






Article

# Automatic Assessment of Prosodic Quality in Down Syndrome: Analysis of the Impact of Speaker Heterogeneity †

Mario Corrales-Astorgano <sup>1,\*</sup>, Pastora Martínez-Castilla <sup>2,\*</sup>, David Escudero-Mancebo <sup>1,\*</sup>, Lourdes Aguilar <sup>3,\*</sup>, César González-Ferreras <sup>1,\*</sup> and Valentín Cardeñoso-Payo <sup>1,\*</sup>

<sup>1</sup> Department of Computer Science, University of Valladolid, 47002 Valladolid, Spain

<sup>2</sup> Department of Developmental and Educational Psychology, UNED, 28040 Madrid, Spain

<sup>3</sup> Department of Hispanic Philology, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

\* Correspondence: mcorrales@infor.uva.es (M.C.-A.); pastora.martinez@psi.uned.es (P.M.-C.); descuder@infor.uva.es (D.E.-M.); lourdes.aguilar@uab.cat (L.A.); cesargf@infor.uva.es (C.G.-F.); valen@infor.uva.es (V.C.-P)

† This paper is an extended version of our paper published in the conference IberSPEECH2018.

Received: 19 February 2019; Accepted: 3 April 2019; Published: 5 April 2019



**Abstract:** Prosody is a fundamental speech element responsible for communicative functions such as intonation, accent and phrasing, and prosodic impairments of individuals with intellectual disabilities reduce their communication skills. Yet, technological resources have paid little attention to prosody. This study aims to develop an automatic classifier to predict the prosodic quality of utterances produced by individuals with Down syndrome, and to analyse how inter-individual heterogeneity affects assessment results. A therapist and an expert in prosody judged the prosodic appropriateness of a corpus of Down syndrome' utterances collected through a video game. The judgments of the expert were used to train an automatic classifier that predicts prosodic quality by using a set of fundamental frequency, duration and intensity features. The classifier accuracy was 79.3% and its true positive rate 89.9%. We analyzed how informative each of the features was for the assessment and studied relationships between participants' developmental level and results: interspeaker variability conditioned the relative weight of prosodic features for automatic classification and participants' developmental level was related to the prosodic quality of their productions. Therefore, since speaker variability is an intrinsic feature of individuals with Down syndrome, it should be considered to attain an effective automatic prosodic assessment system.

**Keywords:** prosody; automatic classification; Down syndrome; educational video games

## 1. Introduction

Prosody is a fundamental speech element that contributes to conveying important communicative functions. For example, it contributes to establishing sentence-modality and conversational turns, conveying emotions, segmenting the speech-chain and expressing the focus of an utterance [1]. The importance of these functions highlights how communication can be negatively affected in individuals who present prosodic deficits [2]. Furthermore, in such cases, communication problems may lead to social isolation, especially when other linguistic components are also affected [2]. This is often the case of individuals with intellectual disability [2]. Intellectual disability can be caused by different factors and, among those, genetics plays a relevant role. This is well illustrated when considering Down syndrome, which is the most frequent genetic cause of intellectual disability [3]. Specifically, Down syndrome is caused by the presence of a third copy of chromosome 21 (usually called "trisomy 21"). The syndrome provokes a cascade of effects: to mention a few, middle ear disease;

immune and endocrine abnormalities; skeletal, heart and digestive system defects; cognitive, learning and attentional limitations; and our concern, language delays [4]. All the areas of language may be impaired, but not in the same degree, as described by Martin et al. [5]. Although lexical acquisition is delayed, difficulties with morphology and syntax appear to be more pronounced (e.g., incorrect use of morphemes; use of short sentences) [6]. With regard to pragmatics, individuals with Down syndrome have trouble producing and understanding questions and emotions, signaling turn-taking, or keeping to topics in conversation; while the study of Smith et al. [7] demonstrated that children with Down syndrome are impaired relative to norms from typically developing children in all areas of pragmatics. At the phonological level, speech intelligibility is seriously damaged by the presence of errors on producing some phonemes, the loss of consonants and the simplification of syllables [8]. In spite of this general description, variability in the different linguistic skills of individuals with Down syndrome has often been documented [4].

As far as prosody is concerned, Kent and Vorperian [9] report disfluencies (stuttering and cluttering) and impairments in the perception, imitation and spontaneous production of prosodic features; while Heselwood et al. [10] have connected some of the speech errors with difficulties in the identification of boundaries between words and sentences. Nevertheless, characterizing prosodic impairments in populations with developmental disorders is a hard task [11]. To fulfill such an aim, prosody assessment procedures appropriate for use with individuals with intellectual and/or developmental disabilities need to be employed. The Profiling Elements of Prosody in Speech-Communication (PEPS-C) test has proved to be successful in this respect [12,13]. PEPS-C follows a psycholinguistic approach by assessing both the skills needed to understand and express prosodic functions and those required to discriminate and imitate prosodic forms [14]. When used with English-speaking children with Down syndrome, a lower performance than expected by chronological age is observed in all prosody tasks [15]. After comparisons with typically developing children matched for mental age, impairments are also found for the discrimination and imitation of prosody [15].

To tackle linguistic impairments from a clinical perspective, technological tools aimed to facilitate speech and language therapy have been developed. These tools are called Computer-Aided Speech and Language Therapy (CASLT) tools and deal with a large variety of language problems. There are tools that are focused on training basic phonation skills in children with neuromuscular disorders, training the articulatory level of language or introducing the impaired child population to language understanding [16,17]. Other tools incorporate speech technologies to assist automated speech therapy in childhood apraxia of speech (CAS) [18], whereas others give a visual feedback of the positioning of the articulatory elements to produce different sounds [19]. Diagnosing and training tools for stuttering problems in children have also been developed [20]. However, despite the positive impact that prosodic training would have on communication abilities, little attention has been paid to the development of technological resources that specifically consider the learning of prosody in students with special needs, in particular those with Down syndrome. This can be explained by considering the difficulty of assigning a change in a suprasegmental feature to an intonational meaning in a unique and unambiguous way (since those features-tone, intensity, duration- co-occur to express a wide range of linguistic and paralinguistic meanings), together with the multiplicity of correct possibilities to reach the same intonational meaning. To advance in the line of developing specific resources to minimize the limitations concerning prosody and pragmatics in individuals with developmental and intellectual disabilities, we have developed an educational video game to train prosody, "PRADIA: Mystery in the city" [21,22] (see Section 2.1).

Automatic assessment of pathological speech has also been researched, but, in general, the studies on the topic are related to specific aspects and populations. Some works focus on the speech intelligibility of people with aphasia [23,24] or speech intelligibility in pathological voices [25,26]. Others try to identify speech disorders in children with cleft lip and palate [27] or to predict automatically some dysarthric speech evaluation metrics, such as intelligibility, severity and articulation impairment [28,29]. In addition, the recognition of speech emotions and autism spectrum disorders has also been

investigated [30]. All these works include a subjective evaluation carried out by experts as a reference to train the classification systems.

In this work, we analyze the difficulties of automatically predicting the quality of the prosody of the speech produced by individuals with Down syndrome and propose a new approach that will serve as a baseline for future work. Recordings of individuals with Down syndrome collected in different sessions using the educational video game “PRADIA: Mystery in the city” gave us information about the relevant features needed to make an automatic classification of the productions. The speech corpus obtained during the game was examined by a therapist, who evaluated in real time the quality of the oral productions, and by a prosody expert, who carried out an off-line evaluation. The judgments of the expert were used to train an automatic classifier that predicts quality by using acoustic features extracted from the recordings of the corpus. Results were related to measurements of participants’ developmental level, prosody perception and production performance, obtained in the PEPS-C test.

