Biomedical Signal Processing and Control 69 (2021) 102913

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Analysis of atypical prosodic patterns in the speech of people with Down syndrome

Mario Corrales-Astorgano [a], David Escudero-Mancebo [a], César González-Ferreras [a,*], Valentín Cardeñoso Payo [a], Pastora Martínez-Castilla [b]

[a] Departamento de Informática, Universidad de Valladolid, Valladolid, Spain
[b] Department of Developmental and Educational Psychology, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain

ARTICLE INFO

Keywords:
Down syndrome speech
Prosody assessment
PEPS-C
Sequential feature selection

ABSTRACT

The speech of people with Down syndrome (DS) shows prosodic features which are distinct from those observed in the oral productions of typically developing (TD) speakers. Although a different prosodic realization does not necessarily imply wrong expression of prosodic functions, atypical expression may hinder communication skills. The focus of this work is to ascertain whether this can be the case in individuals with DS. To do so, we analyze the acoustic features that better characterize the utterances of speakers with DS when expressing prosodic functions related to emotion, turn-end and phrasal chunking, comparing them with those used by TD speakers. An oral corpus of speech utterances has been recorded using the PEPS-C prosodic competence evaluation tool. We use automatic classifiers to prove that the prosodic features that better predict prosodic functions in TD speakers are less informative in speakers with DS. Although atypical features are observed in speakers with DS when producing prosodic functions, the intended prosodic function can be identified by listeners and, in most cases, the features correctly discriminate the function with analytical methods. However, a greater difference between the minimal pairs presented in the PEPS-C test is found for TD speakers in comparison with DS speakers. The proposed methodological approach provides, on the one hand, an identification of the set of features that distinguish the prosodic productions of DS and TD speakers and, on the other, a set of target features for therapy with speakers with DS.

1. Introduction

Prosody is an important component of speech communication because it is responsible for fundamental functions such as parsing the speech chain, expressing sentence type (e.g., declarative, interrogative, exclamatory or imperative), emotions or focus marking [1,2]. A low control of prosody or its inappropriate or atypical production can limit the options of speakers to integrate in society [3]. Such could be the case of people with intellectual disabilities in general and speakers with Down syndrome (DS) in particular. Although heterogeneity has been reported, a large number of individuals with DS present severe speech and language disorders [4,5]. When analyzing the speech of children and adults with DS, reduced communication effectiveness is found, and this is not only accounted for by articulatory impairments, but also by prosodic deficits [5–7].

Prosodic difficulties, observed throughout the lifespan [8], have been attested through different methods. Auditory-perceptual ratings of speech have shown pitch, intonation and rhythm atypicalities in children and adults with DS [6,9]. Stress and speech rate are also impaired in children and adolescents with the syndrome [9,7]. By adulthood, the prosody produced by individuals with DS contributes to their speech being perceived as atypical by typically developing (TD) listeners [10]. Acoustic analyses have also proved to be very informative, as they have shown how children with DS use a lower fundamental frequency when talking [8,11], while the opposite pattern is observed in adults [12,10,13]. Higher speech loudness and longer pauses have also been reported for adults with DS [10].

When speech prosody is assessed by considering communicative functions, difficulties have also been noted. During spontaneous communication, pre-school children show problems to express prosodic contours in interrogative sentences [11]. The "Profiling Elements of Prosody in Speech-Communication" (PEPS-C) test [14] has also been

* Corresponding author.
E-mail addresses: mcorrales@infor.uva.es (M. Corrales-Astorgano), descuder@infor.uva.es (D. Escudero-Mancebo), cesargf@infor.uva.es (C. González-Ferreras), valen@infor.uva.es (V. Cardeñoso Payo), pastora.martinez@psi.uned.es (P. Martínez-Castilla).

https://doi.org/10.1016/j.bspc.2021.102913
Received 26 December 2020; Received in revised form 6 May 2021; Accepted 5 June 2021
Available online 2 July 2021

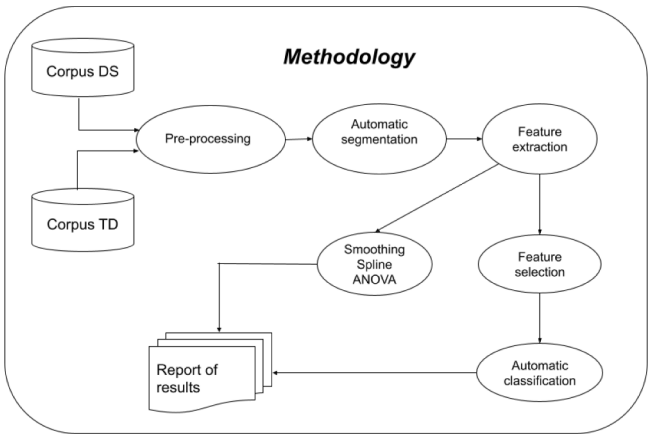

Fig. 1. Schematic flow diagram of the processes carried out in this work.

used to evaluate prosody in DS. This battery is specifically designed to assess the use of prosodic functions through ecological tasks and involves auditory-perceptual judgments. Through PEPS-C, it has been shown that children with DS score lower than TD peers of the same chronological age, both in perception and production prosody tasks, although performance on the former tasks is better than on the latter [15]. The same has been found for adolescents with the syndrome [16].

Speech has shown to be useful as a biomedical signal related to different syndromes and diseases such as Parkinson's [17–19], autism spectrum disorder [20], depression [21], Alzheimer's [22], ataxia [23], aphasia [24], dysarthria [25–27] or bipolar disorder [28], among others. This is also the case of DS [29,30,13,31]. Most of these works use prosodic characteristics of speech and provide information about speech rate, pausing, intonation and general communication skills of people that are generally affected by the disorder, as previously mentioned. Two types of works can be distinguished here: those aimed at assessing speech and detecting pathologies on the one hand, and those aimed at characterizing atypical speech on the other. When the goal of the investigation is the identification of atypical speech, different prosodic features related with pitch, energy and duration are added to the input of an automatic classifier that is trained with typical and pathological speech. Different types of classifiers have been used for this purpose: the support vector machine (SVM) [10,18,22,23,20], support vector regression (SVG) [21], neural networks (NN) [10,27,20] and decision trees [24,31]. In spite of the high classification rates that can be attained with these procedures, very little information about the relative importance of the different prosodic characteristics is obtained. In this paper, we use a gradient boosting tree (GBT) classifier focused on the classification of prosodic function instead of atypical speech identification. In doing so, we obtain information about the prosodic characteristics that are the most relevant for speakers with DS to accurately produce different prosodic functions.

Works on the characterization of atypical speech contrast with those of identification of anomalous speech in that the goal now is to better understand the impact of the disorder on speech [31,29,30,26,28]. These works are complex because acoustic features are difficult to be retrieved from pathological speech [25] and because prosody largely varies depending on the linguistic (the literal content of the message) and paralinguistic context (emotion, attitude…). The general methodological approach is to analyze differences on the use of prosodic characteristics among TD speakers and the population that is under study by controlling the type of message to be uttered by the informants: short pre-established sentences in [10,26,28], words in [32,33], groups of vocal-consonant-vocal productions in [29,30]; studies using spontaneous speech are the exceptions [13,11]. The PEPS-C test lies in between the two possibilities, representing a means to record oral utterances under controlled realistic conditions. In this work we use this test to analyze the production of the specific prosodic functions of affect, turn-taking and chunking. Additionally, we do not limit our study to provide a set of statistics with typical values of the different groups, but we use tools to visualize differences between the prosodic patterns that illustrate the problems of producing the prosodic functions under analysis.

Our goal in this work is threefold. First, we aim to ascertain whether, when successfully expressing a set of basic prosodic functions, prosodic features are produced in a different way by speakers with DS, in comparison to TD speakers. Second, we seek to investigate whether the eventual differences may cause problems for the automatic separation of the minimal categories of each prosodic function. Finally, we devise a visualization of prosodic patterns that reinforces the importance of data presentation techniques that allow separate comparisons of different prosodic functions, in order to obtain solid conclusions. To our knowledge, no prior study has tackled these aims.

