PROGRAMA DE DOCTORADO EN INFORMÁTICA

TESIS DOCTORAL:

ENTRENAMIENTO DE LA PROSODIA EN PERSONAS CON SÍNDROME DE DOWN MEDIANTE EL USO DE UN VIDEOJUEGO EDUCATIVO

Presentada por MARIO CORRALES ASTORGANO para optar al grado de Doctor por la Universidad de Valladolid

Dirigida por:
David Escudero Mancebo
César González Ferreras

Agradecimientos

Esta tesis ha sido realizada en el grupo de investigación ECA-SIMM del Departamento de Informática de la Escuela de Ingeniería Informática de la Universidad de Valladolid.

Quiero agradecer a David Escudero Mancebo y a César González Ferreras su apoyo y dedicación durante todo el transcurso de la tesis. También agradecerles especialmente su búsqueda de financiación que me ha permitido completar el trabajo. Agradecerles también el buen ambiente de trabajo que ha ayudado mucho en los momentos más complicados. Sumo a este agradecimiento a Valentín Cardeñoso Payo, que me ha apoyado mucho con sus conocimientos y experiencia.

A Cristian Tejedor García, que ha sido un compañero de laboratorio estupendo. Con él he compartido multitud de conversaciones y viajes, y espero seguir compartiéndolas en el futuro.

Quiero agradecer especialmente a todas las personas con las que he colaborado a lo largo del trabajo de esta tesis. En primer lugar, agradecer a Lourdes Aguilar, profesora del departamento de Filología Hispánica de la Universidad Autónoma de Barcelona y experta en prosodia, su confianza en el año 2014 para que yo fuera el técnico informático encargado de desarrollar la primera versión de un videojuego para la práctica de la prosodia en personas con síndrome de Down, en el transcurso del proyecto Resercaixa. Mi trabajo fin de Máster se

basó en el trabajo desarrollado en los comienzos de este proyecto, y mi tesis doctoral continuó a partir de ese momento. Lourdes siguió presente durante los siguientes años, con el proyecto PRADIA, cuyo objetivo era introducir mejoras el videojuego educativo desarrollado en Resercaixa, y participando en algunas publicaciones y comunicaciones a congresos realizados durante este tiempo. También quiero agradecerle a Yurena María Gutiérrez González, lingüista y profesora asociada en la Universidad Autónoma de Barcelona, su paciencia y dedicación en la configuración de las actividades de entrenamiento y primeros niveles del videojuego, así como en la revisión de las implementaciones del mismo junto a Lourdes Aguilar. Agradecer a Ferran Adell, experto en videojuegos y educación, sus aportes en el diseño de la narrativa y otros elementos del videojuego. También agradecer a Patricia Sinobas su colaboración realizando los diseños de todos los elementos gráficos que componen el videojuego. Agradecerles a Valle Flores Lucas, profesora titular en el departamento de Psicología de la Universidad de Valladolid y experta en desarrollo del lenguaje en personas con síndrome de Down, y a Pastora Martínez Castilla, profesora en el departamento de Psicología Evolutiva y de la Educación de la UNED y experta en trastornos del neurodesarrollo, su aportación en todo lo referente a las personas con discapacidad intelectual, que ha sido imprescindible para alcanzar los objetivos de esta tesis, así como su gran dedicación en la escritura de las diferentes publicaciones derivadas del trabajo de esta tesis y de sus proyectos relacionados. Agradecerles también a todos los centros de educación especial y a su personal laboral su colaboración en las diferentes sesiones de juego y, por supuesto, a los chicos y chicas que acuden a estos centros, que son

una parte imprescindible sin la que esta investigación no tendría sentido. Los centros que han colaborado en esta investigación han sido la escuela Niu (Barcelona), la fundación Aura (Barcelona), la fundación síndrome de Down de Madrid, la escuela de educación especial "El Pino de Obregón" (Valladolid) y la asociación de síndrome de Down de Valladolid. Dentro del personal de estos centros de educación especial, agradecerle especialmente su colaboración a Yolanda Martín, que ha estado presente desde las primeras versiones del juego y que sigue estando presente con los proyectos en vigor. Sus aportaciones y experiencia han sido fundamentales para el éxito de esta investigación.

Y por supuesto a Viki, sin la que no estaría aquí. Su apoyo y comprensión durante estos años ha sido fundamental en los buenos y malos momentos. Y a mi perro Sarri, compañero fiel siempre. Gracias por todo.

Resumen

Las personas con síndrome de Down tienen una serie de limitaciones cognitivas que afectan, entre otras cosas, al desarrollo de sus habilidades lingüísticas. Dentro de las áreas del lenguaje afectadas se encuentra la prosodia, que es la rama de la lingüística que estudia aspectos suprasegmentales del habla, como la entonación, el tono, el acento o el ritmo. La utilización de forma combinada de estos aspectos complementa la intención comunicativa de una producción oral, aportando matices fundamentales para la correcta transmisión del mensaje requerido por el hablante, como pueden ser la expresión de emociones o la adecuación a un contexto social concreto. La utilización incorrecta de la prosodia puede afectar negativamente a la integración social de las personas que presentan problemas en su uso. La presencia de estas deficiencias en la población con síndrome de Down motiva la existencia de terapias del habla orientadas a la mejora de sus habilidades comunicativas. Los profesionales que realizan estas terapias pueden contar con el apoyo de herramientas tecnológicas que facilitan la realización de las mismas. Entre este tipo de herramientas se encuentran los videojuegos educativos, que están centrados principalmente en el entrenamiento de determinadas habilidades por encima del mero entretenimiento. Dentro de las limitaciones cognitivas que presenta esta población se encuentran los problemas de memoria a corto plazo o el déficit de atención, que les dificultan realizar tareas complejas durante