This methodology was followed to achieve two main objectives: (1) developing an automatic classifier to predict the prosodic quality of the utterances produced by individuals with Down syndrome when using the PRADIA video game; and (2) analyzing the impact of the speaker heterogeneity in the classification results. The paper is structured following these objectives. In Section 2, the experimental procedure is described, which includes a description of the video game, the procedure for corpus collection and evaluation, the processing of speech material and the classification of the samples. Section 3 describes the classification results (Objective 1) and the impact of the speaker variability on the classification results (Objective 2). We end the paper with a discussion of the relevance of the results (Section 4) to the assessment of the prosodic quality on people with Down syndrome (Objective 1) and the dependence of the speaker in this assessment (Objective 2). Finally, a conclusions (Section 5) is included.

## 2. Methodology

Figure 1 describes the experimental procedure followed in this work. The corpora are collected using the PRADIA video game (described in Section 2.1). The video game allows real time therapist decisions to be collected concerning whether the user has to repeat the production activity or continue playing (Section 2.2 details the different corpora collected). Next, an expert evaluates the acceptability of each of the utterances offline (Section 2.3 details both the therapist and expert evaluations). Section 2.4 describes the way prosodic features are computed and which features are selected to be included in the classifiers. The automatic assessment process is thus made up of the classic feature selection, training and testing sequence; details are given in Section 2.5.

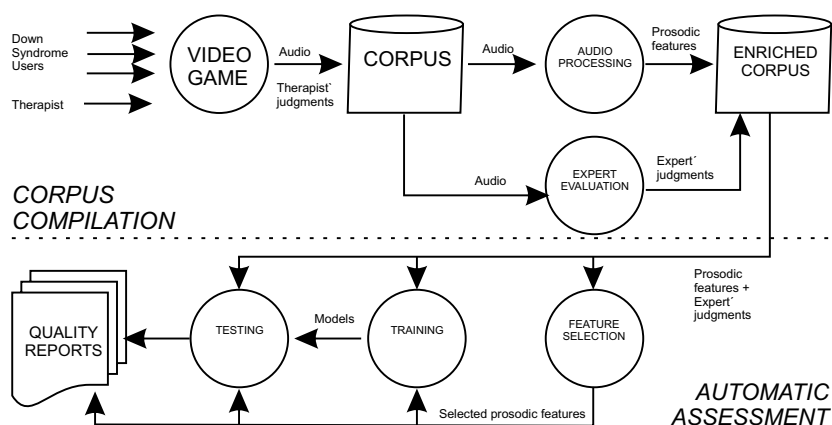


Figure 1. Experimental procedure scheme.

### 2.1. Game Description

PRADIA is an educational video game in which the learning objectives are integrated in the process of the game, implemented as a graphic adventure. The graphic adventure is a genre in

which the player assumes the role of a main character in an interactive story driven by exploration and problem solving. This allows a process of immersion in the virtual world shown, causing an identification of the individual with their character, which increases the user's involvement in the resolution of the story. This fact, at the same time, makes it easier for the person to integrate the learning content into their daily life.

The main way in which the player interacts with the game is through the voice. Although some of the activities may require some reading skill, the player is never asked to write. In order for the player to advance in the game, they must interact with the rest of the non-player characters (an elderly lady, a friend, a bus driver, etc.) and behave adequately in different communicative circumstances, where prosodic features are the most relevant to achieve a correct pragmatic interpretation (Figure 2). Priority is given to the suprasegmental features over segmental ones. Although the intelligibility of the utterances of speakers with Down syndrome is seriously affected [7], prosodic mistakes are what is mainly responsible for pragmatic failures in spoken communication, and this can lead to insecurities and low self esteem in people with Down syndrome. Therefore, we argue that any improvement in the suprasegmental domain will lead to an improvement in communication. The video game includes both prosodic comprehension and prosodic expression tasks. The focus of the video game is to enable the learner to communicate effectively and appropriately in the various situations in which they could find themselves. To do this, it is important to differentiate between the prosodic content according to the purpose pursued and to produce information with the appropriate prosodic features.



**Figure 2.** Example of a type of perception activity in the video game: the player must choose which of the two productions (Could I have a ticket, please? or A ticket!) is pragmatically adequate according to the communicative situation.

Although the video game was designed with the aim of training prosody in individuals with Down syndrome, it became a tool to collect their oral productions and thus to construct a prosodic corpus. In its current version, the video game needs the constant presence of a person (ideally a therapist) who guides the gamer throughout the adventure and who evaluates the success in resolving the production activities. The assistance of the therapist has proved crucial to motivate individuals with Down syndrome. Even so, it would be desirable to improve their autonomy and help trainers in their therapies with new functionalities by including a module of automatic assessment of prosodic quality. This is a difficult task, due to the high number of variables included in prosodic analysis and the heterogeneity of the cognitive and learning capabilities of this population.

### 2.2. Corpus Description

Table 1 describes the contents of the corpus compiled with the video game. A total of 966 utterances were collected corresponding to the oral turns of 23 players with Down syndrome. Although all the audio samples were collected in similar conditions and with the same recording device (a Logitech PC Headset 960 USB microphone), we distinguish the subcorpora C1, C2 and C3 as they were recorded in different sessions that we detail in the following paragraphs. To build the subcorpus C1, five young adults with Down syndrome (mean age 16.5 years) were recruited from a local Down syndrome Foundation located in Madrid (Spain). For sample selection, teachers working at the Foundation were asked to choose individuals with Down syndrome of different developmental levels. To account for the variability of individuals with Down syndrome and get measurements of different developmental variables, all of the participants were given the following tests. The Peabody Picture Vocabulary Scale-III [31] was used to assess verbal mental age, the forward digit-span subtest included in the Wechsler Intelligence Scale for Children-IV [32] was used to evaluate verbal short-term memory and Raven’s Coloured Progressive Matrices [33] served as a means to measure non-verbal cognitive level. The descriptive characteristics and scores obtained are shown in Table 2. The full PEPS-C battery in its Spanish version [34] was also administered to participants in order to have specific measurements of prosody level. The mean percentage of success in perception and production PEPS-C tasks is also presented in Table 2. Once these assessments had been completed, participants played the PRADIA video game. Each participant used PRADIA for a total duration of 4 h, distributed through four sessions of 1 h per week. The participants were supported by a speech and language therapist who knew them in advance and was an expert at working with individuals with Down syndrome. The therapist explained the game, helped participants when needed and took notes about how each session developed. Importantly, the therapist also assessed participants’ speech production and thus monitored their rhythm of progress within the video game.

**Table 1.** Corpus description. Concerning the therapist decision, Cont.R (Continue Right) means that the activity was rightly resolved, Cont (Continue) means that the activity was resolved but the response could be better and Rep. (Repeat) means that the activity was faultily resolved. Concerning the expert judgment, Right means that the recording was rightly produced and Wrong means that the recording was wrongly produced.