To achieve the aims of the study we follow the methodology described in Fig. 1. The speech recordings obtained during the administration of the PEPS-C test to speakers with DS are compared with recordings obtained when TD speakers perform the same test (Section 2.1 presents a description of the corpus). Before recordings are parameterized by computing a set of prosodic features, a pre-processing task selects the appropriate audios for the study and they are segmented into phonemes (described in Section 2.2). Then, the prosodic function information is analyzed together with the acoustic information to obtain the most relevant prosodic features per function and per type of speaker (as detailed in subSection 2.3). A set of reports are obtained, including the output of a task of smoothing the F0 and intensity contours (as detailed in Section 2.4) and an automatic classification of the prosodic profiles. Section 3 examines the results and compares the informative power of the prosodic features with the evidence obtained when sequential feature selection is applied. Finally, we discuss the differences between the oral production of prosodic functions of TD speakers and speakers with DS, and how these differences can affect the appropriate discrimination of the minimal categories of a given prosodic function. We end the paper by suggesting how these observations can pave the way for future training activities.

2. Materials and methods

2.1. Corpus collection

The PEPS-C test, originally developed in English, has been designed to assess prosodic skills in individuals with speech and language disorders; as such, it has been used in prior studies with individuals with DS, as previously mentioned [34,16,35]. The test includes different prosodic function tasks, both in perception and production, to assess the functions of affect, turn-end, chunking and focus. For this study, we have used the Spanish version of the test [36], which is adapted from the English one [14]. In general, both the prosodic functions and forms used in the English and Spanish versions of PEPS-C are parallel [36], although some cross-linguistic differences have been reported. In [37], a study of the most representative features of the affect task of the PEPS-C test in the Spanish language is presented. The F0 contour was identified as the most representative feature to differentiate between liking and disliking and the pattern for the expression of disliking was shown to be different from that employed in English. The Spanish version of PEPS-C has proven to be a sound tool for the assessment of prosodic skills in Spanish-speaking individuals [38] and has been successfully used in individuals with intellectual and developmental disabilities [39].

As already pointed out, we have selected only the recordings of the affect, turn-end and chunking tasks for the PEPS-C test. Focus has not been included since, although it evaluates a well-attested function in Spanish (i.e., prefinal contrastive accent), there is a high variability in the way this function is addressed by prosodic means in Spanish-speaking individuals [36]. In the affect task, users look at a picture of a food item on the screen and have to produce the name of the food by

Table 1
Basic corpus statistics for the affect, turn-end and chunking tasks (BF means boundary in the first lexical item, BS means boundary in the second lexical item).

AFFECT	#Speakers	#Like	#Dislike	Total
TD speakers	41	449	188	637
DS speakers	33	272	102	374
Total	74	721	290	1011
TURN-END	**#Speakers**	**#Question**	**#Affirmative**	**Total**
TD speakers	41	328	318	646
DS speakers	33	135	212	347
Total	74	463	530	993
CHUNKING	**#Speakers**	**#BF**	**#BS**	**Total**
TD speakers	41	162	160	322
DS speakers	29	35	63	98
Total	70	197	223	420

expressing with their voice (i.e., by prosodic means) whether they like it or not. Then users have to confirm their opinion by choosing a corresponding picture of an emotional facial expression. In the turn-end task, users look at a picture of a boy on the screen, who is naming or offering a particular food. Then users have to produce the words as shown in the pictures: offering the food item to someone (question) or stating what is shown in a book being read (statement). In the chunking task, users look at some pictures on the screen and have to say what they are looking at in order of appearance. The key to this task is that users have to use prosody to distinguish segmental information by expressing a boundary after the first or the second lexical item, depending on what they are looking at on the screen, thus distinguishing between minimal pairs (e.g., "*barcopirata y agua*" vs. "*barco, pirata, y agua*" ["pirate-ship and water" vs. "pirate, ship, and water"]). A more detailed description of the PEPS-C test and the three tasks can be found in [36].

The recording process was carried out at different locations. The TD speakers corpus was recorded in a quiet room at each participant's house, and 41 users participated in this campaign (aged 18 to 51 years, mean 27). The DS corpus was recorded at a special education school of the Madrid DS Foundation and at two schools of special education located in Valladolid (33 participants aged 13 to 42 years, mean 22; and verbal mental age from 3.58 to 9.33 years, mean 6.45, assessed with the Peabody Picture Vocabulary Scale-III [40]). Both cities belong to the same linguistic region of Castilian Spanish. Audios were recorded at 22050 Hz using the built-in audio card of the computer where the software was installed. We used a Logitech PC Headset 960 USB in Valladolid, the laptop built-in micro for speakers with DS in Madrid and a Sony Lapel microphone for TD users. Although the audios were recorded in schools, background noise was controlled resulting in 53.6 dB (TD) and 42.3 dB (DS) signal-to-noise ratio (SNR). Differences in SNR between recordings at the two schools were below 6% on average, with 41.0 dB (Valladolid) vs. 43.6 dB (Madrid) SNR mean values. All SNR have been computed from the `stats` function results of the Sound eXchange (SoX) tool.[1]

All the recordings were evaluated by the therapist who led the test. Each of the PEPS-C tasks includes a minimal pair within a prosodic function. In the affect task, the minimal pair refers to the expression of liking and disliking; in the turn-end task it refers to statements versus questions; in the chunking task, the minimal pair refers to identifying either two or three semantic items with the same segmental information. To evaluate the correctness of the prosodic utterances produced by participants, the therapist judges which of the two communicative categories (i.e., minimal pair) is being expressed by the speaker. The therapist's judgments are then compared to the speaker's communicative intention (as requested by the pictures presented in the test) and this results into correct and wrong recordings: when the therapist judgment is the same as the communicative intention of the speaker, the recording is considered as correct; the opposite holds for the wrong recordings. In this work, the recordings evaluated as correct were the focus of analysis. In addition, all audio recordings were processed with the aim of removing the defective ones (errors in wording, corrupted recording or noise disturbance). For the chunking task, half the recordings were selected. Specifically, only the compound-noun items were analyzed, as they are easier to process automatically. One of the therapists re-evaluated a subset of 20% of the previously evaluated samples, both by her and by the other therapist. Thus, an inter- and intra-reliability tests were performed to assess the degree of consistency between the evaluations, resulting in a Kappa index of 0.790 for intra-annotator evaluations and 0.784 for inter-annotator evaluations.

Table 1 shows the contents of the resulting corpus after the filtering process had been applied. Note that the number of recordings reflects the utterances perceived as correct by the therapist (after the removal of the defective ones). As expected from the literature [34], the number of correct utterances is lower in the group with DS compared to the TD group. It should also be noted that four participants with DS are missing in the chunking task. The production of this function is more demanding and is thus acquired later than affect and turn-end in typical development [36]. This explains why the task can be more difficult for some individuals with DS. The same has been reported in prior work with English-speaking individuals with the syndrome [15].

2.2. Prosodic features extraction

The steps described in this section were applied to each audio recording of the speech corpus in order to extract a set of prosodic features. Firstly, the WebMaus web service [41] was employed to automatically generate the segmentation of each audio file. To do this, the `.wav` file and the phonetic transcription in BPF format [42] were provided to the web service. Secondly, a PitchTier and an IntensityTier were generated for each recording using the Praat software [43]. The algorithm used to extract the pitch estimations was based on an acoustic periodicity detection on the basis of an accurate autocorrelation method proposed in [44]. To ensure the quality of the fundamental frequency (F0) contours, pitch outliers and jumps were reduced using the procedure proposed in [45]. Thirdly, as in [46], we removed the speaker dependence on F0 contours and used a perceptual scale by transforming the F0 values to semitones with respect to the mean F0 of the speaker, using the formula:

$$F0' = 12 * log_2\left(\frac{F0}{\langle F0\rangle_{Speaker}}\right) \quad (1)$$

where $\langle F0\rangle_{Speaker}$ is the average of the F0 across all recordings of each speaker. Then, the IntensityTier values were normalized per user by subtracting the intensity mean from each value and dividing the result by the standard deviation. The initial and final silence intervals were excluded from this parameterization procedure. The silence and sounding intervals were calculated using the default values of Praat.

