Acoustic Analysis of Anomalous Use of Prosodic Features in a Corpus of People with Intellectual Disability

Mario Corrales-Astorgano, David Escudero-Mancebo, and César González-Ferreras

Department of Computer Science, University of Valladolid, Spain
mario.corrales;descuder;cesargf@infor.uva.es

Abstract. An analysis of the prosodic characteristics of the voice of people with intellectual disability is presented in this paper. A serious game has been developed for training the communicative competences of people with intellectual disability, including those related with prosody. An evaluation of the video game was carried out and, as a result, a corpus with the recordings of the spoken turns of the game has been collected. This corpus is composed of a set of utterances produced by the target group of people with intellectual disability. The same set of sentences is pronounced by another group of people without intellectual disability. This allows us to compare the prosodic profiles between the target and control groups. Prosodic features (F0, energy and duration) are automatically extracted and analyzed, revealing significant differences between the two groups. We trained an automatic classifier using exclusively prosodic features and 80% of the sentences were correctly discriminated.

Keywords: Prosody, speech, intellectual disabilities, Down Syndrome

1 Introduction

Some intellectually disabled (ID) people have problems in their social relationships because of their communication difficulties [4, 12, 3].

Speech in general [11] and prosody in particular [16] can be affected, resulting in a limited control of many communicative functions. The work presented in this paper is framed in a project [1] whose goal is to develop a serious game for ID young people (mainly but not exclusively people with Down syndrome (DS)) to train and improve their pronunciation. The users of the video game are

[1] This work has been partially funded by Recercaixa, ACUP, Obra Social La Caixa (project "¡Juguemos a comunicar mejor! La mejora de la competencia prosódica como vía de integración educativa e inclusión social del alumnado con Necesidades Específicas de Apoyo Educativo" http://prado.uab.es/recercaixa/es/index.html) and by Ministerio de Economía y Competitividad (project TIN2014-59852-R "Videojuegos Sociales para la Asistencia y Mejora de la Pronunciación de la Lengua Española")

invited to perform a set of perception and oral production activities that have been designed to practise a number of communicative functions of speech, mainly those related with prosody. In this paper, we show that the use of the video game permits us to collect a corpus that is used to detect ID people's anomalous use of prosody when their utterances are compared with those produced by people without ID.

There are few works in the literature that have analyzed the speech of ID people using a corpus based approach. Most of the studies in the state of the art have followed an approach based on tests of perception, not on measuring and comparing the values of acoustic variables across corpora [11]. Among the reasons for this sparsity of works, we can point out the lack of specific corpora and the difficulties for analysing ID affected speech. Concerning the lack of corpora, the recording of ID people's voices in controlled contexts is not an easy task due to the short term memory problems, attention deficit and language development problems of these people [3]. Our video game incorporates a number of game elements in the interface [5, 9] to increase the motivation of the users, resulting in a corpus that contains the sentences produced by the players during the interaction.

Concerning the difficulties of analyzing ID people's altered speech, it must be taken into account that this type of voice usually contains disfluencies and abnormal productions that can originate in physiological and/or perception problems. As a result, the quality of the phonetic production is low, which limits the use of automatic speech recognition systems so that automatic segmentation is, in many cases, unaffordable [7]. Prosody operates at a supra-segmental level (for example word, phrase or sentence levels), so that capturing the prosodic features is a more robust process. In this work, we show that computing a set of acoustic variables related with fundamental frequency, energy and pauses allows us to detect unusual prosodic patterns in the spoken utterances of the corpus.

The possibility of analyzing the prosodic productions of users in real time and the training of automatic systems for identifying prosodic problems is allowing the game to be enriched by providing useful information for the therapists to assist the players and for the players to improve their interaction experience. In this work we have used the prosodic features to train automatic classifiers that permit the sentences of the corpus to be classified in terms of the type of speaker with about 80% success rates.

In section 2, we present the serious game activities. In section 3, the experimental procedure is described, which includes the procedure for collecting data, the processing of the speech material and the classification of the samples. The results section shows the effectiveness of the procedure. We end the paper with a discussion about the relevance of the results and the conclusions and future work section.