un largo periodo de tiempo, por lo que es muy importante que este tipo de herramientas puedan motivar a los usuarios con el objetivo de potenciar el uso de las mismas. Se ha observado que los videojuegos pueden ser útiles para incrementar la motivación de personas con síndrome de Down a la hora de realizar tareas de entrenamiento. Sin embargo, la existencia de este tipo de herramientas enfocadas a la práctica de la prosodia en personas con síndrome de Down es limitada.

El objetivo principal de esta tesis es la definición de un videojuego educativo enfocado al entrenamiento de algunos aspectos del lenguaje, especialmente los relacionados con la prosodia. Para lograr este objetivo, es necesario tener en cuenta las limitaciones cognitivas de las personas con síndrome de Down que afectan al uso de este tipo de herramientas y sus particularidades en el empleo de la prosodia. Tanto el diseño del videojuego como la interacción entre éste y el jugador deben motivar a los usuarios a completar las actividades de entrenamiento. Además, otro de los objetivos es definir un sistema de análisis de la calidad de las producciones orales centrado en la prosodia, con el objetivo de poder evaluar automáticamente dichas producciones de los jugadores y potenciar el uso autónomo de la herramienta.

En el trabajo de esta tesis se realiza la definición del videojuego, con el foco en el diseño de las actividades de entrenamiento, la interfaz y la interacción con el usuario. Además, se realiza una evaluación del uso del videojuego por parte de usuarios reales con el objetivo de analizar la idoneidad de las decisiones de diseño tomadas, así como detectar los posibles problemas no resueltos inicialmente. Por otro lado, es necesario definir objetivamente los criterios que definen la calidad prosódica de una producción oral y cómo se relacionan con las características

acústicas de la señal sonora. La existencia de corpus de habla de personas con síndrome de Down es limitada, especialmente los orientados al estudio de la prosodia, por lo que la herramienta sirve también para recolectar muestras de voz de esta población. Utilizando las grabaciones obtenidas mediante el uso del videojuego, se realiza una comparación entre algunas características acústicas extraídas de estas grabaciones con las extraídas de grabaciones realizadas por personas con desarrollo típico, con el objetivo de detectar las particularidades en el uso de la prosodia de las personas con síndrome de Down. Estas características están relacionadas con la frecuencia fundamental, la energía, el ritmo o el espectro de la señal acústica. Además, se realiza un análisis de la relación entre los criterios de calidad prosódica definidos y las características acústicas de las grabaciones de personas con síndrome de Down. Esto se ha efectuado tomando como referencia la evaluación de las grabaciones realizada por expertas en prosodia, cuyo criterio se pretende reproducir utilizando las características acústicas extraídas de las grabaciones. Para ello se han utilizado métodos estadísticos y de aprendizaje automático, como tests no paramétricos para comprobar las diferencias entre los grupos o el uso de clasificadores automáticos para clasificar las muestras como las máquinas de soporte vectorial, los árboles de decisión o los perceptrones multicapa.

Los resultados obtenidos en los experimentos realizados muestran que la herramienta consigue mantener la atención de los jugadores con síndrome de Down durante toda la sesión de juego y que las actividades desarrolladas cumplen el propósito de detectar y entrenar los problemas relacionados con la prosodia en esta población. Además, se han identificado diferencias significativas en los dominios de la frecuencia,

energía, temporal y espectral cuando se comparan las características acústicas extraídas de las grabaciones de personas con síndrome de Down y de las grabaciones de personas con desarrollo típico. La tasa de clasificación obtenida en la tarea de clasificar una grabación como proveniente de una persona con síndrome de Down o de una persona con desarrollo típico está por encima del 95 % utilizando combinaciones de las características acústicas extraídas. Por último, al analizar la relación entre las evaluaciones realizadas por expertas en prosodia y las características acústicas de las grabaciones de personas con síndrome de Down, se consigue una tasa de clasificación de 79,3 % cuando se utilizan estas características para entrenar algunos clasificadores automáticos con el objetivo de predecir las evaluaciones de las expertas en prosodia. Además, la tasa de falsos positivos, que es el porcentaje de grabaciones clasificadas como incorrectas por el clasificador pero evaluadas como correctas por la experta en prosodia, es del 10.1 %. Este resultado es importante para reducir la frustración de los jugadores en una posible implementación de la evaluación automática dentro del videojuego. Cuando se analizan los resultados para cada usuario, se ha detectado que la heterogeneidad inherente a esta población afecta considerablemente a los resultados, obteniendo mayores tasas de concordancia cuando el nivel cognitivo de los hablantes es alto y menores tasa de concordancia cuando el nivel cognitivo de los hablantes es bajo.