Speaker	#Utterances	Therapist Decision (Real Time)			Expert Judgment (Offline)		Corpus
		Cont.R	Cont.	Rep.	Right	Wrong	
S01	120	70	33	17	87	33	C1
S02	106	90	16	0	81	25	C1
S03	97	93	3	1	78	19	C1
S04	131	19	51	61	75	56	C1
S05	151	21	54	76	77	74	C1
S06	30	x	x	x	19	11	C2
S07	34	x	x	x	13	21	C2
S08	28	x	x	x	23	5	C2
S09	43	x	x	x	20	23	C2
S10	33	x	x	x	29	4	C2
S11	57	x	x	x	31	26	C3
S12	12	x	x	x	7	5	C3
S13	7	x	x	x	2	5	C3
S14	11	x	x	x	3	8	C3
S15	33	x	x	x	19	14	C3
S16	10	x	x	x	6	4	C3
S17	8	x	x	x	5	3	C3
S18	11	x	x	x	6	5	C3
S19	10	x	x	x	6	4	C3
S20	10	x	x	x	6	4	C3
S21	9	x	x	x	1	8	C3
S22	7	x	x	x	3	4	C3
S23	8	x	x	x	3	5	C3
Total	966	293	157	155	465	302	

**Table 2.** Description of the C1 subcorpus. For each speaker, this table shows Chronological age (CA), Verbal mental age (VA), Short-term verbal memory (STVM), and Non-verbal cognitive level (NVCL). Ages are expressed in months. In addition, the mean percentage of success in perception (MPercT) and production (MProdT) PEPS-C tasks are included.

Speaker	Gender	CA	VA	STVM	NVCL	MPercT	MProdT
S01	f	195	84	94	17	69.8%	48.3%
S02	m	204	99	134	18	76%	72.1%
S03	f	178	96	78	20	74%	74.7%
S04	m	190	60	below 74	10	60.4%	49.8%
S05	m	223	69	below 74	13	56.3%	45.7%

The C2 subcorpus was also recorded using PRADIA software. These recordings were obtained through the video game in one session of software testing with real users. This test session was done in a school of special education located in Valladolid (Spain). Five adults with Down syndrome, aged 18 to 25, participated in this test. The judgments obtained during this game session were discarded for this work because the speech productions were not evaluated by a therapist. The oral productions were judged in an offline mode by the expert in prosody.

The C3 subcorpus was recorded using an older version of PRADIA software, the Magic Stone [35], with fewer types of production activities. Eighteen young adults with Down syndrome participated in the different game sessions, which focused on how these users interacted with the video game. Five of these 18 speakers also participated in the recordings of the C2 subcorpus, so their productions were discarded from the C3 subcorpus. As in the C2 subcorpus, the judgments obtained from the assistant that helped players complete the adventure were not considered in the classifications. Instead, the oral productions were judged in an offline mode by the expert in prosody.

### 2.3. Corpus Evaluation

This section focuses on the methodology and criteria used to evaluate the prosodic quality of speech samples by a therapist (C1) and by an expert (C1, C2, C3).

#### 2.3.1. Evaluation Criteria

Following the categories of intonational phonology (that is, intonation, accent and prosodic organization) [36] and the learning objectives included in PRADIA, the following criteria were used to judge the participant's production by both the prosodic expert and the therapist:

- **Intonation:** adjustment to the expected modality. That is, if the target sentence must be interrogative and the speaker manages to model the intonation of a question, it is labeled as correct; otherwise, for instance, in the set of exclamatory phrases, if the speaker fails to reproduce an exclamatory intonation (within a range of intonation possibilities), the sentence is labeled as incorrect.
- **Accent:** preservation of the difference between lexical stress (stressed versus unstressed syllables) and accent (accented versus unaccented syllables). The loss of this difference can occur in three directions: (a) when tonal prominence appears in all the syllables, creating an undesired rhythmic effect; (b) when the speaker does not discriminate between stressed and unstressed syllables, as shown by the absence of variation in any of the acoustic parameters of intensity, duration and pitch; and (c) when there is tonal prominence variability but the syllable stress is inappropriately allocated.
- **Phrasing:** adjustment to the organization in prosodic groups and distinction between function and content words. The sentence is labeled as incorrect if every word is pronounced as if it were in an isolated context, without distinguishing between unstressed and stressed words. The sentence is also considered incorrect when the pauses are inappropriately allocated within the speech chain.

### 2.3.2. Therapist Evaluations

During the game sessions, a speech therapist sat next to the player and evaluated the production activities in real time (in addition to guiding the player in the game and providing help if needed). Due to the limited attention span of young people with Down syndrome and the varied motivational and emotional states they demonstrated throughout the play sessions, the therapist could allow the player to advance in the game and prevent them from getting frustrated and leaving the session. This was achieved through the scale of three evaluation options offered by the video game. This allowed the result of the oral activities to be evaluated by using the computer keyboard where the game is installed, in which each assessment value is associated to a key. If the evaluation was Cont.R (Continue with right result) or Cont. (Continue but the oral activity could be better), the video game advanced to the next activity. If the evaluation was Rep. (Repeat), the game offered a new attempt in which the player had to repeat the activity. For each activity, there was a predetermined number of attempts: when the attempts finished, the video game went to the next screen to avoid frustration, even if the activity was not successfully completed (and the therapist continued judging with Rep.).

Beside the criteria described in Section 2.3.1, for providing her judgments, the therapist also took into consideration the motivational and emotional status of each participant in each session. For example, if the participant was getting bored, anxious, or frustrated, the therapist, whenever possible, made more use of the category Cont. to allow the speaker to continue playing in an attempt to reduce any negative valence of the therapy context.

### 2.3.3. Expert Judgments

An expert in prosody evaluated the three subcorpora of oral productions of 23 speakers with Down syndrome in an offline mode. In the offline evaluation, the external components implied in the development of the game (level of frustration, among others) were left aside in benefit of the examination of the prosodic variables: as a consequence, an evaluation system based on a binary decision (Right or Wrong production) was used. With a website support, the prosody expert listened to each audio file and decided whether the speaker had not achieved the required quality; or their production was satisfactory. The judgments were made relying on a purely auditory basis focused only on the intonational and prosodic structure of the recording, without any acoustic analysis of the sentences. Related to this, factors of intelligibility, quality in pronunciation or adjustment to the expected sentence were not taken into account. Even in the case of speakers with a low cognitive level and serious problems of intelligibility, the main criterion was whether they had modeled prosody with a certain success, even if the message was not understood. Just like the therapist's evaluations, the sentences were judged as right or wrong according to the categories of intonational phonology and the learning objectives of PRADIA.

## 2.4. Feature Extraction and Selection

The openSmile toolkit [37] was used to extract acoustic features from each recording of the C1, C2 and C3 subcorpora. The GeMAPS feature set [38] was selected due to the variety of acoustic and prosodic features contained in this set, which includes frequency related features, energy related features, spectral features and temporal features. The arithmetic mean and the coefficient of variation were calculated on these features. Furthermore, four additional temporal features were added: the silence and sounding percentages, silences per second and the mean silences. The complete description of these features can be found in previous research [39]. In this work, only prosodic features (frequency, energy and temporal) were used because spectral features improve speaker identification, and classifiers can be adapted to each speaker in the classification process. In total, 34 prosodic features were employed (see Appendix A).

Finally, to rank features by their importance to each C1 speaker, the Caret R package [40] was used. The importance of features was estimated by building an SVM model and ordered using the ROC

curve value of each feature. The Receiver operating characteristic curve (ROC curve) is a graphical representation of the results of a binary classifier system where the discrimination threshold is varied. Some studies recommend using the area under the ROC curve (AUC) in preference to overall accuracy for evaluation of machine learning algorithms, especially when the classes are unbalanced [41]. We also used feature selection before training the classifiers: the features were selected by measuring the information gain of the training set and discarding the ones in which the information gain equals zero (column Feat. in Table 3).