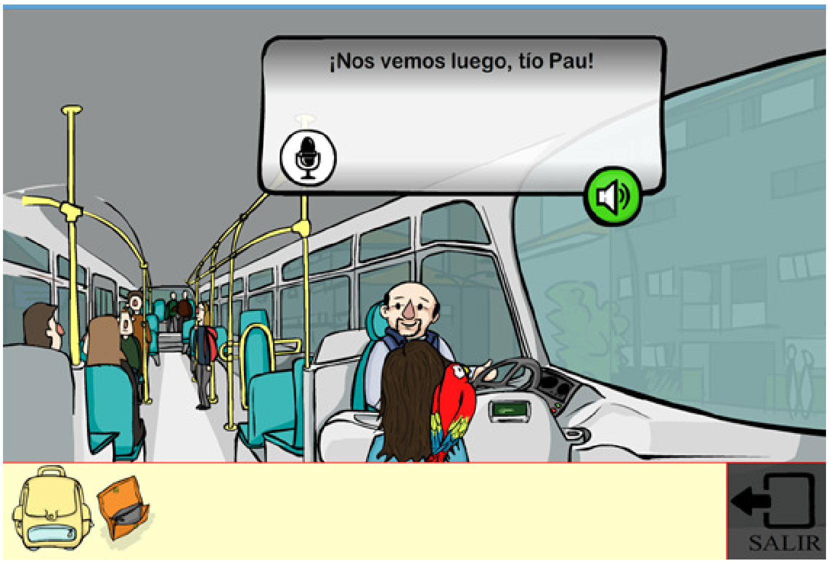

Fig. 1. The user is represented in this game screen by a personalized avatar (backwards position in this case). The parrot acts as the virtual assistant. User and game characters (the bus driver in this case) interact with voice.

2 Game description

The video game has the structure of a graphic adventure game, including conversations with characters, getting and using items and navigating through scenarios. A number of activities are included in the general context of the graphic adventure and players need to solve them in order to progress in the game. Three activity types are defined to practice speech, communication and prosodic skills:

- **Comprehension activities** that are focused on lexical-semantic comprehension and on improving prosodic perception in specific contexts, such as making a question or asking something politely.
- **Production activities** that are focused on oral production, so the players are encouraged by the game to train their speech, keeping in mind prosodic aspects like intonation, expression of emotions or syllabic emphasis.
- **Visual activities** focused on improving specific aspects of prosody, with the corresponding visual response to the user voice input and other activities designed to add variety to the game and to reduce the feeling of monotony while playing.

Two users interact with the system: the player and the trainer. The player is normally an adolescent or a child with language deficits, specifically in prosodic comprehension and production. The trainer (typically the teacher, a speech therapist or a relative) assists the player during game sessions. When trainer and

Sentence in Spanish	Sentence in English
¡Hasta luego, tío Pau!	See you later, uncle Pau!
¡Muchas gracias, Juan!	Thank you very much, Juan!
¡Hola! ¿Tienen lupas? Quería comprar una.	Hello, do you have magnifiers? I wanted to buy one.
Sí, la necesito, ¿Cuánto vale?	Yes, I need it. How much is it?
¡Hola tío Pau! Ya vuelvo a casa.	Hello uncle Pau! I'll be back home.
Sí, esa es. ¡Hasta luego!	Yes, it is. Bye!
¡Hola, tío Pau! ¿Sabes dónde vive la señora Luna?	Hello uncle Pau! Do you know where Mrs Luna lives?
¡Nos vemos luego, tío Pau!	See you later, uncle Pau!
Has sido muy amable, Juan. Muchas gracias!	You have been very kind, Juan. Thank you very much!
¡Hola! ¿Tienen lupas? Me gustaría comprar una.	Hello, do you have magnifiers? I'd like to buy one.
Sí, necesito una sea como sea. ¿Cuánto vale?	Yes, I really need one. How much is it?
Sí, lo es. Vivo allí desde pequeño. ¡Hasta luego!	Yes, it is. I have lived there since I was a child. Bye!
¡Hola, tío Pau! Tengo que encontrar a la señora Luna ¿Sabes dónde vive?	Hello uncle Pau! I have to find Mrs Luna. Do you know where she lives?