Los resultados relacionados con la evaluación del videojuego muestran la importancia que tiene realizar un diseño de las actividades y de la interacción del mismo teniendo en cuenta las limitaciones de las personas con síndrome de Down. Las decisiones tomadas en este sen-

tido son fundamentales para aumentar la motivación de los jugadores hacia el uso del videojuego y para detectar y entrenar las deficiencias más importantes de estas personas en el uso de la prosodia. Además, la comparación realizada entre las voces de personas con síndrome de Down y personas con desarrollo típico muestran diferencias entre algunas características acústicas extraídas de las grabaciones de ambos grupos. Estas diferencias revelan la importancia de la prosodia a la hora de identificar una voz como atípica. Por último, se ha observado la influencia que tiene el perfil lingüístico y psicológico de los hablantes en la evaluación automática de la calidad prosódica de sus grabaciones. La heterogeneidad de las personas con síndrome de Down debe de ser tenida en cuenta a la hora de realizar este tipo de sistemas de evaluación automática.

Abstract

People with Down syndrome have some cognitive problems that affect the development of their language skills. Prosody is one of the affected areas and it is concerned with some suprasegmental aspects of speech, such as intonation, tone, accent or rhythm. The combination of these aspects can complement the communicative intention of the speaker, adding some fundamental elements to produce, for example, emotions or social context adaptation. The low control of prosody can stigmatize speakers and limit their options to get integrated in society. Due to these prosody problems, some speech therapies have been used to improve the communication skills of people with Down syndrome. Technological tools can be useful to help speech therapist in their work. One of these tools are educational video games, which are focused on training some specific skills instead of just being entertainment. People with Down syndrome present other cognitive limitations, such as short term memory problems or attention deficit. These limitations make it difficult for them to perform hard tasks during a long period of time, so these technological tools have to motivate users with the aim of reducing the impact of these limitations on the use of them. It has been observed that video games are useful to improve motivation in people with Down syndrome. However, little attention has been paid to the development of technological resources that specifically consider the learning of prosody in students with Down syndrome.

The main aim of this work is the definition of a video game focused on training some language skills, specifically those ones related with prosody. The cognitive limitations and prosody deficits of people with Down syndrome have to be taken into account in order to achieve this aim. The video game design as well as the user interaction must motivate users to complete the training activities included in the video game. In addition, the creation of an analysis system to evaluate the prosodic quality of users' utterances is a necessity to improve the use of the video game in an autonomous way.

In this work, the definition of the training activities, the interface and the user interaction of the video game are made. In addition, an evaluation of the use of the video game by real users is made, with the aim of analyzing the efficacy of the design decisions. On the other hand, a prosodic quality criteria have to be defined, as well as its relationship with some acoustic features of the speech. There are few speech corpus of people with Down syndrome focused on prosody, so the video game is used to record some speech samples of this population. A comparison between some acoustic features extracted from the recordings of this corpus and the same features extracted from utterances recorded by people without intellectual disabilities is done. This comparison has the aim of detecting differences on the use of prosody in people with Down syndrome. The acoustic features are related with fundamental frequency, energy, rhythm or the spectral domain of the speech signal. Furthermore, an evaluation of the recordings of the speech corpus of people with Down syndrome is done by a prosody expert using the defined prosodic quality criteria. To find a relationship between this evaluation and the acoustic features extracted from the

recordings, some statistical methods and machine learning techniques are used. Specifically, the acoustic features and the expert evaluation are used to train three automatic classifiers (support vector machines, decision trees and multilayer perceptron) with the aim of predicting automatically the expert evaluation using only the acoustic features.

The results show that the video game is useful to keep the attention of the players with Down syndrome during all game session. The training activities can detect and train the specific problems related with prosody of this population. In addition, some statistically significant differences are found between people with Down syndrome and people without intellectual disabilities in the frequency, energy, temporal and spectral domain. The accuracy of identifying a recording as produced by a person with Down syndrome or by a person without intellectual disabilities is up to 95 %. Finally, an accuracy of 79.3 % is achieved in the task of predicting the prosodic expert evaluation using an automatic classifier trained with the acoustic features extracted from the recordings of people with Down syndrome. The percentage of recordings evaluated as correct by the prosody expert and evaluated as incorrect by the automatic classifier is only 10.1 %. This is an important result because evaluating a correct recording as incorrect can produce frustration on the players in an implementation of an automatic evaluation of prosodic quality inside the video game. The heterogeneity of the people with Down syndrome affects the results when they are analyzed individually. The agreement between the prosodic expert and the therapist depends on the speaker's developmental levels, obtaining better agreement values when the developmental level of the users is high and worse agreement values

when the developmental level of the users is low.

The results related with the evaluation of the video game show the importance of designing the training activities and the user interaction taking into account the limitations of people with Down syndrome. These design decisions are essential to improve the motivation of the users to play the video game and to detect and train the prosodic deficits of this population on the use of prosody. In addition, the differences found between the acoustic features extracted from the recordings of people with Down syndrome and the acoustic features extracted from the recordings of people without intellectual disabilities show the importance of prosody in order to identify a voice as atypical. Finally, the results related with the automatic evaluation of the prosodic quality show the influence of the psychological and linguistic profile of the speakers in these results. The heterogeneity of people with Down syndrome has to be taken into account to develop this kind of automatic evaluation systems.