### 2.5. Automatic Classification

As explained in Section 2.3, the recordings were evaluated by the therapist and the prosody expert. Since the final aim of the module is to decide whether the gamer can continue the game or should repeat the activity (without considering degrees of failure), the evaluation of the expert was used to build the classifier. According to this, the outputs of the different classifiers were Right (R) or Wrong (W), based on the prosody expert scoring. The Weka machine learning toolkit [42] was used as well as three different classifiers to compare their performance: the C4.5 decision tree (DT), the multilayer perceptron (MLP) and the support vector machine (SVM). The stratified 10-fold cross-validation technique was used to create the training and testing datasets.

## 3. Results

### 3.1. Classification Results

Table 3 presents the performance of the different classification systems in the task of automatically predicting the expert judgments. The different cases displayed in the table (case A to F) are based on the different subcorpora. These subcorpora were recorded with a different version of the video game and a different recording context and they were not balanced in terms of sample size and number of speakers, so it was important to know how these differences would affect the classifier accuracy. Therefore, the results of using the recordings of the three subcorpora, as well as all the combinations of these subcorpora, were compared. The SVM classifier works better with all subcorpora and the worst results are obtained using the DT classifier (best case is 79.3% vs. 64.9% baseline). The best results are obtained in Cases A and D by using any of the three classifiers (UAR 0.83 with SVM classifier). The classification accuracy decreases when the C3 corpus is entered (C, E, F and G cases), as the number of speakers substantially increases. On average, we obtain 89.9% of true positives. This will be discussed in the next section as a positive result for real time situations.

**Table 3.** Classification results depending on the corpus and the classifier used. The prosody expert judgments were used to train the classifiers. This table shows the performance baseline (BL) of each group of samples (number of samples of the most populated class divided by all the samples), Decision trees (DT), Support vector machines (SVM), Multilayer Perceptron (MLP), classification rate (CR), Area Under the Curve (AUC) and Unweighted Average Recall (UAR). The number of samples (utt.), the number of speakers (SPK) and the number of features (Feat.) are presented. The output of the different classifiers are Right or Wrong, based on prosody expert scoring.

	Corpora	BL	DT			SVM			MLP			#Utt.	#Feat.	#SPK
			CR	AUC	UAR	CR	AUC	UAR	CR	AUC	UAR			
Case A	C1	65.8%	69.6%	0.68	0.74	78.5%	0.74	0.83	73.2%	0.7	0.79	605	21	5
Case B	C2	61.9%	60.3%	0.58	0.61	72.7%	0.7	0.79	68.5%	0.67	0.73	168	16	5
Case C	C3	50.8%	65.8%	0.66	0.66	61.6%	0.62	0.69	63.7%	0.64	0.64	193	7	13
Case D	C1+C2	64.9%	70.8%	0.68	0.75	79.3%	0.76	0.83	72.6%	0.7	0.78	773	21	10
Case E	C1+C3	62.2%	66.3%	0.65	0.69	72.3%	0.7	0.79	67.2%	0.65	0.74	798	20	18
Case F	C2+C3	56%	60.9%	0.6	0.64	66.5%	0.66	0.75	64%	0.63	0.69	361	13	18
Case G	C1+C2+C3	62.1%	66.9%	0.66	0.71	74.3%	0.71	0.81	69.4%	0.66	0.76	996	20	23



Table 4 shows the prosodic features with more influence in the utterance assessment of each C1 subcorpus speaker. The data for all the speakers of the C1 subcorpus were used to remove the highly correlated features (above 0.8 of Pearson correlation). After this redundant feature deletion, 22 of the 34 features were selected. Within these 22 features, only 10 present a significant ROC area value (above 0.6).

**Table 4.** Ranking of the correlated prosodic features (frequency, energy and temporal) between each C1 speaker and the expert evaluation. The first number represents the order of each feature in the ranking and the second number shows the area under the ROC curve. These values were calculated using an SVM classifier. The features are sorted by their importance when all data are used. Values in bold represent an area under the ROC curve above 0.6.

Feature	S01	S02	S03	S04	S05	All
silencesMean	4 (0.675)	6 (0.638)	3 (0.673)	2 (0.744)	2 (0.662)	1 (0.692)
silencesPerSecond	10 (0.6)	1 (0.754)	2 (0.696)	3 (0.725)	11 (0.581)	2 (0.683)
jitterLocal_sma3nz_amean	1 (0.688)	16 (0.534)	5 (0.618)	8 (0.683)	22 (0.515)	3 (0.65)
F0semitoneFrom27.5Hz_sma3nz_stddevNorm	7 (0.646)	7 (0.633)	22 (0.506)	4 (0.712)	17 (0.559)	4 (0.647)
jitterLocal_sma3nz_stddevNorm	2 (0.683)	2 (0.681)	18 (0.524)	6 (0.689)	9 (0.592)	5 (0.631)
F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope	5 (0.662)	12 (0.578)	7 (0.601)	9 (0.66)	3 (0.651)	6 (0.629)
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	17 (0.572)	3 (0.67)	4 (0.652)	16 (0.548)	1 (0.684)	7 (0.628)
F0semitoneFrom27.5Hz_sma3nz_pctlrange0.2	11 (0.598)	14 (0.559)	15 (0.545)	7 (0.689)	18 (0.544)	8 (0.626)
F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope	3 (0.679)	8 (0.6)	8 (0.6)	14 (0.595)	10 (0.588)	9 (0.625)
StddevVoicedSegmentLengthSec	22 (0.506)	4 (0.66)	16 (0.535)	1 (0.762)	16 (0.561)	10 (0.601)
loudnessPeaksPerSec	12 (0.595)	13 (0.563)	12 (0.558)	17 (0.533)	8 (0.603)	11 (0.586)
shimmerLocaldB_sma3nz_stddevNorm	9 (0.601)	5 (0.642)	11 (0.58)	13 (0.598)	14 (0.563)	12 (0.583)
shimmerLocaldB_sma3nz_amean	6 (0.651)	18 (0.523)	1 (0.698)	20 (0.532)	21 (0.528)	13 (0.583)
loudness_sma3_stddevNorm	18 (0.569)	10 (0.585)	14 (0.548)	11 (0.615)	20 (0.53)	14 (0.579)
MeanUnvoicedSegmentLength	8 (0.617)	9 (0.586)	20 (0.518)	10 (0.628)	19 (0.542)	15 (0.557)
StddevUnvoicedSegmentLength	15 (0.579)	11 (0.581)	21 (0.511)	12 (0.613)	13 (0.575)	16 (0.555)
loudness_sma3_meanFallingSlope	13 (0.59)	15 (0.549)	17 (0.524)	15 (0.578)	7 (0.609)	17 (0.545)
VoicedSegmentsPerSec	16 (0.573)	19 (0.517)	9 (0.599)	21 (0.519)	12 (0.577)	18 (0.521)
loudness_sma3_stddevFallingSlope	19 (0.528)	21 (0.513)	13 (0.555)	18 (0.532)	5 (0.624)	19 (0.519)
loudness_sma3_pctlrange0.2	20 (0.512)	22 (0.504)	10 (0.586)	5 (0.696)	4 (0.63)	20 (0.514)
MeanVoicedSegmentLengthSec	14 (0.582)	20 (0.514)	6 (0.617)	22 (0.503)	15 (0.562)	21 (0.514)
loudness_sma3_stddevRisingSlope	21 (0.509)	17 (0.523)	19 (0.52)	19 (0.532)	6 (0.612)	22 (0.501)

### 3.2. Speakers Variability Results

The 22 selected features were ranked by their importance to each speaker of the C1 subcorpus (Table 4). There is a high variability among speakers in the relevance of the different prosodic features. The feature *silencesMean* is very relevant for all speakers, but the same is not true for the rest of the features. For example, *silencesPerSecond* appears at the top for all speakers except for S05. In addition, the features related to frequency (from 3 to 9 in the ranking) are relevant to all speakers, but the specific feature and its importance greatly varies for each speaker. Finally, the intensity features and other rhythm features are less relevant in general, but are present at the top of the ranking for some speakers (speakers S02 and S03).