Table 1. Sentences recorded during game sessions

player work together on a game activity, the trainer will help the player in the correct use of voice and also to configure the system. Production and prosodic activities allow the trainer to evaluate the players, making them repeat the exercise when the result is not correct. The role of the trainer is essential to maximize the educational potential of the game. The trainer supports and guides players during the game, adapts the difficulty level, encourages players to continue when they have difficulties; helps them to solve such difficulties as understanding the story and the activities and gives clues for helping to improve their performance.

During the game session, information about user interaction is stored, as well as the audio recordings of the production activities. This information can be used by the speech therapist to analyze the evolution of the user in successive game sessions and the audio recordings increase the speech corpus. This user interaction log has information about game time, the attempts to complete a task, number of mouse clicks or the helps showed to the user.

A set of usability tests have been performed showing that the degree of satisfaction of players and trainers is high. The game elements engage the users, motivating them to use the software. It has been analyzed in perception tests and confirmed by the teachers that the oral productions of the players improve with use.

3 Experimental procedure

3.1 Data collection

The game sessions consisted in completing all the game and were done in the facilities of the centers where the players attended their regular classes to assure their comfort. In addition, a staff member of the centers was always with the players. The game sessions were held at the Niu School (Barcelona), Aura Foundation (Barcelona), and the College of Special Education “El Pino de Obregón” (Valladolid). In total, 4 game sessions were held. During the first of them, a usability test was carried out to see how the users interacted with the game and to detect deficiencies in the user interface. The other 3 sessions were carried out by staff members of the center of Valladolid, with the aim of getting more recordings to be analyzed later. During game sessions, the role of the trainer (a teacher or speech therapist) is twofold: on the one hand, he/she evaluates the player’s recordings and on the other hand, he/she helps players if necessary. The trainer has to sit next to the player. To evaluate the player’s recordings, the trainer uses the keyboard of the same computer that the player is using. Besides, to reduce the ambient noise in the recording process, the players use a headset with microphone incorporated.

25 users participated in the game sessions. All of them have a moderate or mild intellectual disability, so they can follow the game in a reliable way. In addition, the number of male and female users was the same and the age range was between 13 and 30 years. Of these 25 users, 18 have Down syndrome and 7 have intellectual disability without diagnosis of a specific syndrome. As the four game sessions were carried out over different periods of time, not all the users participated in all the game sessions. Some users were not available during the development of some game sessions, so we do not have their recordings.

One of the purposes of the tool is to collect examples of sentences with different modalities (i.e. declarative, interrogative and exclamatory) produced by people with intellectual disabilities in order to analyze the most common difficulties. The sentences recorded can be seen in Table 1.

3.2 Processing

The information about the corpus generated during the game sessions can be seen in Table 2. This table shows the number of recordings that each user has made, as well as the duration of these recordings in seconds. To obtain control samples, a series of recordings were made of 20 adult people, 11 men and 9 women.

The prosodic features extracted from recordings are frequency, energy and pauses. To extract frequency from recordings, an algorithm for pitch analysis based on an autocorrelation method implemented by Praat [2] is used. Energy is calculated using rms (root mean square) with a window size of 1024 samples (recordings are performed at 44100Hz). Both rms and F0 are computed every

User	Session 1	Session 2	Session 3	Session 4	Total
U1	8/50				8/50
U2	14/46	16/42	10/20		40/109
U3	12/34				12/34
U4		3/10	7/23	10/59	20/93
U5		12/56			12/56
U6	7/34				7/34
U7	11/93				11/93
U8		13/43	8/29	12/57	33/130
U9		3/14	9/48	11/40	23/103
U10	10/34				10/34
U11	8/24				8/24
U12		9/31	12/49	11/28	32/109
U13	11/44				11/44
U14	9/31	10/28	10/28	12/39	41/127
U15	10/40				10/40
U16	14/55	4/22	10/34	11/48	39/161
U17	13/39	10/30	9/19		32/90
U18	10/38				10/38
U19	7/23	11/29		13/46	31/98
U20		3/9	8/20		11/29
U21	9/46				9/46
U22		11/41	10/40		21/82
U23	13/46	10/38	10/42		33/127
U24	7/37				7/37
U25	8/33				8/33
Total	181/757	115/398	103/357	80/320	479/1,832

Table 2. Number of recordings (first number) and duration (second number) in seconds of each session of user recording

0.01 seconds. Finally, the intervals of silence and sound are also detected using the Praat tools.