In previous works [10,46], we used openSmile [47] to compute an extensive set of prosodic features. Given the simpler nature of the tasks in the PEPS-C test (isolated words in the affect and turn-end tasks and short utterances of three to four words in the chunking task), we selected a reduced set of features which we had successfully used in previous works for similar experiences [48,37]. For F0 and energy contours, we use features that reflect the temporal evolution or the shape of the prosodic pattern, measuring mean values, ranges, slopes and declinations. For duration, we put the focus on the impact of the inner pause in the chunking task and add features that reflect abnormal changes in rhythm [49].

The average (*xMean*), standard deviation (*xSd*), range (*xRange*), difference between maximum and average (*xMaxavg*), difference between average and minimum (*xMinavg*), average of the rising

[1] http://sox.sourceforge.net/

(*xRisingMean*) and of the falling (*xFallingMean*) slopes, were applied to pitch and intensity files (*x* must be replaced by *f0* or *i* when referring to F0 or intensity, respectively). To calculate the rising and falling means for F0 and intensity, each F0 or intensity value across the signal was compared to the next one and the difference was added to the set of rising (difference above 0) or falling (difference equal to or below 0) slopes, which were then averaged, respectively. Two additional features have been included to represent the temporal evolution of the F0 contours, corresponding to declination and excursion, as defined in [37]: *f0Declination* is computed as the difference between the last value and the maximum value divided by the time interval between these values, and *f0ConEx* is computed as the difference between the first value and the maximum value plus the range.

Related to the duration of the segments, the following features were extracted: the rate of speech (*ROS*) as the number of phones per second, the vocalic intervals ratio (*VIR*) as the sum of the lengths of vocalic intervals divided by the total duration of the sentence (excluding pauses), the standard deviation of the duration of vocalic intervals (*dV*) and of consonant intervals (*dC*), the standard deviation of vocalic (*varV*) and consonant (*varC*) interval duration divided by the mean vocalic or consonant duration within the utterance. In addition, two forms of the pairwise variability index (PVI) [50] were computed (raw PVI and normalized PVI), with variants for vocalic and consonant segments:

$$rPVI = 100 \times \frac{\sum_{i=1}^{N-1} |d_i - d_{i+1}|}{N-1} \tag{2}$$

$$nPVI = 100 \times \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|d_i - d_{i+1}|}{(d_i + d_{i+1})/2} \tag{3}$$

where N is the number of segments and d_i the duration of segment i. With these, four utterance-level features are extracted: rPVI.V, nPVI.V, rPVI.C, nPVI.C, where V and C refer to vocalic and consonantal phonemes respectively, and the prefixes r and n mean raw and normalized.

Additionally, three duration features were extracted in the chunking task recordings: *ROSvariation*, computed as the ratio between the ROS of the first word in the sentence divided by the ROS of the other words in the same sentence; the length of the pause that follows the first word of the sentence (*pauseLength*), the maximum duration of the vocalic phones in the sentence z-normalized across the speaker group (*maxVocalLength*). These features are inspired in the works on chunking analysis presented in [51].

2.3. Feature selection and classification

The 29 prosodic features described in the previous section have been analyzed to remove the most closely correlated ones (Pearson correlation > 0.9). With the remaining set of features, a ranking is built by taking into account the informative value of each feature, computed with the information gain (IG) as [52]:

$$IG(F, PF) = H(F) - H(F|PF) \tag{4}$$

$H(F)$ being the entropy of the prosodic function F in the corpora and $H(F|PF)$ the conditional entropy of the function F, taking into account the discretized values of each prosodic feature PF.

$$H(F) = -\sum_{i=1}^{n} p(f_i) \log p(f_i) \tag{5}$$

$$H(F|PF) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(f_i, pf_j) \log \frac{p(f_i, pf_j)}{p(pf_j)} \tag{6}$$

where $n = 2$, as the prosodic function F can only take two possible values. As the prosodic feature PF is a continuous variable, a discretization process is applied [53] and m is the resulting number of different values. The obtained rankings are then analyzed to compare the DS and TD data sets in terms of each studied prosodic function (see Section 3).

Table 2
Ranking of features for the three tasks and two groups of speakers, derived from the information gain metric. Only features with non-zero gain are shown.

TD speakers		DS speakers	
Feature	InfoGain	Feature	InfoGain
AFFECT			
f0Mean	0.3738	iMean	0.2469
f0Minavg	0.2442	f0Mean	0.2394
f0Range	0.2262	f0Declination	0.1309
f0Declination	0.2097	f0Minavg	0.1025
iMean	0.0988	f0RisingMean	0.0941
f0RisingMean	0.0599	f0Range	0.0878
f0FallingMean	0.0596	f0FallingMean	0.0741
iRange	0.0476		
iMinavg	0.0446		
rPVI.V	0.0235		
TURN-END			
f0Range	0.5988	f0Declination	0.1317
f0Mean	0.4687	f0Range	0.104
f0Minavg	0.3128	f0Mean	0.0709
f0Declination	0.1999	f0RisingMean	0.0652
f0RisingMean	0.1663	ROS	0.0593
iMean	0.0757	rPVI.V	0.0479
iMinavg	0.0549	VIR	0.0408
f0ConEx	0.0541	f0Minavg	0.036
f0FallingMean	0.0377		
rPVI.V	0.0232		
iMaxavg	0.0226		
CHUNKING			
pauseLength	0.6044	ROSvariation	0.252
maxVocalLength	0.5012	maxVocalLength	0.175
ROSvariation	0.3379	pauseLength	0.128
ROS	0.2605	iSd	0.103
f0Declination	0.107		
iMean	0.0958		
iMaxavg	0.0956		
f0Mean	0.0708		
rPVI.V	0.07		
rPVI.C	0.0652		
dC	0.0575		
iSd	0.0531		
iRange	0.0473		
f0FallingMean	0.0437		
varC	0.0415		
f0Minavg	0.0409		

In parallel, we apply sequential forward feature selection (SFS) to select the most relevant features. The selection algorithm uses the accuracy of a GBT classifier as the evaluation measure. The order of features selected provides a second ranking that is expected to be consistent with the one obtained from the information gain analysis; however, in this case, once a feature has been selected, others highly correlated with the ones already selected lose importance.

The selection procedure is used for each task in three different scenarios: 1) using the recordings of TD speakers in cross validation (referred to as *TD-TD*); 2) using the recordings of speakers with DS in cross validation (*DS-DS*); 3) using the recordings of TD speakers as training data and the recordings of speakers with DS as testing data (*TD-DS*). The goal of obtaining relevant features in these three scenarios is to find differences that could have to do with both the conventional way to produce the prosodic functions (scenario 1) and the particular way to do so by speakers with DS (scenario 2), as well as the problems that could be found when the conventional classification procedure (scenario 3) is applied. We have used gradient boosting trees classifiers for each of the three tasks. The aim of these classifiers is to automatically identify the prosodic function of the recordings in the affect task (like vs dislike), turn-end task (question vs affirmative) and chunking task (boundary in the first lexical item vs boundary in the second lexical item). Scenarios 1 and 2 used ten fold cross validation in each group of speakers (TD or DS).

Table 3
Ranking of features using sequential forward feature selection for different tasks and different train-test datasets. For each task and datasets, accuracy values represent the classification results obtained when the feature in the corresponding row is added to the features of previous rows.

	TD-TD		DS-DS		TD-DS	
	Feature	Accuracy	Feature	Accuracy	Feature	Accuracy
AFFECT	f0Mean	80.06	iMean	78.07	f0Mean	78.88
	f0Range	88.54	f0Mean	84.48	f0Range	85.83
	f0ConEx	91.04	f0Range	87.16	f0ConEx	87.43
	f0Declination	92.30	iFallingMean	88.50	iRisingMean	88.24
	iMaxavg	92.78	f0RisingMean	88.77	f0Minavg	87.97
TURN-END	f0Range	87.14	f0Declination	62.60	f0Range	68.88
	f0Mean	91.77	f0Range	73.89	f0ConEx	72.05
	iMaxavg	92.39	f0Mean	75.33	f0RisingMean	74.64
	f0ConEx	93.17	iMean	77.61	rPVI.C	74.64
	iRisingMean	93.63	varC	76.46	f0Declination	75.79
CHUNKING	pauseLength	89.41	pauseLength	68.56	maxVocalLength	71.43
	maxVocalLength	92.53	ROSvariation	69.56	nPVI.C	73.47
	f0Declination	94.72	iSd	77.78	nPVI.V	73.47
	dC	95.34	f0RisingMean	78.78	rPVI.C	74.49
	varV	95.03	rPVI.C	76.67	rPVI.V	74.49

R tools [54] were used to detect correlated features. Weka tools [55] were used to compute the information gain and build the prosodic feature rankings. The software libraries scikit-learn [56] and mlxtend [57] were used to train and test the classifiers.