A high difference between speakers is also seen with regard to their developmental level and prosodic skills, as can be inferred if we relate the figures of Table 2 with those of Table 5. S04 and S05 have the lowest scores in *verbal mental age* (60 and 69, respectively), *short-term verbal memory* (below 74 both speakers) and *non-verbal cognitive level* (10 and 13, respectively). In addition, both have the lowest mean percentage of success in perception PEPS-C tasks (60.4% and 56.3%, respectively) and lower mean percentage of success in production PEPS-C tasks (49.8% and 45.7%, respectively). These low scores are related to the quality of the productions, with a higher percentage of *W* assignments from the prosody expert (42.8% and 49% respectively) and higher percentage of *Rep.* from the therapist (47% and 50%, respectively).

**Table 5.** Percentage of coincidence between therapist decision, classifier (SVM in case D) and prosody expert per speaker. Concerning the classifier, R represents the utterances classified as Right by the classifier and W represents the utterances classified as Wrong by the classifier. Each row percentage is relative to the number of each type of utterances of prosody expert evaluation.

Speaker	#Total utt	Expert Judgment		Classified as		Therapist Decision		
		Type	#utt	R	W	Cont.R	Cont.	Rep.
S01	120	R	87	83.9%	16.1%	69%	24.1%	6.9%
		W	33	57.6%	42.4%	30.3%	36.7%	33.3%
S02	106	R	81	87.7%	12.4%	85.2%	14.8%	0.0%
		W	25	28.0%	72.0%	84.0%	16.0%	0.0%
S03	97	R	78	97.4%	2.6%	94.9%	3.9%	1.3%
		W	19	73.7%	26.3%	100.0%	0.0%	0.0%
S04	131	R	75	94.6%	5.3%	21.3%	44.0%	34.7%
		W	56	41.1%	58.9%	5.4%	32.1%	62.5%
S05	151	R	77	87%	13%	20.8%	50.7%	28.6%
		W	74	29.7%	70.3%	6.8%	20.3%	73%
Total	605	R	398	89.9%	10.1%	80.2%	68.8%	35.5%
		W	207	41.1%	58.9%	19.8%	31.2%	64.5%

In order to see the influence of the speaker in the classification results, we present results per speaker in Table 5 and we focus on Case D to comment on them. Only the samples of corpus C1 are analyzed because they were evaluated by the two evaluators and because measurements of their developmental level were available. Comparing the judgments of the expert with the classifier predictions, there is a high recall in the R-R case for all speakers (S01 83.9%, S02 87.7%, S03 97.4%, S04 94.6%, S05 88%). The coincidence in the W-W case is lower: while S02 and S05 present a reasonable classification rate (72% and 70.3%, respectively), results for S03 go down to 26.3%. Furthermore, most of the utterances judged as wrong by the expert were rated as right by the therapist (100% in cell W-Cont.R for S3).

Concerning the therapist's judgments, the *Cont.R* decision could be identified as a *Right* assignment in a high percentage of cases for S01, S02 and S03 speakers (69%, 85.2% and 94.9% respectively). These are the speakers with a higher developmental level, according to Table 2. Of these three participants, the first, with more disagreement between the therapist and the prosody expert, showed the lowest prosodic level from the outset. In general, the correspondence between real time decisions and expert judgment is not straightforward, with a high variety in the contingency table. Concerning the therapist's *Rep.* decision, the highest percentages of agreement are obtained for S04 and S05 speakers (62.5% and 73.0%, respectively), who are the speakers with the lowest developmental level in all the variables measured, as seen in Table 2.

To deepen our analysis of how the inter-individual heterogeneity can affect the assessment results, Pearson correlation coefficients were calculated between the profile of the speakers of the C1 subcorpus and the assessment values. To ensure the appropriateness of the use of this correlation coefficient, the normality assumption was first checked for all the variables under analysis [43]. The Kolmogorov-Smirnov test showed that the assumption was fulfilled for all cases ( $p$ -value > 0.05). Table 6 shows the Pearson correlation results. Short-term verbal memory (STVM) is not included in this analysis because the values of the STVM of S04 and S05 were not high enough to be taken into account. Verbal mental age (VA) is highly and significantly correlated with Non-verbal cognitive level (NVCL), Prosodic perception (MPercT), percentage of Right expert evaluations (RRate) and all therapist evaluations. The correlation is positive for the NVCL, MPercT, RRate and ConRRate features, while the correlation is negative for the ContRate and RepRate features. NVCL is significantly correlated with RRate and ContRRate (positive) and with ContRate and RepRate (negative). In addition, MPercT is positively and significantly correlated with RRate and ContRRate, and negatively and significantly

correlated with RepRate. Prosodic production (MProdT) is significantly correlated with ContRRate in a positive way and is significantly correlated with ContRate in a negative way. Finally, the automatic classification rate (CR) is highly correlated with the prosodic production competences (MProdT), but the correlation did not reach statistical significance.

**Table 6.** Correlation values (Pearson correlation coefficient) between C1 speakers' profile variables (Table 2) and evaluators' assessment (Table 5). RRate means the percentage of Right judgments of the prosody expert. ContRRate, ContRate and RepRate mean the percentage of Cont.R, Cont. and Rep. evaluated by the therapist, respectively. CR means the classification rate of the SVM classifier. Values in bold represent statistically significant correlations with  $p$ -value < 0.05.

	CA	VA	NVCL	MPercT	MProdT	RRate	ContRRate	ContRate	RepRate	CR
CA	1.0	−0.29	−0.36	−0.53	−0.49	−0.64	−0.52	0.53	0.5	−0.16
VA		1.0	<b>0.96</b>	<b>0.93</b>	0.84	<b>0.91</b>	<b>0.97</b>	<b>−0.91</b>	<b>−0.96</b>	0.42
NVCL			1.0	0.87	0.76	<b>0.9</b>	<b>0.95</b>	<b>−0.92</b>	<b>−0.93</b>	0.29
MPercT				1.0	0.83	<b>0.98</b>	<b>0.96</b>	−0.86	<b>−0.99</b>	0.36
MProdT					1.0	0.81	<b>0.89</b>	<b>−0.93</b>	−0.83	0.80

## 4. Discussion

### 4.1. Analysis of the Classification Results

The results presented in Table 3 show different accuracy results depending on the classifier used and the subcorpus included to train the classifiers. Focusing on SVM, which is the most accurate classifier, the best results were obtained with the subcorpus C1, and C1+C2 subcorpora. However, the classifier trained with C3 subcorpus presented the worst results. The C3 subcorpus included more speakers than the other subcorpora, but much fewer samples of each speaker. This result indicates that the sample size is more important than the number of speakers to obtain a better classification accuracy.