To analyze the prosody of the recordings and to compare them to the recordings of people without ID, a set of features were computed. The features concern frequency: within word F0 range (`f0_range`), difference between maximum and average within word F0 (`f0_maxavg_diff`), difference between average and minimum within word F0 (`f0_minavg_diff`); energy: within word energy range (`e_range`), difference between maximum and average within word energy (`e_maxavg_diff`), difference between average and minimum within word energy (`e_minavg_diff`); and duration: number of silences (`num_silences`), silence duration in percent with respect to total duration (`silence_percent`). These sets of features we have used have been shown to be effective in previous experiments on analyzing prosody for the automatic prosodic labeling of spoken utterances [1, 8]. Besides, we use semitones and decibels to give a perceptual interpretation to these ranges and differences.

		Target		
	Control	DS	$\overline{DS}$	ID
f0_maxavg_diff (st)	4.74± 1.7	**3.71± 1.6**	**3.54± 1.3**	**3.66± 1.5**
f0_minavg_diff (st)	4.49± 2.1	**5.42± 2.7**	**5.64± 3.3**	**5.49± 2.9**
f0_range (st)	9.23± 3.2	9.13± 3.6	9.18± 4.0	9.15± 3.7
e_maxavg_diff (dB)	9.53± 1.9	**8.52± 2.5**	8.86± 2.4	**8.63± 2.4**
e_minavg_diff (dB)	40.58± 16.0	**32.14± 6.9**	**31.16± 5.9**	**31.82± 6.6**
e_range (dB)	50.11± 16.2	**40.65± 7.3**	**40.02± 6.3**	**40.44± 7.0**
num_silences (#)	0.88± 0.8	**2.02± 2.0**	**1.59± 1.6**	**1.88± 1.9**
silence_percent [0-1]	0.13± 0.1	**0.21± 0.2**	0.13± 0.1	**0.18± 0.2**

Table 3. Average and standard deviation of recording features of people without intellectual disabilities (control group) and people with intellectual disabilities (target group). DS refers to users with Down syndrome, $\overline{DS}$ refers to users with an intellectual disability that is different to Down syndrome and $ID = DS \cup \overline{DS}$. The boldfaced cells indicate significant differences with $p - value < 0.01$ applying the Mann-Whitney U test.

In order to make an automatic classification of the recordings, the Weka machine learning toolkit [10] was used. We made use of the implementations of the classifiers C4.5 decision trees (DT), Multilayer perceptron (MLP) and Support Vector Machine (SVM). We were interested in contrasting the behavior of these three types of classifiers, as we observed in [6] that they behave differently in prosodic labeling tasks.

4 Results

Table 3 shows the variables of the different groups. In general, all the variables in Table 3 present high mean and sd values due to the type of sentences to be uttered, which include many questions, exclamations and orthographically marked pauses. As for duration variables (*num_silences* and *silence_percent*), Down syndrome users need more pauses to complete their turns (0.88 vs 2.02), which is also projected in the percent of silences that they use (21% vs 13%). The impact of the use of pauses seems to be lower in the case of $\overline{DS}$ users: 0.88 vs 1.5 for *num_silences*.

Concerning the variables that refer to $F0$, both Target and Control users modulate the pitch with similar ranges (9.23 vs 9.15 st). *f0_maxavg_diff* is higher in the Control groups (4.74 vs 3.66 st), indicating higher F0 excursions. When energy is analyzed, the range is significantly different (50.11 vs 40.44 dB) and the excursions are shorter both from the mean value to the maximum (9.53 vs 8.63 dB) and to the minimum (40.58 vs 31.81 dB). No relevant differences are observed between DS and $\overline{DS}$ users.