2.4. Smoothing Spline ANOVA for F0 and intensity contour

A smoothing spline ANOVA was conducted on the F0 and Intensity data to generate the smoothing spline fit for the three tasks and the two groups of speakers. This technique has been used successfully in other works to compare the F0 and Intensity contours [58,59]. The R package gss [60] was used to calculate the smoothing spline with 95% confidence intervals and the R `ggplot2` package [61] was used to plot the splines. Due to the different time intervals of the F0 and intensity files, a min–max scaling was applied to each file with the aim of obtaining all values in the interval [0,1]. Time values (x) were scaled to normalized $x' \in [0, 1]$:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{7}$$

3. Results

Table 2 shows the ranking of features in terms of their informative power to classify affect, turn-end and chunking, respectively. For the affect task, F0 related features stand out with respect to intensity and duration for TD speakers; intensity seems to play a more important role for speakers with DS, while duration features disappear from the ranking. As for the turn-end task, features belonging to the three domains (F0, intensity and duration) appear in both rankings, with higher relevance for the features of the F0 domain. Concerning the chunking task, the three most relevant features are the same for both groups of speakers; other features seem to be relevant in the TD case, but they disappear from the ranking in the DS case.

Information gain values are higher for TD speakers in the three tasks. Compared with the values for the DS group, the most relevant feature is 0.13 points higher in the affect task, 0.46 points higher in the turn-end task and 0.35 points higher in the chunking task. Also, there are more features with *InfoGain* > 0 for TD speakers than for speakers with DS, in all tasks.

Table 3 shows the five selected features obtained using SFS. High classification rates (over 90% accuracy) are obtained for TD speakers using just a few features, for the three prosodic functions. Poorer results are obtained for speakers with DS, showing a better performance for the affect function. Results for the turn-end function and for chunking are below 79% accuracy.

Table 4
Confidence intervals (95%) of different features for different tasks. L means like, D means dislike, Q means question, AF means affirmative, BF means boundary in the first lexical item and BS means boundary in the second lexical item. The intervals were computed using scipy library [62].

	TD		DS	
AFFECT	**L**	**D**	**L**	**D**
f0Mean	(0.34, 0.75)	(−3.94, −3.19)	(0.01, 0.57)	(−3.55, −2.64)
f0Range	(12.02, 12.75)	(7.53, 8.89)	(10.06, 11.11)	(6.86, 8.8)
iMean	(0.83, 0.88)	(0.59, 0.67)	(0.57, 0.65)	(0.11, 0.24)
TURN-END	**Q**	**AF**	**Q**	**AF**
f0RisingMean	(0.84, 0.94)	(0.67, 0.86)	(0.42, 0.51)	(0.43, 0.73)
f0Range	(12.99, 13.81)	(6.09, 7.04)	(8.38, 9.8)	(5.93, 7.11)
f0Mean	(0.91, 1.47)	(−2.66, −2.02)	(−0.48, 0.28)	(−1.74, −1.08)
f0Declination	(−76.34, −52.0)	(−26.66, −18.14)	(−42.91, −23.88)	(−23.58, −18.49)
CHUNKING	**BF**	**BS**	**BF**	**BS**
pauseLength	(0.23, 0.32)	(−0.0003, 0.0009)	(0.21, 0.44)	(0.04, 0.16)
maxVocalLength	(1.22, 1.66)	(−0.66, −0.43)	(0.32, 0.92)	(−0.42, 0.06)
ROSvariation	(0.85, 0.92)	(1.31, 1.42)	(0.75, 0.98)	(1.25, 1.58)

The order of the selected features reported in Table 3 differs from the rankings presented in Table 2 because, once a feature has been selected by the classification algorithm, other features correlated with the one selected can lose importance.

Table 4 shows the 95% confidence intervals of the most informative features in the classification results for each task and speaker group (Table 3). In addition, the Mann–Whitney U test was used to check if there were statistically significant differences between the minimal pairs in each task and between groups of speakers. The results of the Mann–Whitney test can be seen in Table 5. Regarding the affect task, there are statistically significant differences ($p-value < 0.05$) in all cases except in TD-L vs. DS-L (f0Mean), in TD-D vs. DS-D (f0Mean) and in TD-D vs. DS-D (f0Range). In the turn-end task, there are statistically significant differences in all cases except for TD-AF vs. DS-AF (f0Range) and TD-AF vs. DS-AF (f0Declination). Finally, in the chunking task, there are statistically significant differences in all cases except for TD-BF vs. DS-BF (pauseLength and ROSvariation) and TD-BS vs. DS-BS

Table 5
Mann–Whitney results for each feature and case. The values in bold are the ones with $p-value < 0.05$. L means Like, D means Dislike, Q means Question, AF means Affirmative, BF means Boundary in the first lexical item and BS means Boundary in the second lexical item.

AFFECT	TD-L vs. TD-D	DS-L vs. DS-D	TD-L vs. DS-L	TD-D vs. DS-D
f0Mean	**0.0**	**0.0**	0.15	0.06
f0Range	**0.0**	**0.0**	**0.0**	0.16
iMean	**0.0**	**0.0**	**0.0**	**0.0**
TURN-END	**TD-Q vs. TD-AF**	**DS-Q vs. DS-AF**	**TD-Q vs. DS-Q**	**TD-AF vs. DS-AF**
f0RisingMean	**0.0**	**0.0**	**0.0**	**0.0**
f0Range	**0.0**	**0.0**	**0.0**	0.87
f0Mean	**0.0**	**0.0**	**0.0**	**0.0**
f0Declination	**0.0**	**0.02**	**0.01**	0.19
CHUNKING	**TD-BF vs. TD-BS**	**DS-BF vs. DS-BS**	**TD-BF vs. DS-BF**	**TD-BS vs. DS-BS**
pauseLength	**0.0**	**0.0**	0.99	**0.0**
maxVocalLength	**0.0**	**0.0**	**0.0**	**0.01**
ROSvariation	**0.0**	**0.0**	0.19	0.53

(ROSvariation).

In the following subsections we analyze the results obtained per type of prosodic function.

3.1. Affect task

In the affect task, the classifier trained and tested with samples of TD speakers obtains a rate of 91.04% using only three features of the F0 domain (Table 3). In the case of the samples of DS speakers, the features of F0 also have an important role, but the introduction of a feature related to intensity average is indicative that the intensity replaces or complements F0. When the samples of DS speakers are tested with the classifier trained with samples of TD speakers, the classifier obtains classification rates similar to the previous case, but in this case, the features of F0 are again the most relevant for the classification.

The results of Table 4 show that the mean values of the fundamental frequency in the expression of the feeling of liking are higher than the mean of each user and close to 0; while, in the expression of the feeling of disliking, the values are less than the average of each user and are far from 0. This happens for both groups of speakers. The higher mean values of the f0Range in the fundamental frequency for the case of liking, compared to the case of disliking, indicate that there is a greater variation of the F0 values in the case of liking for both groups of speakers.

The results of the statistical tests carried out for these features (Table 5) show that there are no significant differences between groups in the samples of liking (f0Mean) and those of disliking (f0Mean and f0Range), while they do exist when the samples of liking and disliking are compared within each group. In the case of disliking, there are statistical and significant differences in all cases except in TD-D vs. DS-D (f0Mean and f0Range). However, the p-value is 0.06 for f0Mean, which indicates that, although there are no significant differences, there is a trend towards small differences in the expression of this function.