Índice general

Índice de figuras

ÍNDICE DE FIGURAS

Índice de tablas

ÍNDICE DE TABLAS

Capítulo 1

Introducción

El trabajo de tesis que se desarrolla en esta memoria se centra en la definición de una herramienta tecnológica que ayude en el entrenamiento de ciertas habilidades lingüísticas, en especial la prosodia, en personas con síndrome de Down. Como herramienta concreta se propone un videojuego educativo, con el objetivo de mitigar las limitaciones cognitivas que presentan las personas con síndrome de Down en relación a realizar tareas de entrenamiento que requieran una dedicación cognitiva alta. Las decisiones de diseño relacionadas con las actividades de entrenamiento incluidas en el videojuego, la interfaz y la interacción con el usuario son claves para motivar a los usuarios a completar dichas tareas de entrenamiento. La caracterización de las voces de personas con síndrome de Down en el uso de la prosodia y la evaluación automática de las actividades de entrenamiento con el objetivo de promover el uso autónomo de la herramienta son otros aspectos fundamentales del trabajo de esta tesis.

La sección 1.1 describe los problemas en el uso de la prosodia de las personas con síndrome de Down, las soluciones existentes para paliar estas dificultades y las propuestas que este trabajo de tesis plantea para abarcar las limitaciones que estas soluciones presentan. La sección 1.2 describe los objetivos iniciales de esta tesis en relación a las limitaciones encontradas previamente. La sección 1.3 detalla el marco de trabajo propuesto y la sección 1.4 especifíca la metodología a desarrollar para alcanzar los objetivos definidos.

1.1. Motivación

El síndrome de Down es la causa más frecuente de discapacidad intelectual de origen genético [21]. Concretamente, el síndrome de Down se produce por la presencia de una tercera copia del cromosoma 21 (trisomía 21). En 2008, se estimó que aproximadamente 8 de cada 10.000 personas tenían este síndrome en Estados

Unidos [41] mientras que en España se estimó que existían unas 35.000 personas con este síndrome en el mismo año [26]. El síndrome de Down puede aparecer en la descendencia de personas de todas las razas y niveles socioeconómicos, pero la probabilidad de tener descendencia que presente este síndrome se incrementa significativamente en mujeres mayores de 35 años. También se ha observado una relación entre la edad paterna y la probabilidad de tener hijos con síndrome de Down [22]. El aumento de la edad en la que se tienen hijos en Europa ha producido un aumento de los nacimientos de personas con síndrome de Down, de una media de 16 por cada 10.000 nacimientos en 1990 a una media de 23 por cada 10.000 nacimientos en 2015 [29]. Este síndrome provoca una serie de efectos fisiológicos y psicológicos en las personas que lo padecen, como pueden ser limitaciones cognitivas, hipotonía muscular o enfermedades de oído. Estos problemas afectan, entre otros aspectos, al desarrollo de sus habilidades lingüísticas [32]. Aunque las personas con síndrome de Down presentan problemas en la adquisición de vocabulario, es en el uso correcto de la sintaxis donde tienen más dificultades [17]. Con respecto al uso de la pragmática, entendida como la adecuación del uso del lenguaje a un contexto específico, las personas con síndrome de Down presentan problemas en la distinción entre preguntas y enunciados neutros, en la discriminación de emociones, en el uso coherente del discurso o en el entendimiento y producción de figuras retóricas [27; 32]. A nivel fonológico, la inteligibilidad está afectada debido a una producción incorrecta de algunos fonemas, a la eliminación de algunas consonantes o la simplificación silábica [30].

Otra de las áreas del lenguaje que también está afectada es la prosodia. La prosodia estudia aspectos supra-segmentales del habla, como la entonación, el tono, el acento o el ritmo. Estos aspectos combinados se utilizan para complementar el significado de una oración, destacando otros elementos del lenguaje que no están representados en la gramática o en el vocabulario concreto utilizado [15].

Por esta razón, la prosodia es un aspecto fundamental en la comunicación, ya que contribuye, por ejemplo, a la identificación de la modalidad de una frase o a la expresión de emociones. Un uso incorrecto de la prosodia puede afectar negativamente a la integración social de las personas que presentan problemas en su uso [50]. Según diversos estudios, en el caso concreto del síndrome de Down, estas personas presentan disfluencias en el habla (tartamudeo) así como limitaciones en la percepción, imitación o producción espontánea de algunas funciones prosódicas [27; 39; 46]. De la propia señal acústica generada al hablar se pueden extraer una serie de características que están relacionadas con los aspectos que estudia la prosodia. Cuando se analiza la prosodia utilizando la señal de voz, se emplean la frecuencia fundamental (entonación y el tono), la duración (ritmo), la energía (acento) y algunas características espectrales. Existen estudios que han encontrado diferencias en algunas de estas características acústicas cuando se comparan la voces de las personas con síndrome de Down y de las personas con desarrollo típico [27]. Debido a las dificultades cognitivas de la población con síndrome de Down, la grabación de un corpus de voz enfocado en esta población es una tarea complicada, lo que dificulta la realización de un estudio centrado en la medición y comparación de variables acústicas extraídas directamente de los audios incluidos en un corpus.