Table 4 presents a ranking of the acoustic features in terms of how accurately they permit the type of users to be distinguished. *e_range* and *num_silence* are the most relevant features, both for classifying the DS and the $\overline{DS}$ users. Energy

	DS vs Control		$\overline{DS}$ vs Control	
Attribute	InfoGain	GainRatio	InfoGain	GainRatio
e_range	0.1311	0.1119	0.1523	0.1105
num_silences	0.1291	0.1743	0.113	0.205
e_minavg_diff	0.0987	0.2812	0.151	0.1105
f0_maxavg_diff	0.0737	0.0739	0.1076	0.1229
e_maxavg_diff	0.0707	0.1538	0.087	0.1476
silence_percent	0.0463	0.0628	0	0
f0_minavg_diff	0.0365	0.0483	0.0565	0.0875
f0_range	0	0	0	0

Table 4. *InfoGain* and *GainRatio* are entropy based metrics that indicate how relevant the acoustic features are for classifying the analyzed utterances in terms of the type of user that produced them. The ranking has been computed with the Weka tools [10].

	Control vs *DS*	Control vs $\overline{DS}$	Control vs *ID*
DT	78.1%;25.8%;19.6%	76.4%;22.2%;25.3%	80.5%;41.8%;10.4%
MLP	78.1%;44.8%;8.1%	79.3%;19.1%;22.8%	80.2%;61.3%;2.9%
SVM	80.6%;26.8%;15.0%	79.8%;14.9%;26.6%	81.0%;40.2%;10.4%

Table 5. Classification results of the three classifiers. In every cell, Ac;FPC;FPT are the accuracy, the false positive classification rates of Control samples and the false positive classification rates of the Target samples.

related variables are more discriminant than the F0 ones. Thus, for example, *f0_range* is not discriminant at all. The duration of the silences (*silence_percent* variable) again seems to be similar for Control and $\overline{DS}$ users.

Table 5 shows that the classification of the users through the prosodic features permits us to identify correctly about 80% of the utterances. The SVM classifier permits us to obtain the best classification results in all the cases, with an accuracy of 81% and with the lowest (on average) false positive rates in the Control vs *DS* case. The worst results are obtained with the *MLP* classifiers with false positive rates of the target group close to or over 50%.

5 Discussion

The results presented in Tables 3 and 4 shows us that *ID* players have problems with correctly using the pauses. This is an expected result as many authors have reported fluency disorders, including stuttering and cluttering that mainly affects *DS* speakers [11].

Additionally, we have observed a different use of the prosodic variables of F0 and energy, which could indicate an anomalous production of intonation and emphatic stress by *ID* speakers. Limitations in the production of prosodic features have already been reported in works like the one presented in [13] and others [11]. The difference between our proposal and the previous works is that we use acoustic features instead of perceptual tests for evidencing the abnormal use of prosody.

Of particular interest is the fact that the control of energy seems to be more problematic than the control of F0 for intellectually disabled speakers. [15] also observed this fact when analyzing the Alborada corpus [14].

To discuss about the practical implications of these results, let us go back to the video game. Our goal is that the interface [5, 9] permits both the user to operate autonomously or with the supervision of the trainer. In the assisted mode, the trainer decides whether the user should change the activity or continue practicing it. In the autonomous mode, the program must take the decision automatically. Users are invited to keep trying until their pronunciation is estimated to be good enough (or until a given number of repetitions is exceeded to avoid user frustration and abandonment). In the assisted mode, the trainer can receive real time information about the deviation of the prosodic features of the user with respect to the expectations. This information could be used by the trainer to decide whether the user should repeat the sentence production or continue the game. In the autonomous mode, the automatic classification results could be used to judge the quality of the production: the higher the probability of being a sample of the Control group, the better the pronunciation is. This practical concern will determine the type of classifier to be used, because a wrong classification of a user of the Control group can decrease dramatically the overall perception of the quality of the system.