3.2. Turn-end task

As in the affect task, the features related to the fundamental frequency are the most informative when classifying between Q and AF (Table 3). In the case of the TD group, a rate of 91.77% is reached only using two features, f0Range and f0Mean. These two features indicate that the range and the mean of the fundamental frequency are different for the Q samples and the AF samples. In addition, a feature related to the difference between the maximum and average intensity appears in the third position (iMaxavg), but this feature increases the classification rate by less than one percent.

The results of Table 4 show that the differences between the minimal and maximal values (range) of the fundamental frequency in the expression of questions are higher than the range of the fundamental frequency in the expression of affirmative sentences for both groups. However, these differences are higher for TD speakers than for speakers with DS. These results indicate that TD speakers produced the question sentences with more differences in the use of the fundamental frequency than the speakers with DS.

The results of Table 5 show that there are statistically significant differences in all cases except for TD-AF vs. DS-AF (f0Range) and TD-AF vs. DS-AF (f0Declination), which indicates that the production of affirmative sentences is more similar than the production of question sentences when both groups of speakers are compared.

3.3. Chunking task

In the chunking task, the features which convey the highest classification rates are those related to duration for both groups of speakers (Table 3). The duration of the pause that follows the first word of each sentence (pauseLength) and the maximum duration of the vowel phonemes (maxVocalLength) are the two most informative features and allow a rate of 92.53% to be obtained in the TD group (Table 3). In the DS group, the pauseLength feature is also the most informative, and maxVocalLength is replaced by ROSvariation. However, in the case of the DS group, the classification rates obtained are lower than in the TD case.

The confidence intervals and the statistical tests presented in Tables 4 and 5 show that there are differences in the features pauseLength, maxVocalLength and ROSvariation in both groups between the different types of sentences, but the differences in pauseLength and maxVocalLength between BF and BS in TD speakers are higher than those in speakers with DS. The statistical differences observed in the TD-BS vs. DS-BS case, but not in the TD-BF vs. DS-BF case, show that speakers with Down syndrome can produce the boundary in the first lexical item in a similar way to TD speakers, but this is not the case in the production of the boundary in the second lexical item.

4. Discussion

4.1. Differences between TD and DS prosodic patterns

Experimental results show clear differences between the production of prosodic features of TD speakers and speakers with DS. The present discussion will be supported by plots of prosodic patterns (Fig. 2) representing the smoothing spline ANOVA (F0 and Intensity) of the word *yogur* (affect task), the word *uvas* (turn-end task) and the sentence *barco pirata y agua* (chunking task) of the two speaker groups. The contours are similar in the different items of the different tasks, so we have selected one item as an example with the aim of illustrating the contour pattern of the F0 and intensity curves.

In the affect task, the contours are similar in shape in both groups of speakers, not only in the F0 curves, but also in the intensity curves. In the like case, the F0 curve has an inverted-u form and the intensity curve has a high rise in the beginning and a fall at the end, higher in the TD group than in the DS group. The U-pattern is prototypical for this kind of productions, as already reported in [37]. In line with our observations, [63] reports findings of increases in the F0 mean, F0 ranges and F0 variability and mean energy related to *joy*. Concerning the dislike case, [63] reports a decrease in the mean, F0 range and mean energy and downwards-directed F0 contours on the acoustic features related to *sadness*, while our corpus shows that the F0 curve is much flatter than in the like case, with similar values in the TD group and the DS group, while the intensity curve has a higher fall in the DS group than in the TD group. Numerically, these differences have already been contrasted in Tables 4 and 5: f0Mean, f0Range and iMean are lower in DS than in TD,

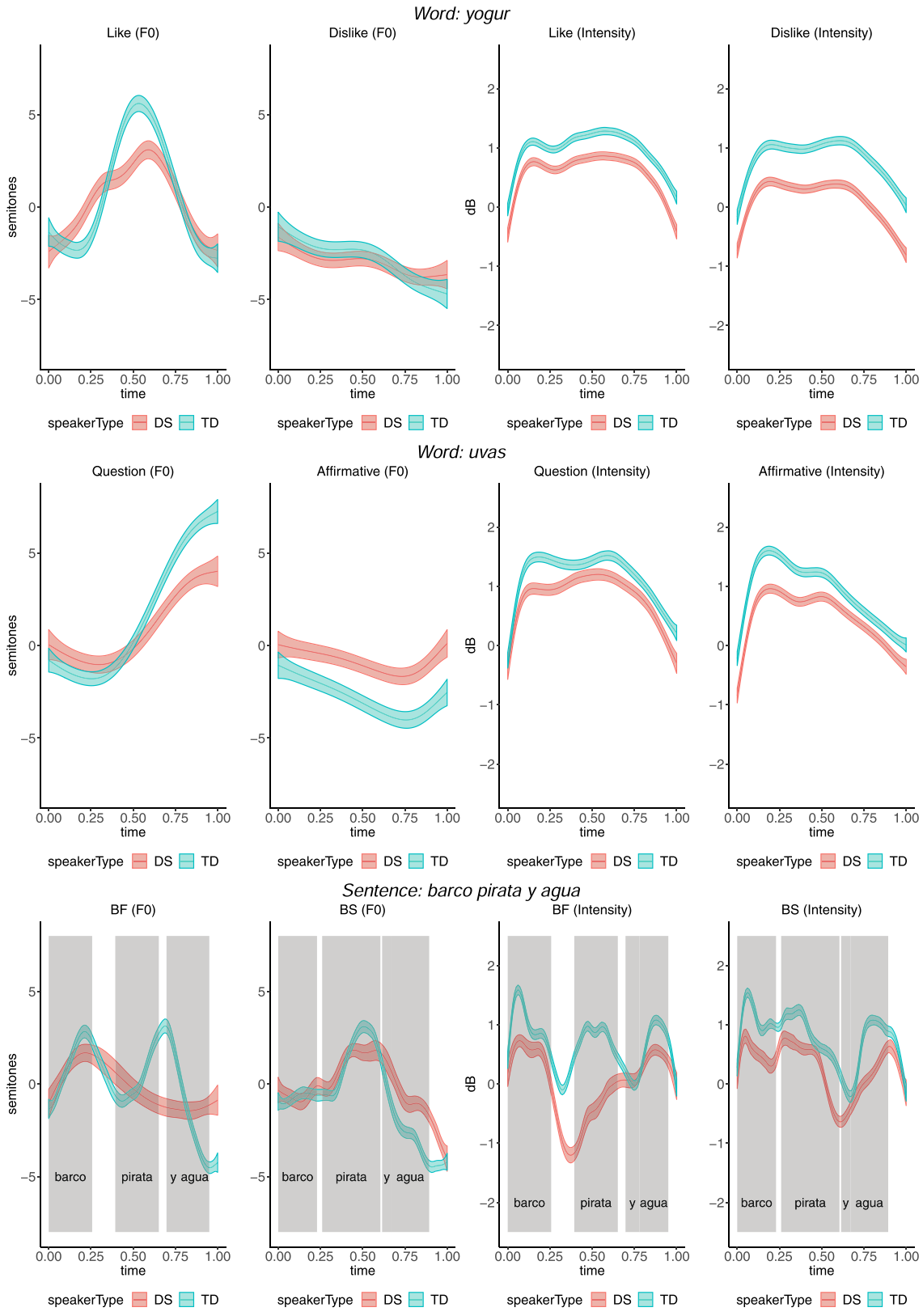

Fig. 2. Typical F0 and intensity patterns (with 95% confidence interval computed with smoothing spline ANOVA described in Section 2.4) of the different group of speakers and prosodic functions under analysis, corresponding to the oral productions of the word *yogur* (yoghurt) in the affect task, the word *uvas* (grapes) in the turn-end task, the sentence *barco pirata y agua* (pirate ship and water) in the chunking task.

both in L and D categories. It seems that speakers with DS produce correct affect patterns with a lower intensity, affecting both energy and F0.