Existen terapias del habla cuyo objetivo es entrenar y mejorar diferentes habilidades comunicativas en personas que presenten alguna disfunción en las mismas [4]. Centrándonos en el ámbito tecnológico, existen herramientas denominadas *Herramientas de terapia del habla y lenguaje asistidas por ordenador* (Computer-Aided Speech and Language Therapy, CASLT, por sus siglas en inglés), que tienen como objetivo el entrenamiento de personas con problemas en el lenguaje mediante el uso de la tecnología. Estas herramientas abarcan diferentes problemas del lenguaje, como el entrenamiento de habilidades básicas de fonación y compren-

sión del lenguaje en personas con alteraciones en el lenguaje [44; 45], la diagnosis y entrenamiento de personas con problemas de tartamudeo [47] o la mejora de la pronunciación de un idioma por parte de personas no nativas en el mismo [48]. También existen herramientas tecnológicas centradas en el entrenamiento de la prosodia orientadas a la lectura expresiva [37]. Sin embargo, no existen apenas herramientas exclusivamente orientadas al entrenamiento de la prosodia en personas con discapacidad intelectual, carencia que motiva la realización de este trabajo de tesis.

La caracterización de la prosodia es una tarea complicada debido, entre otras razones, a la dificultad de establecer un modelo de validez prosódica que pueda ser utilizado para evaluarla. Es complicado establecer qué aspectos del uso de la prosodia pueden ser considerados como indicadores de calidad, ya que la prosodia puede ser utilizada de diferentes formas para expresar un mismo significado [25]. La adecuación al contexto en el que se produce una oración también puede influir en la evaluación de la calidad prosódica. Además, la tarea se vuelve más complicada cuando se analiza la prosodia en voces patológicas [38]. Estas dificultades pueden explicar esta falta de herramientas específicas. Para potenciar su utilidad dentro de la terapia del habla, es recomendable que este tipo de herramientas incluyan algún tipo de evaluación de la validez de las producciones orales que realizan los usuarios en las mismas, ya que esta información puede ser muy útil tanto para los propios usuarios como para los terapeutas que acompañan a los usuarios en estas terapias.

Los problemas cognitivos específicos de las personas con síndrome de Down pueden condicionar la efectividad de las herramientas tecnológicas desarrolladas para esta población. Entre estos problemas cognitivos se encuentran el déficit de atención [33], una baja tolerancia a la frustración [20], dificultad para procesar información que proceda de varios canales simultáneamente [28], falta de motiva-

ción [51], problemas con la memoria a corto plazo [6] o dificultades para entender el significado de cierta iconografía [52]. A pesar de todas estas dificultades, las personas con síndrome de Down son capaces de utilizar la tecnología para realizar diversas tareas en su día a día [20], por lo que el uso de herramientas tecnológicas enfocadas al entrenamiento de personas con síndrome de Down es posible, pero hay que tener en cuenta todas las limitaciones mencionadas anteriormente a la hora de diseñar dichas herramientas.

Dentro de las herramientas tecnológicas enfocadas al aprendizaje, se ha observado que los videojuegos tienen gran potencial para aumentar la motivación y la disposición a realizar actividades de los usuarios en entornos educativos [35]. Existen videojuegos orientados al entrenamiento de diferentes habilidades en personas con discapacidad intelectual, como, por ejemplo, los videojuegos que tienen como objetivo promover la interacción con otras personas en entornos laborales o el uso del transporte público [49]. Otros videojuegos se centran en la terapia de niños con trastornos en el habla [5]. Con respecto a la población con síndrome de Down, existen videojuegos orientados a la práctica de diferentes habilidades, como el entrenamiento de la vocalización en edades tempranas o el entrenamiento de la lectura [3], o la práctica de habilidades matemáticas [23]. Sin embargo, estas propuestas presentan las actividades de entrenamiento de forma aislada, sin introducir un contexto general donde estas actividades estén directamente integradas.

En resumen, el uso de herramientas tecnológicas, incluidos videojuegos, para el entrenamiento del lenguaje en personas con síndrome de Down es un enfoque a analizar. Sin embargo, es imprescindible conocer las limitaciones de esta población para poder desarrollar cualquier herramienta tecnológica enfocada específicamente a personas con síndrome de Down, poniendo especial énfasis en el diseño de la interfaz de usuario y en la manera de plantear la interacción entre la

herramienta y el usuario de la misma. Para que este tipo de herramientas puedan ser utilizadas en terapias del habla es necesario ofrecer algún tipo de evaluación de las actividades que se realicen con la herramienta. Comprender las características específicas del habla de las personas con síndrome de Down, así como los indicadores que influyen a la hora de determinar la validez de una producción concreta, específicamente los relacionados con la prosodia, es fundamental en este tipo de herramientas. La grabación de un corpus de voz centrado en la prosodia que contenga las frases que se reproducen en las actividades del videojuego es necesario, ya que las grabaciones contenidas en este corpus pueden ser analizadas para la caracterización y evaluación de las voces de personas con síndrome de Down.