6 Conclusions and future work

As a conclusion, prosodic features are able to discriminate speakers with intellectual disabilities with a high success rate. The most discriminant features are the number of pauses and the duration of the silences, because of their less fluent speech. Moreover, we have observed that these speakers make an anomalous use of F0 and energy features.

We have used global statistical metrics of F0 and intensity that do not allow a study of details about their local temporal evolution. We are currently working on obtaining information about the dynamics of those acoustic correlates, in order to learn more about the particular acoustic deviations of speakers with ID. A deeper knowledge of the type of prosodic deviations will allow us to design specific exercises to correct the anomalies.

The approach used in this work to identify anomalous utterances contrasts with the alternative approach of using a reference speaker or golden speaker, which is used to compare each utterance. We are currently undertaking a comparative study of both techniques in order to analyze the advantages and disadvantages of each alternative.

References

1. Ananthakrishnan, S., Narayanan, S.S.: Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. Audio, Speech, and Language Processing,

IEEE Transactions on 16(1), 216–228 (2008)

2. Boersma, P., et al.: Praat, a system for doing phonetics by computer. Glot international 5(9/10), 341–345 (2002)
3. Chapman, R.S.: Language development in children and adolescents with down syndrome. Mental Retardation and Developmental Disabilities Research Reviews 3(4), 307–312 (1997)
4. Cleland, J., Wood, S., Hardcastle, W., Wishart, J., Timmins, C.: Relationship between speech, oromotor, language and cognitive abilities in children with down's syndrome. International journal of language & communication disorders 45(1), 83–95 (2010)
5. Corrales, M., Escudero, D., Flores, V., González, C., Gutiérrez, Y.: Arquitectura para la interacción en un videojuego para el entrenamiento de la voz de personas con discapacidad intelectual. In: Actas del XXI Congreso Internacional de Interación Persona-Ordenador. pp. 445–448 (Septembre 2015)
6. Escudero-Mancebo, D., González-Ferreras, C., Vivaracho-Pascual, C., Cardeñoso Payo, V.: A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling. Computer Speech and Language 28(1), 326 – 341 (2014)
7. Feng, J., Lazar, J., Kumin, L., Ozok, A.: Computer Usage by Children with Down Syndrome: Challenges and Future Research. ACM Transactions on Accessible Computing 2(3), 13 (2010)
8. Gonzalez-Ferreras, C., Escudero-Mancebo, D., Vivaracho-Pascual, C., Cardeñoso Payo, V.: Improving automatic classification of prosodic events by pairwise coupling. Audio, Speech, and Language Processing, IEEE Transactions on 20(7), 2045 –2058 (sept 2012)
9. González-Ferreras, C., Escudero-Mancebo, D., Corrales-Astorgano, M., Gutiérrez-González, Y., Aguilar-Cuevas, L., Flores-Lucas, V., Cardeñoso-Payo, V.: Engaging adolescents with down syndrome in an educational video game. ACM Transactions on Computer-Human Interaction (TOCHI) Under revision (2016)
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10–18 (2009)
11. Kent, R.D., Vorperian, H.K.: Speech impairment in down syndrome: A review. Journal of Speech, Language, and Hearing Research 56(1), 178–210 (2013)
12. Martin, G.E., Klusek, J., Estigarribia, B., Roberts, J.E.: Language characteristics of individuals with down syndrome. Topics in Language Disorders 29(2), 112 (2009)
13. Pettinato, M., Verhoeven, J.: Production and perception of word stress in children and adolescents with down syndrome. Down Syndrome Research & Practice 13, 48–61 (2008)
14. Saz, O., Lleida, E., Vaquero, C., Rodríguez, W.R.: The alborada-i3a corpus of disordered speech. In: LREC (2010)
15. Saz, O., Simón, J., Rodríguez, W., Lleida, E., Vaquero, C., et al.: Analysis of acoustic features in speakers with cognitive disorders and speech impairments. EURASIP Journal on Advances in Signal Processing 2009, 1 (2009)
16. Stojanovik, V.: Prosodic deficits in children with down syndrome. Journal of Neurolinguistics 24(2), 145–155 (2011)