In the turn-end task, the F0 contours have a final high rise in questions (with a higher slope in the TD group) and a flatter contour in the affirmative case. The final rise is a prototypical pattern that allows for the difference between declarative and question sentences to be marked [64,65]. As reported for affect, the differences between DS and TD speakers are related to the extent of the pattern with higher values in the case of TD speakers in all the cases except for affirmative sentences (F0), where speakers with DS have higher values than TD speakers. Numerically, these differences can be seen in Table 4 in that, for questions, f0RisingMean, f0Mean and f0Range are higher values in the TD group (the differences being statistically significant, as shown in the TD-Q vs DS-Q comparison in Table 5). In the TD-AF vs. DS-AF case, there are also significant differences in f0RisignMean and f0Mean; in this last case, the values of speakers with DS are higher than those of TD speakers. A similar behavior is observed in the energy patterns with respect to the one observed in the affect task, with lower energy values for the speakers with DS; however, in these cases, no significant differences are observed.

The chunking task seems to be the most problematic case for speakers with DS. The F0 curve has two consecutive rises in the BF case in the TD group, while the DS group does not show this pattern. In the BS case, there is a unique rise in the F0 curve in both groups. A similar effect can be seen in the intensity curve, with an initial fall in the BF case, and without an initial fall in the BS case for both groups. The role of the position and number of intonation groups seems to be determinant. The typical pattern for chunking in Spanish is using a short pause that is frequently preceded by an F0 inflection [64,66,48]. TD speakers mark two intonation boundaries in the BF case and only one in the BS case (two peaks versus only one). This is not the case in speakers with DS, who seem to mark only one. Numerically, this fact is observed in the feature maxVocalLength of the case TD-BF vs. DS-BF, as the intermediate prosodic boundaries are also marked with a lengthening that is clear for TD speakers, but not for speakers with DS. Once again, in both cases, the intensity is lower in the DS group than in the TD group.

To sum up, while the expression of chunking clearly differed between groups, similar patterns were found in the TD and DS groups for the affect and turn-end tasks (with lower values of either F0 or intensity in the case of DS). However, classification accuracy is lower in speakers with DS than in TD speakers. This can be explained by considering that the automatic classification of the prosodic function for a given task is more difficult in speakers with DS because the differences between the values of the prosodic features within a minimal pair are smaller than those in the TD speakers, as can be seen in Table 4 and Fig. 2. This will be discussed below.

4.2. Less interclass separation in speakers with Down syndrome

The results obtained in the classification tasks seem to indicate that the separation of the different classes is greater in TD speakers than in speakers with DS because the classification rates are better in the TD-TD case than in the case of DS-DS: Table 3 shows that while the TD-TD case exceeds 90% accuracy in all functions, the classification rate with DS samples drops considerably. This fact affects the three prosodic functions, but it is less pronounced in the case of affect, where classification rates of 88% are reached in the DS-DS case.

In Section 3, we saw that there were similarities in both the DS and TD patterns in the Affect and Turn-end cases, but this was not the case in the Chunking with the BF function. This fact is also reflected in the results of Table 3. The classification rates obtained by the classifier trained by TD data and tested with DS data are similar to those obtained by the classifier trained by DS data and tested with DS data, although the differences are bigger for the chunking task. In this task, the difference between the best DS-DS rate (78.78%) and TD-DS rate (74.49%) is 4.29 percentual points, and it is 0.53 and 1.82 percentual points in the case of affect and turn-end, respectively.

Analyzing each function independently, we find (see Table 4) that, for the affect function, the differences in the average values of the features are bigger in the TD-L vs. TD-D case than in the DS-L vs. DS-D case, for all the features except in the case of *iMean*, where the separation is greater in the energy graphs for speakers with DS. This points to the fact that speakers with DS do not reach the ranges of variation in the F0 curve that TD speakers do, but that they could complement this information with a different energy modulation. The use of this feature seems to be more effective when expressing the prosodic function of affect than when expressing the other two prosodic functions.

As for the turn-end function, we see in Table 4 that the differences in the DS-Q vs. DS-AF are smaller than in the TD-Q vs. TD-AF ones. Although these distances are statistically significant (Table 5), they are not so effective for the automatic classification of the minimal pairs. In this case, there is also less inflection in the curve of F0 and the compensation of energy does not seem to be enough. In the case of the chunking function, we also see that the poor production of the pauses of the speakers with DS reported in the previous section makes the differences in the DS-BF vs. DS-BS excessively small, causing the task of the chunking type function identification to be very difficult in the case of speakers with DS.

4.3. Implications for intervention

Deriving implications for clinical practice related to how to expressively use prosody by people with DS is the goal of many studies [16]. The information provided by the experimental procedure opens the way to the always challenging task of preparing personalized exercises aimed at improving the communication skills of speakers with DS. First, it is important to focus on the prosodic features that are less discriminant within a given prosodic function; in a second step, it would be important to design exercises that allow speakers with DS to get closer to the performance of productions made by TD individuals.

Although designing specific training exercises is outside the scope of this paper, we consider that, of the two aforementioned objectives, the separation between prosodic categories within the same function, on the one hand, and an approach to typical speakers, on the other, the most important is the first. Getting speakers with DS to clearly distinguish the categories within the same prosodic function is more important than the fact that their production should be as similar as possible to that of TD speakers, because the ultimate goal of oral communication is to make themselves understood and, for that aim, it should not be necessary to precisely follow a given standard. Different speakers would assume that there are oral productions that can be different and that all can be valid. It is especially important to take this into account, because we are facing a group of users with muscular hypotonia problems [67], which makes it difficult for them to reach certain inflections of F0. They also have short-term memory limitations [68], and these can make it difficult for them to correctly produce pauses in a given sequence.

5. Conclusions

The methodology presented in this work has allowed us to show that differences can be detected between the way prosodic functions are produced by TD speakers and those with DS. We have shown that the ability to express differentiated prosodic functions (as presented in the PEPS-C test) is lower in speakers with DS. In addition, the results provide information concerning the prosodic features that better discriminate the categories of prosodic function, and also, to what extent the use of these prosodic features differs between TD speakers and speakers with DS.

There are differences in the way speakers with DS produce prosodic functions, and those differences depend on the prosodic function itself: speakers with DS make more use of energy than TD speakers in order to produce affect; they have problems to articulate declination to produce

Table A.6
Description of the acoustic features used in the paper.

Feature	Description
PITCH (F0) DOMAIN	
f0Mean	The average of the fundamental frequency (semitones)
f0Sd	The standard deviation of the fundamental frequency (semitones)
f0Range	The difference between maximum and minimum of the fundamental frequency (semitones)
f0Maxavg	The difference between maximum and average of the fundamental frequency (semitones)
f0Minavg	The difference between average and minimum of the fundamental frequency (semitones)
f0RisingMean	Average of the rising segments of the fundamental frequency (semitones)
f0FallingMean	Average of the falling segments of the fundamental frequency (semitones)
f0Declination	The difference between the last value and the maximum value divided by the time interval between these values (semitones)
f0Contour	The difference between the first value and the maximum value plus the range (semitones)
INTENSITY DOMAIN	
iMean	The average of the energy (decibels)
iSd	The standard deviation of the energy (decibels)
iRange	The difference between maximum and minimum of the energy (decibels)
iMaxavg	The difference between maximum and average of the energy (decibels)
iMinavg	The difference between average and minimum of the energy (decibels)
iRisingMean	Average of the rising segments of the energy (decibels)
iFallingMean	Average of the falling segments of the energy (decibels)
TEMPORAL DOMAIN	
ROS	Rate of speech, the number of phones per second
VIR	The vocalic intervals ratio, as the sum of the lengths of vocalic intervals divided by the total duration of the sentence, excluding pauses (seconds)
dV	The standard deviation of the duration of vocalic intervals (seconds)
dC	The standard deviation of the duration of consonant intervals (seconds)
varV	The standard deviation of vocalic interval duration divided by mean vocalic duration within the utterance (seconds)
varC	The standard deviation of consonant interval duration divided by mean consonant duration within the utterance (seconds)
rPVI.V	First form of the Pairwise Variability Index with variants for vocalic segments (seconds)
rPVI.C	First form of the Pairwise Variability Index with variants for consonant segments (seconds)
nPVI.V	Second form of the Pairwise Variability Index with variants for vocalic segments (seconds)
nPVI.C	Second form of the Pairwise Variability Index with variants for consonant segments (seconds)
ROSvariation	The ratio between the ROS of the first word in the sentence divided by the ROS of the other words in the same sentence (seconds)
pauseLength	The pause length of the pause that follows the first word of the sentence (seconds)
maxVocalLength	The maximum of the duration of vocalic phones in the sentence z-normalized across the speaker group (seconds)

declarative sentences; and, in chunking, the variation of speech rate is higher in speakers with DS than in TD individuals.