1.2. Objetivos

En el apartado anterior se exponen los problemas relacionados con el uso de la prosodia que presentan las personas con síndrome de Down a la vez que se reflejan los beneficios que los videojuegos educativos producen en el entrenamiento de ciertas habilidades en personas con discapacidad intelectual. Además, también se plantean algunas limitaciones de las soluciones existentes. La falta de un diseño específico orientado a paliar las dificultades de las personas con síndrome de Down a la hora de interactuar con herramientas tecnológicas o el entrenamiento específico de la prosodia en esta población son algunas de ellas. Estas limitaciones motivan el objetivo general de esta tesis, que es:

- Definir un videojuego educativo que permita entrenar algunos aspectos del lenguaje, especialmente los relacionados con la prosodia y que se adapte a las limitaciones cognitivas de las personas con síndrome de Down.

Como se ha visto en la sección 1.1, las personas con síndrome de Down presentan una serie de limitaciones físicas y psicológicas que influyen en el uso de

herramientas tecnológícas así como en la capacidad de atención para realizar una tarea compleja. La elección del género del videojuego, la adecuación de las actividades de entrenamiento a las dificultades principales de la población con síndrome de Down en el uso de la prosodia y el diseño de la interacción entre el usuario y la herramienta son aspectos clave a tener en cuenta. Además, la evaluación automática centrada en la prosodia de las locuciones realizadas en las actividades de entrenamiento puede ayudar al uso autónomo de la herramienta por parte de los usuarios. Teniendo en cuenta estos aspectos, el objetivo principal se puede desgranar en los siguientes objetivos específicos:

- O1: Diseñar y evaluar un videojuego educativo centrado en el entrenamiento de la prosodia en personas con síndrome de Down.
  - O1.1 Definir un diseño del videojuego, de las actividades de entrenamiento y de la interacción con el usuario que motive el uso del mismo a las personas con síndrome de Down.
  - O1.2 Evaluar la adecuación de las actividades y diseño del videojuego con respecto a las limitaciones de las personas con síndrome de Down.
- O2: Definir un sistema de análisis de la calidad de las producciones orales de los jugadores centrado en la prosodia.
  - O2.1 Analizar las características acústicas más importantes a la hora de diferenciar una voz de persona con síndrome de Down con respecto a una voz de persona con desarrollo típico.
  - O2.2: Realizar una evaluación automática del uso de la prosodia utilizando las características acústicas extraídas de las locuciones realizadas durante las sesiones de juego.

Figura 1.1: Marco de trabajo general planteado para el desarrollo de herramientas educativas enfocadas en personas con síndrome de Down.

1.3. Marco de trabajo

La figura 1.1 muestra el marco de trabajo general planteado en esta tesis para el desarrollo de herramientas educativas orientadas a personas con síndrome de Down. Como eje central, se plantea la herramienta como un videojuego educativo que incluye las actividades de entrenamiento enfocadas a la práctica de una habilidad específica, en este caso la prosodia. Además, con el objetivo de que las herramientas desarrolladas basándose en este marco de trabajo puedan ser utilizadas en terapias del habla, es necesario tener en cuenta no solo a los usuarios que presentan problemas en el habla, sino también a los terapeutas que desarrollan las propias terapias.

Por un lado, es necesario que el género del videojuego elegido, las activida-

des diseñadas y la narrativa en la que se incluyen se ajusten a los objetivos de aprendizaje definidos para la herramienta y a las características específicas de las personas con síndrome de Down (elemento *Narrativa y actividades* de la figura 1.1). Con respecto al diseño de las actividades, es necesario analizar los problemas relacionados con el uso de la prosodia que presentan las personas con síndrome de Down, tanto en la percepción como en la producción de la misma. Partiendo de estas limitaciones, las actividades deben ser diseñadas con el objetivo de que sean efectivas para el entrenamiento de la prosodia en esta población. Además, es necesario desarrollar una narrativa que incluya estas actividades.

Relacionado con el punto anterior, un correcto diseño de la interfaz de usuario y de la interacción entre el videojuego y el jugador es clave a la hora de construir estas herramientas (elemento *Interfaz e interacción* de la figura 1.1). El diseño de la interfaz de usuario debe adaptarse también a los problemas cognitivos de esta población. Por ello, es necesario analizar las particularidades de esta población a la hora de interactuar con el videojuego así como validar que las decisiones de diseño tomadas se adecuan a esta población. Con respecto a la interacción entre el jugador y el videojuego, se han de aplicar conceptos de usabilidad generales adaptando lo necesario para paliar las dificultades que presentan las personas con síndrome de Down a la hora de interactuar con dispositivos electrónicos. La presencia del terapeuta puede ayudar en el caso de que se produzcan bloqueos o desmotivación por parte de los jugadores, motivando y facilitando el uso de la herramienta a estos últimos. Además, el terapeuta también tiene que interactuar con el videojuego para evaluar algunas actividades, por lo que se debe diseñar una estrategia de interacción que no sea intrusiva para los jugadores.

Por otro lado, es importante definir los criterios de evaluación de las actividades realizadas por los jugadores y cómo estos criterios son implementados dentro del videojuego (elemento *Evaluación* de la figura 1.1). Un aspecto clave

es la evaluación de las actividades de producción oral que se desarrollen durante el transcurso del videojuego, ya que son las más complicadas de evaluar debido a que hay que extraer la información de una señal sonora y analizarla. Por este motivo, es necesario analizar las características acústicas de las producciones realizadas por los jugadores y relacionarlas con los criterios de validez definidos para una producción oral correcta, específicamente los relacionados con la prosodia.

