the identification of specific issues. The goal is to assist therapists in designing training therapies tailored to each user's particular problems.

The study's limitation lies in the restricted number of speakers used. However, collecting a voice corpus poses challenges, particularly when targeting populations with intellectual disabilities. Therefore, as part of future work, we consider it crucial to continue expanding the voice corpus by incorporating a more diverse range of users and recordings to achieve better generalization of results.

Author Contributions: Conceptualization, M.C.-A., D.E.-M., V.C.-P. and C.G.-F.; methodology, M.C.-A., D.E.-M., V.C.-P. and C.G.-F.; software, M.C.-A. and D.E.-M.; validation, D.E.-M., V.C.-P. and C.G.-F.; formal analysis, D.E.-M.; investigation, M.C.-A. and C.G.-F.; resources, M.C.-A., D.E.-M., V.C.-P. and C.G.-F.; data curation, M.C.-A.; writing—original draft preparation, M.C.-A.; writing—review and editing, M.C.-A., D.E.-M., V.C.-P. and C.G.-F.; visualization, M.C.-A.; supervision, D.E.-M.; project administration, V.C.-P. and C.G.-F.; funding acquisition, V.C.-P. and C.G.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out in the Project PID2021-126315OB-I00 that was supported by MCIN/AEI/10.13039/501100011033/FEDER/EU.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of the University of Valladolid (protocol code PI 20-1639 NO HCUV approved on 11 June 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the collected data consists of recordings from individuals with intellectual disabilities, and voice is a biometric data that can be used to identify a person, which goes against data protection laws.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DS	Down syndrome
F0	Fundamental frequency
TD	Typical development
DT	Decision tree
MLP	Multilayer perceptron
SVM	Support vector machine

References

- Roach, P. *English Phonetics and Phonology Fourth Edition: A Practical Course*; Cambridge University Press: Cambridge, UK, 2010.
- Batliner, A.; Möbius, B. Prosody in Automatic Speech Processing. In *The Oxford Handbook of Language Prosody*; Oxford University Press: Oxford, UK, 2020. <https://doi.org/10.1093/oxfordhb/9780198832232.013.42>.
- Wells, B.; Peppé, S.; Vance, M. Linguistic assessment of prosody. In *Linguistics in Clinical Practice*; Taylor & Francis: London, UK, 1995; pp. 234–265.
- Chapman, R.S.; Hesketh, L. Language, cognition, and short-term memory in individuals with Down syndrome. *Down Syndr. Res. Pract.* **2001**, *7*, 1–7.
- Kent, R.D.; Vorperian, H.K. Speech impairment in Down syndrome: A review. *J. Speech Lang. Hear. Res.* **2013**, *56*, 178–210.
- Stojanovik, V. Prosodic deficits in children with Down syndrome. *J. Neurolinguist.* **2011**, *24*, 145–155.
- Heselwood, B.; Bray, M.; Crookston, I. Juncture, rhythm and planning in the speech of an adult with Down's syndrome. *Clin. Linguist. Phon.* **1995**, *9*, 121–137.
- O'Leary, D.; Lee, A.; O'Toole, C.; Gibbon, F. Perceptual and acoustic evaluation of speech production in Down syndrome: A case series. *Clin. Linguist. Phon.* **2020**, *34*, 72–91.
- Corrales-Astorgano, M.; Escudero-Mancebo, D.; González-Ferreras, C. Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome. *Speech Commun.* **2018**, *99*, 90–100.
- Cano, A.R.; García-Tejedor, Á.J.; Alonso-Fernández, C.; Fernández-Manjón, B. Game Analytics Evidence-Based Evaluation of a Learning Game for Intellectual Disabled Users. *IEEE Access* **2019**, *7*, 123820–123829.
- García, L.E.; Mejía, R.J.; Salazar, A.; Gómez, C.E. Un videojuego para estimular habilidades matemáticas en personas con síndrome de Down. *Rev. Espac.* **2019**, *40*, 1–15.
- Prena, K.; Sherry, J.L. Parental perspectives on video game genre preferences and motivations of children with Down syndrome. *J. Enabling Technol.* **2018**, *12*, 1–9.
- Del Rio Guerra, M.S.; Martin-Gutierrez, J.; Acevedo, R.; Salinas, S. Hand Gestures in Virtual and Augmented 3D Environments for Down Syndrome Users. *Appl. Sci.* **2019**, *9*, 2641.
- Boone, D.R.; McFarlane, S.C.; Von Berg, S.L.; Zraick, R.I. *The Voice and Voice Therapy*; Pearson/Allyn & Bacon: Boston, MA, USA, 2005.
- Rodríguez, W.R.; Saz, O.; Lleida, E. A prelingual tool for the education of altered voices. *Speech Commun.* **2012**, *54*, 583–600.
- González-Ferreras, C.; Escudero-Mancebo, D.; Corrales-Astorgano, M.; Aguilar-Cuevas, L.; Flores-Lucas, V. Engaging adolescents with Down syndrome in an educational video game. *Int. J. Hum.-Interact.* **2017**, *33*, 693–712.
- Martínez, M.H.; Duran, X.P.; Navarro, J.N. Attention deficit disorder with or without hyperactivity or impulsivity in children with Down's syndrome. *Int. Med. Rev. Down Syndr.* **2011**, *15*, 18–22.
- Chapman, R.S. Language development in children and adolescents with Down syndrome. *Ment. Retard. Dev. Disabil. Res. Rev.* **1997**, *3*, 307–312.
- Escudero-Mancebo, D.; Corrales-Astorgano, M.; Cardeñoso-Payo, V.; Aguilar, L.; González-Ferreras, C.; Martínez-Castilla, P.; Flores-Lucas, V. PRAUTOCAL corpus: A corpus for the study of Down syndrome prosodic aspects. *Lang. Resour. Eval.* **2021**, *56*, 191–224.