While for TD speakers the acoustic features allow the minimal pairs that are compared for each prosodic function to be accurately separated, the automatic classifier performance decreases in the case of speakers with DS. Although the shorter distances between pairs of each prosodic function for DS speakers do not prevent the identification of the intended prosodic category by the therapist, within a given prosodic function, we have found that, in general terms, speakers with DS separate prosodic categories less than TD speakers. In the case of affect, the separation is better than in the case of turn-end and chunking. However, for accurately producing affect, speakers with DS seem to follow an alternative strategy: using energy to complement F0 excursions.

The proposed methodology gives some cues about the prosodic features that need to be trained in speakers with DS, both for distinguishing more clearly different communicative functions and for getting closer to the typical production patterns. This opens a path to prepare specific exercises for speakers with DS to be trained in prosodic skills with the goal of improving the production of prosodic functions.

CRediT authorship contribution statement

Mario Corrales-Astorgano: Conceptualization, Software, Investigation, Data curation, Writing - original draft, Visualization. **David Escudero-Mancebo:** Methodology, Software, Resources, Writing - original draft, Funding acquisition. **César González-Ferreras:** Conceptualization, Software, Investigation, Writing - original draft. **Valentín Cardeñoso Payo:** Methodology, Validation, Investigation, Resources, Writing - review & editing, Funding acquisition. **Pastora Martínez-Castilla:** Investigation, Resources, Writing - review & editing.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgments

This work has been supported in part by the Ministerio de Economía y Competitividad and the European Regional Development Fund FEDER under Grant TIN2017-88858-C2-1-R and by the Consejería de Educación de la Junta de Castilla y León under Grant VA050G18.

The authors would like to thank the students and therapists of the special education centers “Pino de Obregón” and “Asociación Down Valladolid (ASDOVA)”, in Valladolid, and the “Fundación Síndrome de Down Madrid (Down Madrid)” for their valuable participation and collaboration in the experimental activities carried out for this work. We would also like to thank Valle Flores, Jesús Gómez, Yolanda Martín and Alfonso Rodríguez.

Appendix A. Description of the features

The acoustic features used in the paper are shown in Table A.6. Section 2.2 describes the feature extraction procedure and tools used to calculate them.

References

[1] P. Roach, English phonetics and phonology fourth edition: A practical course, Ernst Klett Sprachen, 2010.
[2] S.J. Peppé, Why is prosody in speech-language pathology so difficult? Int. J. Speech-Language Pathol. 11 (4) (2009) 258–271.
[3] B. Wells, S. Peppé, M. Vance, Linguistic assessment of prosody, Linguist. Clin. Practice (1995) 234–265.
[4] R.S. Chapman, L. Hesketh, Language, cognition, and short-term memory in individuals with Down syndrome, Down Syndr. Res. Practice 7 (1) (2001) 1–7.
[5] R.D. Kent, H.K. Vorperian, Speech impairment in Down syndrome: A review, J. Speech Language Hear. Res. 56 (1) (2013) 178–210.
[6] R.D. Kent, J. Eichhorn, E.M. Wilson, Y. Suk, D.M. Bolt, H.K. Vorperian, Auditory-perceptual features of speech in children and adults with down syndrome: A speech profile analysis, J. Speech Language Hear. Res. 64 (4) (2021) 1157–1175, https://doi.org/10.1044/2021_JSLHR-20-00617.
[7] E.M. Wilson, L. Abbeduto, S.M. Camarata, L.D. Shriberg, Speech and motor speech disorders and intelligibility in adolescents with Down syndrome, Clin. Linguist. Phonet. 33 (8) (2019) 790–814.
[8] P. Zanchi, L. Zampini, F. Panzeri, Narrative and prosodic skills in children and adolescents with Down syndrome and typically developing children, Int. J. Speech-Language Pathol. (2020) 1–9, https://doi.org/10.1080/17549507.2020.1804618.
[9] H.N. Jones, K.D. Crisp, M. Kuchibhatla, L. Mahler, J. Risoli, Thomas, C.W. Jones, P. Kishnani, Auditory-perceptual speech features in children with down syndrome, Am. J. Intellect. Dev. Disab. 124 (4) (2019) 324–338. doi:10.1352/1944-7558-124.4.324.