In the PRADIA video game, it is very important to avoid evaluating as wrong a correct utterance; otherwise, frustration may arise. This is even more important when individuals with Down syndrome are the players, since they can be particularly susceptible to this feeling [4]. Bearing in mind that the video game aims to engage and motivate the users, the percentage of false negatives must be as low as possible. Table 5 shows that only 10.1% of the samples evaluated as Right by the expert are classified as Wrong by the classifier.

An additional strength of the classifier developed here arises when comparing it with prior work on automatic assessment of disordered speech, such as aphasic speech [44] or dysarthric speech [45], where prosodic features and pronunciation scores are combined. In our study, instead, prosody was assessed by leaving aside the well-known difficulties of pronunciation of individuals with Down syndrome. Regardless of their intelligibility problems, prosody alone makes the speech of these individuals sound atypical [39]. Therefore, the development of an automatic speech assessment only focused on prosody in Down syndrome represents a relevant contribution. In addition, in the context of the PRADIA video game used in this study, the development of an automatic prosody assessment system in which pronunciation problems are not considered is important. Thus, for this video game to become a valuable prosody self-learning tool, communicative skills that provide an appropriate handling of prosody (production and distinction of sentence modalities and accents, among others) need to be prioritized.

### 4.2. Impact of Variability on Assessment

As shown in Table 2, the chronological age of the participants for whom both the therapist and prosody expert evaluations were available was similar. However, their skills for reasoning, recalling auditory verbal material and understanding vocabulary were clearly different. Given the sample selection criterion (see Section 2.2), the heterogeneity found in these developmental measurements was expected.

As shown in Table 6, when the developmental level is low, the quality of the prosodic productions is also low (positive high correlation with the VA-RRate and the VA-ContRRate and negative high correlation with the VA-ContRate and the VA-RepRate). This has an impact on the raters' assessment and on the likelihood of their agreement as to the appropriateness of the output. The speakers S01, S02 and S03—who had the highest developmental and prosodic level (Table 2)—present higher values of agreement in right cases (R-Cont.R) than the S04 and S05 speakers (Table 5). This shows the difficulties inherent to the task being carried out. Furthermore, even in the cases of a higher cognitive level, variability in the linguistic profile can also play a role. Thus, levels of vocabulary are not necessarily paired with those of prosody perception and production (Table 2). A high prosodic perception level seems to help players to obtain a better assessment in their speech productions (positive highly correlated MPercT-RRate, MPercT-ContTRate; negative highly correlated MPercT-RepRate). A high prosodic production level seems to be related to a good assessment of the recordings, but the correlation is only significant with ContRRate and ContRate (Table 6). The lack of statistical significance in other high correlation coefficients (e.g., MPercT-ContRate) can be explained by considering the small sample size.

In short, agreement between the prosodic expert and the therapist depends on the speaker's developmental levels and the type of sentence produced (right or wrong). In addition, differences in the evaluation context can also explain raters' disagreements. Thus, while the expert only based her decisions on intonational criteria, the therapist also took into consideration the progress of the player while playing the video game. In doing so, avoiding frustration was a priority; therefore, levels of frustration tolerance and number of failures influenced the therapist's decisions.

The high variability of the speech of individuals with Down syndrome has also been shown in our experimental results regarding the use of prosodic features. Table 4 shows the differences in the correlated features with the expert evaluation per speaker. Some rhythm and frequency features appear above in the ranking of all the speakers (from *silencesMean* to *F0semitoneFrom27.5Hz\_sma3nz\_stddevFallingSlope* on Table 4). However, intensity features seem to be very important to the S03 and S05 speakers, but present lower importance to the others. Speaker S04 presents higher ROC area values than the other speakers in his features, so the lower importance of these features to the other speakers can affect the performance of the automatic classifiers. This heterogeneity complicates the automatic assessment of prosody quality, because automatic classifiers show poorer generalization power.

In addition, this inherent inter-speaker variability of acoustic features in Down syndrome may have also been a source of the disagreements between the therapist and the expert. In fact, prior research has shown that, as compared to typical voices, the speech signal in pathological voices may be characterized by a higher variability of specific acoustic parameters. As a consequence, a lower level of agreement among perceptual judgments may be found when evaluating pathological voices [46].

#### 4.3. Limitations and Future Work

Although the amount of prosodic productions analyzed was large (605 utterances, as shown in Table 5), the number of informants, especially in C1 subcorpus, is low. We have reported statistically significant correlations by using these samples, but the statistical power of these results needs to be strengthened with future recordings of more participants. Nevertheless, the sample size has not prevented us from fulfilling the first aim of the paper. Thus, an accurate automatic classification system with a low rate of false negatives was successfully developed. Moreover, the criterion for sample selection (see Section 2.2) by which teachers were asked to choose individuals with Down syndrome of different developmental levels ensured that the sample was representative of the variability inherent to the population of individuals with Down syndrome. This allowed us to analyze how the heterogeneity of these individuals can affect the assessment results and therefore reach our second aim. Even so, it should be noted that further research should compile a bigger and more balanced corpus of the speech of individuals with Down syndrome and should also record a reference corpus of people with typical

development. A bigger corpus will have to be compiled in order to explore new approaches such as end-to-end deep learning methods, which have shown promising results in the assessment of atypical prosody in other populations with intellectual disabilities, such as Autism Spectrum Disorder [47].

The differences between the therapist and prosody expert evaluations highlight the importance of evaluation contexts. If the automatic evaluation module aims to be included in a real time video game, aspects different from prosody should also be considered, in the line of what the therapist did in her evaluation. In addition, the evaluation scale can be improved by adding more dimensions to be scored by the experts. Instead of having a global score of the prosody of a recording, the experts could assign a different score to different prosodic dimensions (intonation, accent, phrasing), with the aim of making a more precise classification. These features could then also be automatically classified.

As already mentioned, the fact that 10.1% of the samples evaluated as Right by the expert are classified as Wrong by the classifier is a good result, because evaluating a recording as Wrong when the recording was Right can affect the motivation of the player. Yet, reducing this rate of false negatives in order to obtain the best possible reliable evaluation system is work in progress. To reach this goal, inter-speaker variability should be taken into account as an intrinsic feature of individuals with Down syndrome, so that both the reference for correct pronunciation and the particular limitations of the speaker should be taken into account to ensure an effective automatic prosodic assessment.

Knowing the variables that contribute to variability in the quality of the prosodic productions of individuals with Down syndrome paves the way for the design of the best possible automatic classification system for the PRADIA video game as an intervention tool, in particular, or for other future intervention programs, in general. Having such an automatic classification module would allow the player to have more autonomy, which in turn would have a positive impact on his/her self-confidence and motivation. At the same time, the automatic classification module would release resources for the therapist, who could use the time needed to support individuals with Down syndrome in the PRADIA intervention tool for other intervention activities. Therefore, our results represent a first step for the future development of useful intervention materials. Thus, our results could benefit future clinical practices.

## 5. Conclusions

In this study, we have developed an automatic classifier to predict the prosodic quality of the utterances produced by individuals with Down syndrome. The study has also analyzed how the heterogeneity of people with Down syndrome can affect the assessment of the prosody quality of their utterances. By doing this, the study shows some of the variables that contribute to accounting for the difficulties of conducting an automatic evaluation of prosody in speakers with Down syndrome.