20. Le, D.; Provost, E.M. Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation. In Proceedings of the INTERSPEECH, Singapore, 14–18 September 2014.
21. Tu, M.; Berisha, V.; Liss, J. Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. In Proceedings of the INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 1849–1853.
22. Li, M.; Tang, D.; Zeng, J.; Zhou, T.; Zhu, H.; Chen, B.; Zou, X. An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. *Comput. Speech Lang.* **2019**, *56*, 80–94.
23. Kent, R.D.; Eichhorn, J.; Wilson, E.M.; Suk, Y.; Bolt, D.M.; Vorperian, H.K. Auditory-perceptual features of speech in children and adults with Down syndrome: A speech profile analysis. *J. Speech Lang. Hear. Res.* **2021**, *64*, 1157–1175.
24. Corrales-Astorgano, M.; Martínez-Castilla, P.; Escudero-Mancebo, D.; Aguilar, L.; González-Ferreras, C.; Cardeñoso-Payo, V. Automatic Assessment of Prosodic Quality in Down Syndrome: Analysis of the Impact of Speaker Heterogeneity. *Appl. Sci.* **2019**, *9*, 1440.
25. Martínez-Castilla, P.; Peppé, S. Developing a test of prosodic ability for speakers of Iberian Spanish. *Speech Commun.* **2008**, *50*, 900–915.
26. Oates, J. Auditory-perceptual evaluation of disordered voice quality: Pros, cons and future directions. *Folia Phoniatr. Logop.* **2009**, *61*, 49–56.
27. Kreiman, J.; Gerratt, B.R.; Ito, M. When and why listeners disagree in voice quality assessment tasks. *J. Acoust. Soc. Am.* **2007**, *122*, 2354–2364.
28. Yoon, T.J.; Chavarria, S.; Cole, J.; Hasegawa-Johnson, M. Intertranscriber reliability of prosodic labeling on telephone conversation using toBI. In Proceedings of the Interspeech 2004, Jeju Island, Republic of Korea, 4–8 October 2004; pp. 2729–2732. <https://doi.org/10.21437/Interspeech.2004-659>.
29. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 835–838.
30. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202.
31. Ekberg, M.; Stavrinou, G.; Andin, J.; Stenfelt, S.; Dahlström, Ö. Acoustic Features Distinguishing Emotions in Swedish Speech. *J. Voice* **2023**. <https://doi.org/10.1016/j.jvoice.2023.03.010>.
32. Boersma, P. Praat: Doing Phonetics by Computer. Amsterdam, The Netherlands, 2006. Available online: <http://www.praat.org/> (accessed on 15 September 2023).
33. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18.
34. Hall, M.A. Correlation-Based Feature Subset Selection for Machine Learning. Ph.D. Thesis, University of Waikato, Hamilton, New Zealand, 1998.
35. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. <https://doi.org/10.1109/MCSE.2007.55>.
36. Orozco-Arroyave, J.R.; Vásquez-Correa, J.C.; Vargas-Bonilla, J.F.; Arora, R.; Dehak, N.; Nidadavolu, P.; Christensen, H.; Rudzicz, F.; Yancheva, M.; Chinaei, H.; et al. NeuroSpeech: An open-source software for Parkinson’s speech analysis. *Digit. Signal Process.* **2018**, *77*, 207–221. <https://doi.org/https://doi.org/10.1016/j.dsp.2017.07.004>.
37. Finnegan, D. *referenceIntervals: Reference Intervals*, 2020. R package version 1.2.0. Available online: <https://CRAN.R-project.org/package=referenceIntervals> (accessed on 21 September 2023).
38. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. <https://doi.org/10.1038/s41592-019-0686-2>.
39. Jang, S.J.; Choi, S.H.; Kim, H.M.; Choi, H.S.; Yoon, Y.R. Evaluation of performance of several established pitch detection algorithms in pathological voices. In Proceedings of the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 22–26 August 2007; pp. 620–623.
40. Van Borsel, J.; Vandermeulen, A. Cluttering in Down syndrome. *Folia Phoniatr. Logop.* **2008**, *60*, 312–317.
41. Devenny, D.; Silverman, W. Speech dysfluency and manual specialization in Down’s syndrome. *J. Intellect. Disabil. Res.* **1990**, *34*, 253–260.
42. Eggers, K.; Van Eerdenbrugh, S. Speech disfluencies in children with Down Syndrome. *J. Commun. Disord.* **2017**, *71*, 72–84.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.