[10] M. Corrales-Astorgano, D. Escudero-Mancebo, C. González-Ferreras, Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome, Speech Commun. 99 (2018) 90–100.
[11] L. Zampini, M. Fasolo, M. Spinelli, P. Zanchi, C. Suttora, N. Salerni, Prosodic skills in children with Down syndrome and in typically developing children, Int. J. Language Commun. Disorders 51 (1) (2016) 74–83.
[12] M.T. Lee, J. Thorpe, J. Verhoeven, Intonation and phonation in young adults with down syndrome, J. Voice 23 (1) (2009) 82–87, https://doi.org/10.1016/j.jvoice.2007.04.006.
[13] D. O'Leary, A. Lee, C. O'Toole, F. Gibbon, Perceptual and acoustic evaluation of speech production in Down syndrome: A case series, Clin. Linguist. Phonet. 34 (1–2) (2020) 72–91.
[14] S. Peppé, J. McCann, Assessing intonation and prosody in children with atypical language development: the PEPS-C test and the revised version, Clin. Linguist. Phonet. 17 (4–5) (2003) 345–354.
[15] V. Stojanovik, J. Setter, Prosody in two genetic disorders: Williams and Down's syndrome, in: V. Stojanovik, J. Setter (Eds.), Speech Prosody in Atypical Populations: Assessment and Remediation, J&R Press, 2011, pp. 25–43.
[16] S.J. Loveall, K. Hawthorne, M. Gaines, A meta-analysis of prosody in autism, williams syndrome, and down syndrome, J. Commun. Disord. 106055 (2020).
[17] A.M. García, F. Carrillo, J.R. Orozco-Arroyave, N. Trujillo, J.F.V. Bonilla, S. Fittipaldi, F. Adolfi, E. Nöth, M. Sigman, D.F. Slezak, et al., How language flows when movements don't: an automated analysis of spontaneous discourse in Parkinson's disease, Brain Language 162 (2016) 19–28.
[18] J. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, E. Nöth, Automatic detection of Parkinson's disease in running speech spoken in three different languages, J. Acoust. Soc. Am. 139 (1) (2016) 481–500.
[19] B. Karan, S.S. Sahu, J.R. Orozco-Arroyave, K. Mahto, Hilbert spectrum analysis for automatic detection and evaluation of Parkinson's speech, Biomed. Signal Process. Control 61 (2020), 102050.
[20] M. Li, D. Tang, J. Zeng, T. Zhou, H. Zhu, B. Chen, X. Zou, An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder, Comput. Speech Language 56 (2019) 80–94.
[21] P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, Assessing speaker independence on a speech-based depression level estimation system, Pattern Recogn. Lett. 68 (2015) 343–350.
[22] E.L. Campbell, L. Docío-Fernández, J.J. Raboso, C. García-Mateo, Alzheimer's dementia detection from audio and text modalities, arXiv preprint arXiv:2008.04617.
[23] B. Kashyap, M. Horne, P.N. Pathirana, L. Power, D. Szmulewicz, Automated topographic prominence based quantitative assessment of speech timing in cerebellar ataxia, Biomed. Signal Process. Control 57 (2020), 101759.
[24] D. Le, E.M. Provost, Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation, in: INTERSPEECH, 2014, pp. 1563–1567.
[25] H.-D. Huici, H.A. Kairuz, H. Martens, G. Van Nuffelen, M. De Bodt, Speech rate estimation in disordered speech based on spectral landmark detection, Biomed. Signal Process. Control 27 (2016) 1–6.
[26] V. Mendoza Ramos, H.A. Kairuz Hernandez-Diaz, M.E. Hernandez-Diaz Huici, H. Martens, G. Van Nuffelen, M. De Bodt, Acoustic features to characterize sentence accent production in dysarthric speech, Biomed. Signal Process. Control 57 (2020), 101750.
[27] M. Tu, V. Berisha, J. Liss, Interpretable objective assessment of dysarthric speech based on deep neural networks., in: INTERSPEECH, 2017, pp. 1849–1853.
[28] A. Guidi, J. Schoentgen, G. Bertschy, C. Gentili, E. Scilingo, N. Vanello, Features of vocal frequency contour and speech rhythm in bipolar disorder, Biomed. Signal Process. Control 37 (2017) 23–31.
[29] A. Rochet-Capellan, M. Dohen, Acoustic characterisation of vowel production by young adults with Down syndrome, in: 18th International Congress of Phonetic Sciences (ICPhS 2015), Glasgow, United Kingdom, 2015.
[30] A. Hennequin, A. Rochet-Capellan, M. Dohen, Auditory-Visual Perception of VCVs Produced by People with Down Syndrome, Preliminary Results, in: Interspeech 2016, 2016, pp. 213–217, https://doi.org/10.21437/Interspeech.2016-1198.
[31] M. Corrales-Astorgano, D. Escudero-Mancebo, C. González-Ferreras, Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome, Speech Commun. 99 (2018) 90–100.
[32] O. Saz, J. Simón, W. Rodríguez, E. Lleida, C. Vaquero, et al., Analysis of acoustic features in speakers with cognitive disorders and speech impairments, EURASIP J. Adv. Sig. Process. 2009 (2009) 1.
[33] G. Albertini, S. Bonassi, V. Dall'Armi, I. Giachetti, S. Giaquinto, M. Mignano, Spectral analysis of the voice in Down syndrome, Res. Dev. Disabil. 31 (5) (2010) 995–1001.
[34] V. Stojanovik, Prosodic deficits in children with Down syndrome, J. Neuroling. 24 (2) (2011) 145–155.
[35] M. Corrales-Astorgano, P. Martínez-Castilla, D. Escudero-Mancebo, L. Aguilar, C. González-Ferreras, V. Cardeñoso-Payo, Automatic assessment of prosodic quality in Down syndrome: Analysis of the impact of speaker heterogeneity, Appl. Sci. 9 (7) (2019) 1440.
[36] P. Martínez-Castilla, S. Peppé, Developing a test of prosodic ability for speakers of iberian Spanish, Speech Communication 50 (11) (2008) 900–915, iberian Languages. doi: 10.1016/j.specom.2008.03.002.
[37] P. Martínez-Castilla, S. Peppé, Intonation features of the expression of emotions in Spanish: preliminary study for a prosody assessment procedure, Clin. Linguist. Phonet. 22 (4–5) (2008) 363–370.
[38] P. Martínez-Castilla, S. Peppé, Assessment of spanish prosody in clinical populations: The case of williams syndrome, in: Intonational Grammar in Ibero-Romance, John Benjamins, 2016, pp. 351–368.
[39] P. Martínez-Castilla, M. Sotillo, R. Campos, Prosodic abilities of spanish-speaking adolescents and adults with williams syndrome, Language Cognit. Process. 26 (8) (2011) 1055–1082.
[40] L. Dunn, L. Dunn, D. Arribas, Test de vocabulario en imágenes peabody, Madrid: TEA.
[41] T. Kisler, U. Reichel, F. Schiel, Multilingual processing of speech via web services, Comput. Speech Language 45 (2017) 326–347, https://doi.org/10.1016/j.csl.2017.01.005.
[42] F. Schiel, S. Burger, A. Geumann, K. Weilhammer, The partitur format at bas, in: Proceedings of the First International Conference on Language Resources and Evaluation, 1998, pp. 1295–1301.
[43] P. Boersma, Praat: doing phonetics by computer, 2006. URL:http://www.praat.org/.
[44] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in: Proceedings of the institute of phonetic sciences, Vol. 17, Citeseer, 1993, pp. 97–110.
[45] D.J. Hirst, A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation, in: Proceedings of the XVIth International Conference of Phonetic Sciences, Vol. 12331236, sn, 2007, pp. 1223–1236.
[46] M. Corrales-Astorgano, P. Martínez-Castilla, D. Escudero-Mancebo, L. Aguilar, C. González-Ferreras, V. Cardeñoso-Payo, Towards an automatic evaluation of the prosody of people with Down syndrome, in: Proc. IberSPEECH 2018, 2018, pp. 112–116. doi:10.21437/IberSPEECH.2018-24.
[47] F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in opensmile, the munich open-source multimedia feature extractor, in: Proceedings of the 21st ACM international conference on Multimedia ACM, 2013, pp. 835–838.
[48] L. Aguilar, A. Bonafonte, F. Campillo, D. Escudero, Determining intonational boundaries from the acoustic signal, in: Tenth Annual Conference of the International Speech Communication Association, 2009, pp. 2447–2450.
[49] V. Cardeñoso-Payo, C. González-Ferreras, D. Escudero-Mancebo, Assessment of Non-native Prosody for Spanish as L2 using quantitative scores and perceptual evaluation, in: LREC, 2014, pp. 3967–3972.
[50] E. Grabe, E.L. Low, Durational variability in speech and the rhythm class hypothesis, Papers in laboratory phonology 7 (515–546).
[51] J.P. Van Santen, E.T. Prud'hommeaux, L.M. Black, Automated assessment of prosody production, Speech Commun. 51 (11) (2009) 1082–1097.
[52] C.M. Bishop, Pattern recognition and machine learning, Springer, 2006.
[53] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous valued attributes for classification learning, in: Thirteenth International Joint Conference on Articial Intelligence, Vol. 2, Morgan Kaufmann Publishers, 1993, pp. 1022–1027.
[54] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019. URL:https://www.R-project.org/.
[55] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The weka data mining software: an update, ACM SIGKDD Explorat. Newsl. 11 (1) (2009) 10–18.
[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[57] S. Raschka, Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack, J. Open Sour. Soft. 3 (24). doi: 10.21105/joss.00638.
[58] L. Liu, M. Jian, W. Gu, Prosodic characteristics of mandarin declarative and interrogative utterances in parkinson's disease, Age (year) 66 (6.00) (2019) 63–64.
[59] S. Yiu, Intonation of statements and questions in cantonese english: Acoustic evidence from a smoothing spline analysis of variance, in: M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith, J.M. Scobbie (Eds.), 18th International Congress of Phonetic Sciences, ICPhS 2015, Glasgow, UK, August 10-14, 2015, University of Glasgow, 2015, pp. X–Y.
[60] C. Gu, et al., Smoothing spline anova models: R package gss, J. Stat. Softw. 58 (5) (2014) 1–25.
[61] H. Wickham, ggplot2: elegant graphics for data analysis, springer, 2016.
[62] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nat. Methods 17 (2020) 261–272, https://doi.org/10.1038/s41592-019-0686-2.
[63] R. Banse, K.R. Scherer, Acoustic profiles in vocal emotion expression, J. Personal. Soc. Psychol. 70 (3) (1996) 614.
[64] J.M. Sosa, La entonación del español: su estructura fónica, variabilidad y dialectología, Anaya-Spain, 1999.
[65] T.L. Face, The intonation of absolute interrogatives in castilian spanish, Southwest J. Linguistics 23 (2) (2004) 65–80.

[66] A. Quilis, Tratado de fonología y fonética españolas, Vol. 2, Gredos Madrid, 1993.
[67] E.A. Malt, R.C. Dahl, T.M. Haugsand, I.H. Ulvestad, N.M. Emilsen, B. Hansen, Y. Cardenas, R.O. Skøld, A. Thorsen, E. Davidsen, Health and disease in adults with Down syndrome, Tidsskrift for den Norske laegeforening: tidsskrift for praktisk medicin, ny raekke 133 (3) (2013) 290–294.
[68] C. Jarrold, A. Baddeley, Short-term memory in Down syndrome: Applying the working memory model, Down Syndr. Res. Practice 7 (1) (2001) 17–23.