Con respecto a la adaptación al usuario (elemento *Adaptación al usuario* de la figura 1.1), la narrativa diseñada debe ajustarse a las limitaciones cognitivas que presentan estas personas, de manera que sea motivante para continuar jugando y que sientan que las actividades insertadas respetan el contexto narrativo donde se están desarrollando. La inclusión de elementos del mundo real, como escenarios, objetos o personajes que pueden encontrarse en la vida cotidiana de estas personas, puede ayudar a la transferencia de los conocimientos aprendidos durante el videojuego a las situaciones reales. Estos dos puntos son claves en la motivación y disposición de los jugadores hacia el videojuego. Además, esta adaptación se refiere a la adecuación de las diferentes actividades del juego al nivel del usuario, teniendo en cuenta tanto la información de la sesión de juego en curso como la de sesiones anteriores.

El marco de trabajo incluye adicionalmente otro aspecto a tener en cuenta, como es la generación de informes de las sesiones de juego (elemento *Generación de informes* de la figura 1.1). La generación de informes se enfoca en la realización de informes sobre las sesiones de juego del jugador, de manera que el terapeuta pueda consultar los resultados de las diferentes sesiones de juego y observar la evolución de los jugadores a lo largo del tiempo. Esta información puede ser útil de cara a utilizar la herramienta en sesiones reales de terapia del habla.

1.4. Metodología

En esta sección, se expone la metodología empleada para alcanzar los objetivos planteados en la sección 1.2. El trabajo realizado está descrito en profundidad en cada uno de los artículos del compendio de publicaciones incluidos en esta tesis (ver sección 2).

Para alcanzar los objetivos O1.1 y O1.2, relacionados con el diseño del videojuego, la evaluación de las actividades y de la interfaz, se desarrolla la metodología representada en la figura 1.2, extraída del primer artículo del compendio de publicaciones incluido en esta tesis, sección 2.1. La definición del videojuego se basa en un diseño centrado en el usuario, participando tanto expertos en síndrome de Down como los propios usuarios y terapeutas. La narrativa y las actividades incluidas en la misma son diseñadas con el apoyo de expertos en síndrome de Down y en terapia del habla. Para validar los diseños de la interfaz de usuario, se ejecutan técnicas como el *recorrido cognitivo* [31; 36; 40], que consiste en ponerse en el lugar del usuario de la herramienta y realizar un recorrido de la interacción prevista sobre prototipos iniciales del diseño de la interfaz. De esta manera se pueden detectar tempranamente algunos errores del diseño. Esta fase cobra especial importancia cuando el usuario objetivo tiene algún tipo de limitación cognitiva, ya que es necesario que la interfaz sea lo más intuitiva posible para que estos usuarios sean capaces de usarla fácilmente. Una vez implementada la herramienta, se realiza una evaluación de la usabilidad con usuarios reales, teniendo en cuenta los aspectos definidos por [42], que son la facilidad de uso, la efectividad, la eficiencia, la tolerancia a errores y la disposición hacia el uso del videojuego. Estas evaluaciones se realizan con 14 usuarios con síndrome de Down, 10 niños con desarrollo típico y 10 adultos sin discapacidad intelectual. Para evaluar cada aspecto se utilizan diferentes estrategias. Por un lado, se realiza un análisis de la interacción de personas con síndrome de Down con el videojuego, comparándola

con la interacción realizada por personas con desarrollo típico (niños y adultos). Para ello, se utilizan los datos recopilados por el juego para medir y comparar diferentes aspectos de la interacción, como son los tiempos requeridos para completar la actividad, los resultados de dicha actividad, las ayudas proporcionadas por el videojuego o el número de clicks realizados. Para realizar esta comparación, se utilizan test estadísticos no paramétricos, con el objetivo de analizar si las diferencias entre los datos de las sesiones de juego de personas con síndrome de Down y de las personas con desarrollo típico son significativas estadísticamente o no. Además, se efectúan una serie de cuestionarios a los jugadores para identificar las sensaciones que los jugadores tuvieron con el videojuego. Las propias observaciones realizadas por el equipo de investigación durante las sesiones de juego también se utilizan para detectar los puntos fuertes y débiles del videojuego. Por otro lado, se realizan entrevistas a los profesores que asistieron a las sesiones de juego para analizar sus impresiones y observaciones sobre el uso del juego por parte de los participantes con síndrome de Down. Por último, para obtener unos indicios de la eficacia y efectividad del videojuego para entrenar la prosodia, se efectúan una serie de tests perceptuales comparando las producciones de los jugadores en las primeras sesiones con el videojuego con las realizadas en las últimas sesiones.