The acoustic features that are important for classifying a recording as Right or Wrong differed depending on each speaker of the C1 subcorpus. Evaluation results were highly dependent on the different speaker profiles. We found significant correlations between VA, NVCL and mean percentage of success in perception PEPS-C tasks with the therapist and expert evaluations. In addition, the coincidence between evaluators was highly dependent on the prosodic quality of the recordings and the speakers' heterogeneity. The agreement was high in the assessment of high prosodic quality utterances (values above 80%). However, the agreement was lower when the prosodic quality of the recordings was poor. To sum up, variability in the cognitive and linguistic skills of individuals with Down syndrome is common. To build an automatic evaluation of these recordings, this variability has to be taken in account.

**Author Contributions:** Conceptualization, M.C.-A., P.M.-C., D.E.-M., L.A. and C.G.-F.; Data curation, M.C.-A., P.M.-C. and L.A.; Formal analysis, M.C.-A. and D.E.-M.; Funding acquisition, D.E.-M., L.A. and V.C.-P.; Investigation, M.C.-A., P.M.-C., D.E.-M., L.A., C.G.-F. and V.C.-P.; Methodology, M.C.-A., D.E.-M., C.G.-F. and V.C.-P.; Project administration, D.E.-M., L.A. and V.C.-P.; Resources, P.M.-C. and L.A.; Software, M.C.-A.; Supervision, D.E.-M., C.G.-F. and V.C.-P.; Validation, M.C.-A., P.M.-C. and L.A.; Visualization, M.C.-A., D.E.-M. and V.C.-P.; Writing—original draft, M.C.-A. and D.E.-M.; Writing—review & editing, M.C.-A., P.M.-C., D.E.-M., L.A., C.G.-F. and V.C.-P.

**Funding:** The activities of Down syndrome speech analysis continue (1/2018-12/2020) in the project funded by the Ministerio de Ciencia, Innovación y Universidades and the European Regional Development Fund FEDER (TIN2017-88858-C2-1-R) and in the project funded by Junta de Castilla y León (VA050G18). Part of this work was funded by BBVA Foundation (2015-2017) in the framework of the project PRADIA: Pragmatics and prosody: the graphic adventure game.

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

## Appendix A. Description of the Features

The tables included in this appendix describe the features used in each of the domains. Frequency features are presented in Table A1. Energy features are described in Table A2. Temporal features are explained in Table A3.

**Table A1.** Frequency features explained. All functionals are applied to voiced regions only. Text in brackets shows the original name of the eGeMAPS features.

Feature	Description
F0_mean (F0semitoneFrom27.5Hz_sma3nz_amean)	Mean of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_stddevNorm (F0semitoneFrom27.5Hz_sma3nz_stddevNorm)	Coefficient of variation of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile20 (F0semitoneFrom27.5Hz_sma3nz_percentile20.0)	Percentile 20-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile50 (F0semitoneFrom27.5Hz_sma3nz_percentile50.0)	Percentile 50-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile80 (F0semitoneFrom27.5Hz_sma3nz_percentile80.0)	Percentile 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_pctrange (F0semitoneFrom27.5Hz_sma3nz_pctrange0-2)	Range of 20-th to 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_meanRisingSlope (F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope)	Mean of the slope of rising signal parts of F0
F0_stddevRisingSlope (F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope)	Standard deviation of the slope of rising signal parts of F0
F0_meanFallingSlope (F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope)	Mean of the slope of falling signal parts of F0
F0_stddevFallingSlope (F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope)	Standard deviation of the slope of falling signal parts of F0
jitter_mean (jitterLocal_sma3nz_amean)	Mean of the deviations in individual consecutive F0 period lengths
jitter_stddevNorm (jitterLocal_sma3nz_stddevNorm)	Coefficient of variation of the deviations in individual consecutive F0 period lengths

**Table A2.** Energy features explained. All functionals are applied to voiced and unvoiced regions together. Text in brackets shows the original name of the eGeMAPS features.

Feature	Description
loudness_mean (loudness_sma3_amean)	Mean of estimate of perceived signal intensity from an auditory spectrum
loudness_stddevNorm (loudness_sma3_stddevNorm)	Coefficient of variation of estimate of perceived signal intensity from an auditory spectrum
loudness_percentile20 (loudness_sma3_percentile20.0)	Percentile 20-th of estimate of perceived signal intensity from an auditory spectrum
loudness_percentile50 (loudness_sma3_percentile50.0)	Percentile 50-th of estimate of perceived signal intensity from an auditory spectrum
loudness_percentile80 (loudness_sma3_percentile80.0)	Percentile 80-th of estimate of perceived signal intensity from an auditory spectrum
loudness_pctlrange02 (loudness_sma3_pctlrange0-2)	Range of 20-th to 80-th of estimate of perceived signal intensity from an auditory spectrum
loudness_meanRisingSlope (loudness_sma3_meanRisingSlope)	Mean of the slope of rising signal parts of loudness
loudness_stddevRisingSlope (loudness_sma3_stddevRisingSlope)	Standard deviation of the slope of rising signal parts of loudness
loudness_meanFallingSlope (loudness_sma3_meanFallingSlope)	Mean of the slope of falling signal parts of loudness
loudness_stddevFallingSlope (loudness_sma3_stddevFallingSlope)	Standard deviation of the slope of falling signal parts of loudness
shimmer_mean (shimmerLocaldB_sma3nz_amean)	Mean of difference of the peak amplitudes of consecutive F0 periods
shimmer_stddevNorm (shimmerLocaldB_sma3nz_stddevNorm)	Coefficient of variation of difference of the peak amplitudes of consecutive F0 periods

**Table A3.** Temporal features explained. The first four features are not included in the eGeMAPS feature set.

Feature	Description
silencePercentage	Duration percentage of unvoiced regions
silencesMean	Mean of unvoiced regions
silencesPerSecond	The number of silences per second
soundingPercentage	Duration percentage of voiced regions
loudnessPeaksPerSec	The number of the loudness peaks per second
VoicedSegmentsPerSec	The number of continuous voiced regions per second
MeanVoicedSegmentLengthSec	Mean of continuously voiced regions
StddevVoicedSegmentLengthSec	Standard deviation of continuously voiced regions
MeanUnvoicedSegmentLength	Mean of unvoiced regions
StddevUnvoicedSegmentLength	Standard deviation of unvoiced regions

## References

1. Roach, P. *English Phonetics and Phonology Fourth Edition: A Practical Course*; Ernst Klett Sprachen: Cambridge, UK, 2010.
2. Wells, B.; Peppé, S.; Vance, M. Linguistic assessment of prosody. *Linguistics in Clinical Practice*; Whurr: London, UK, 1995; pp. 234–265.
3. Fidler, D.J.; Nadel, L. Education and children with Down syndrome: Neuroscience, development, and intervention. *Ment. Retard. Dev. Disabil. Res. Rev.* **2007**, *13*, 262–271. [[CrossRef](#)]

4. Grieco, J.; Pulsifer, M.; Seligsohn, K.; Skotko, B.; Schwartz, A. Down syndrome: Cognitive and behavioral functioning across the lifespan. *Am. J. Med. Genet. Part C Semin. Med. Genet.* **2015**, *169*, 135–149. [[CrossRef](#)] [[PubMed](#)]
5. Martin, G.E.; Klusek, J.; Estigarribia, B.; Roberts, J.E. Language characteristics of individuals with Down syndrome. *Top. Lang. Disord.* **2009**, *29*, 112. [[CrossRef](#)]
6. Eadie, P.A.; Fey, M.; Douglas, J.; Parsons, C. Profiles of grammatical morphology and sentence imitation in children with specific language impairment and Down syndrome. *J. Speech Lang. Hear. Res.* **2002**, *45*, 720–732. [[CrossRef](#)]
7. Smith, E.; Næss, K.A.B.; Jarrold, C. Assessing pragmatic communication in children with Down syndrome. *J. Commun. Disord.* **2017**, *68*, 10–23. [[CrossRef](#)]
8. Laws, G.; Bishop, D.V. Verbal deficits in Down's syndrome and specific language impairment: A comparison. *Int. J. Lang. Commun. Disord.* **2004**, *39*, 423–451. [[CrossRef](#)]
9. Kent, R.D.; Vorperian, H.K. Speech impairment in Down syndrome: A review. *J. Speech Lang. Hear. Res.* **2013**, *56*, 178–210. [[CrossRef](#)]
10. 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. [[CrossRef](#)]
11. Peppé, S.J. Why is prosody in speech-language pathology so difficult? *Int. J. Speech-Lang. Pathol.* **2009**, *11*, 258–271. [[CrossRef](#)]
12. Martínez-Castilla, P.; Sotillo, M.; Campos, R. Prosodic abilities of Spanish-speaking adolescents and adults with Williams syndrome. *Lang. Cogn. Process.* **2011**, *26*, 1055–1082. [[CrossRef](#)]
13. Peppé, S.; McCann, J.; Gibbon, F.; O'Hare, A.; Rutherford, M. Receptive and expressive prosodic ability in children with high-functioning autism. *J. Speech Lang. Hear. Res.* **2007**, *50*, 1015–1028. [[CrossRef](#)]
14. Peppé, S.; McCann, J. Assessing intonation and prosody in children with atypical language development: The PEPS-C test and the revised version. *Clin. Linguist. Phon.* **2003**, *17*, 345–354. [[CrossRef](#)]
15. Stojanovik, V. Prosodic deficits in children with Down syndrome. *J. Neurolinguist.* **2011**, *24*, 145–155. [[CrossRef](#)]
16. Saz, O.; Yin, S.C.; Lleida, E.; Rose, R.; Vaquero, C.; Rodríguez, W.R. Tools and technologies for computer-aided speech and language therapy. *Speech Commun.* **2009**, *51*, 948–967. [[CrossRef](#)]
17. Rodríguez, W.R.; Saz, O.; Lleida, E. A prelingual tool for the education of altered voices. *Speech Commun.* **2012**, *54*, 583–600. [[CrossRef](#)]
18. Shahin, M.; Ahmed, B.; Parnandi, A.; Karappa, V.; McKechnie, J.; Ballard, K.J.; Gutierrez-Osuna, R. Tabby Talks: An automated tool for the assessment of childhood apraxia of speech. *Speech Commun.* **2015**, *70*, 49–64. [[CrossRef](#)]
19. Öster, A.M.; House, D.; Protopapas, A.; Hatzis, A. Presentation of a new EU project for speech therapy: OLP (Ortho-Logo-Paedia). In Proceedings of the XV Swedish Phonetics Conference (Fonetik 2002), Stockholm, Sweden, 29–31 May 2002; pp. 29–31.
20. Tan, T.S.; Ariff, A.; Ting, C.M.; Salleh, S.H. Application of Malay speech technology in Malay speech therapy assistance tools. In Proceedings of the IEEE 2007 International Conference on Intelligent and Advanced Systems, Kuala Lumpur, Malaysia, 25–28 November 2007; pp. 330–334.
21. PRADIA, misterio en la ciudad. Available online: <http://www.pradia.net> (accessed on 18 July 2018).
22. Adell, F.; Aguilar, L.; Corrales-Astorgano, M.; Escudero-Mancebo, D. Proceso de innovación educativa en educación especial: Enseñanza de la prosodia con fines comunicativos con el apoyo de un videojuego educativo. In Proceedings of the I Congreso Internacional en Humanidades Digitales, Valladolid, Spain, 17–19 April 2018.
23. Le, D.; Licata, K.; Persad, C.; Provost, E.M. Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2187–2199. [[CrossRef](#)]
24. Qin, Y.; Lee, T.; Feng, S.; Kong, A.P.H. Automatic Speech Assessment for People with Aphasia Using TDNN-BLSTM with Multi-Task Learning. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3418–3422.
25. Maier, A.; Haderlein, T.; Eysholdt, U.; Rosanowski, F.; Batliner, A.; Schuster, M.; Nöth, E. PEAKS—A system for the automatic evaluation of voice and speech disorders. *Speech Commun.* **2009**, *51*, 425–437. [[CrossRef](#)]
26. Kim, J.; Kumar, N.; Tsiartas, A.; Li, M.; Narayanan, S.S. Automatic intelligibility classification of sentence-level pathological speech. *Comput. Speech Lang.* **2015**, *29*, 132–144. [[CrossRef](#)]



27. Maier, A.; Hönig, F.; Hacker, C.; Schuster, M.; Nöth, E. Automatic evaluation of characteristic speech disorders in children with cleft lip and palate. In Proceedings of the Ninth Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008.
28. Laaridh, I.; Kheder, W.B.; Fredouille, C.; Meunier, C. Automatic prediction of speech evaluation metrics for dysarthric speech. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1834–1838.
29. Martínez, D.; Lleida, E.; Green, P.; Christensen, H.; Ortega, A.; Miguel, A. Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Trans. Access. Comput. (TACCESS)* **2015**, *6*, 10. [[CrossRef](#)]
30. Lee, H.Y.; Hu, T.Y.; Jing, H.; Chang, Y.F.; Tsao, Y.; Kao, Y.C.; Pao, T.L. Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 215–219.
31. Dunn, L.; Dunn, L.; Arribas, D. *Test de vocabulario en imágenes Peabody*; TEA: Madrid, Spain, 2006.
32. Corral, S.; Arribas, D.; Santamaría, P.; Sueiro, M.; Pereña, J. *Escala de Inteligencia de Wechsler para niños-IV*; TEA Ediciones: Madrid, Spain, 2005.
33. Raven, J.; Raven, J.C.; Court, J. *Test de matrices progresivas: Manual/Manual for Raven's progressive matrices and vocabulary scales* Test de matrices progresivas; Number 159.9. 072; J C Raven Ltd.: Buenos Aires, Argentina, 1993.
34. Martínez-Castilla, P.; Peppé, S. Developing a test of prosodic ability for speakers of Iberian Spanish. *Speech Commun.* **2008**, *50*, 900–915. [[CrossRef](#)]
35. 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. Human-Comput. Interact.* **2017**, *33*, 693–712. [[CrossRef](#)]
36. Ladd, D.R. *Intonational Phonology*; Cambridge University Press: Cambridge, UK, 2008.
37. 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; ACM: New York, NY, USA, 2013; pp. 835–838.
38. 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. [[CrossRef](#)]
39. 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. [[CrossRef](#)]
40. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [[CrossRef](#)]
41. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* **1997**, *30*, 1145–1159. [[CrossRef](#)]
42. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newslett.* **2009**, *11*, 10–18. [[CrossRef](#)]
43. Pardo, A.; Ruiz, M.Á. *SPSS 11: Guía para el análisis de datos*; Mc Graw Hill: Madrid, Spain, 2002.
44. Le, D.; Provost, E.M. Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
45. 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.
46. Kreiman, J.; Gerratt, B.R.; Precoda, K.; Berke, G.S. Individual differences in voice quality perception. *J. Speech Lang. Hear. Res.* **1992**, *35*, 512–520. [[CrossRef](#)]
47. 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. [[CrossRef](#)]