Para alcanzar el objetivo O2.1, relacionado con la identificación de las características acústicas y prosódicas que caracterizan el habla de las personas con síndrome de Down, se sigue la metodología mostrada en la figura 1.3, extraída del segundo artículo del compendio de publicaciones incluido en esta tesis, sección 2.2. Las grabaciones obtenidas durante algunas sesiones con el videojuego (349 frases obtenidas de 18 usuarios con síndrome de Down) se utilizan para extraer características acústicas relacionadas con los dominios de la frecuencia, de la energía, temporal y espectral. Para poder realizar una comparación entre el

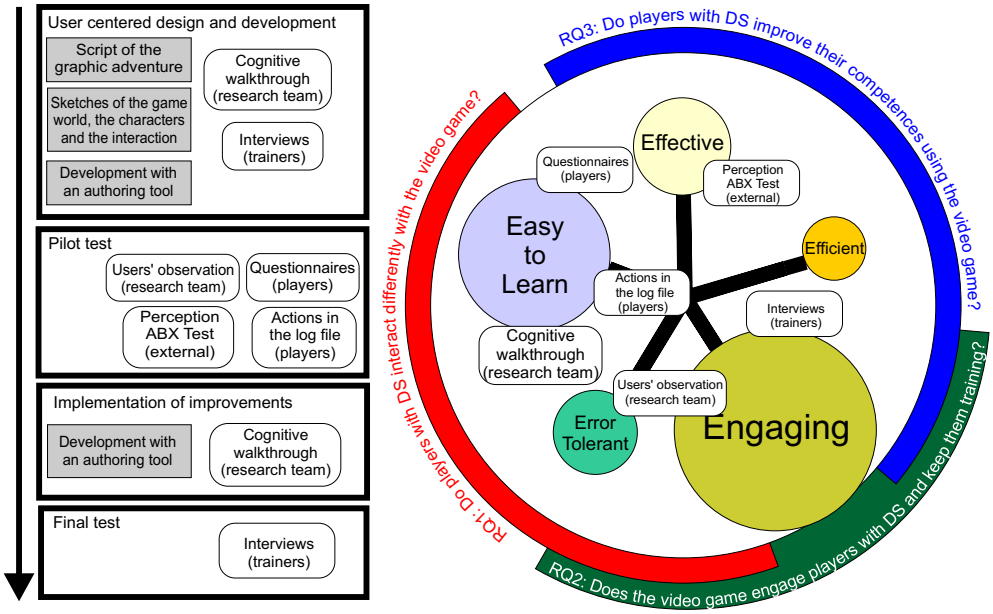

Figura 1.2: Estrategia de evaluación del videojuego. Los 5 círculos representan las 5 dimensiones relacionadas con la usabilidad que se evalúan en el videojuego y los rectángulos representan las diferentes estrategias de evaluación desarrolladas. El tamaño de los círculos representa la importancia de cada dimensión dentro del estudio realizado.

habla de personas con síndrome de Down y el habla de personas con desarrollo típico, se llevan a cabo unas grabaciones de las mismas frases del juego, pero grabadas por personas adultas sin discapacidad intelectual (250 frases obtenidas de 22 usuarios sin discapacidad intelectual). Una vez obtenidas las grabaciones, se extraen de ambos corpus una serie de características acústicas utilizando el software openSMILE [19] y el conjunto de características eGeMAPS [18], que contiene características de cada uno de los dominios a estudiar. Para encontrar las características más discriminativas para diferenciar voces de personas con síndrome de Down y voces de personas con desarrollo típico, se comparan las características extraídas de las grabaciones de ambos grupos. De esta comparación surgen las características con mayor capacidad de discriminación entre los dos grupos, utilizando para ello tests estadísticos. Además, se entrenan tres clasificadores binarios utilizando las características con diferencias significativas y se realiza una com-

paración entre las características de los diferentes dominios para analizar cuál es más relevante a la hora de clasificar una voz como proveniente de una persona con síndrome de Down. Para realizar la clasificación, se utiliza la técnica de validación cruzada, que consiste en dividir todas las muestras en dos subconjuntos mutuamente excluyentes. En este caso, el subconjunto de entrenamiento contiene el 90 % de los datos y el subconjunto de prueba tiene el 10 % restante. El subconjunto de entrenamiento es utilizado para entrenar al clasificador y el subconjunto de prueba es utilizado para evaluar dicho clasificador, obteniendo una tasa de precisión del clasificador con respecto al subconjunto de prueba. La valoración se realiza comparando la evaluación efectuada por la experta en prosodia frente a la evaluación producida por el clasificador. Este proceso se repite 10 veces creando subconjuntos diferentes en cada repetición y posteriormente se realiza la media de la precisión obtenida en cada repetición. Como valores de calidad del clasificador, se utilizan la precisión (porcentaje de elementos clasificados correctamente frente al total de elementos) y el UAR (Unweighted Average Recall), que consiste en calcular la media del porcentaje de elementos de cada tipo clasificados correctamente. Esta última medida es especialmente útil cuando los datos utilizados no están balanceados. Finalmente, se realiza un test perceptual utilizando grabaciones creadas utilizando un algoritmo de transferencia de prosodia entre algunas grabaciones de personas con síndrome de Down y otras grabaciones de personas con desarrollo típico. Para ello, se transfieren, fonema a fonema, el tono, la energía y la duración de una grabación hacia otra perteneciente al grupo contrario. Posteriormente, varias personas sin conocimientos específicos en prosodia evalúan cada grabación utilizando una escala Likert de 5 puntos, indicando la seguridad con la que identifican cada grabación como perteneciente a una persona con discapacidad intelectual. Este test tiene como objetivo analizar perceptualmente la importancia de la prosodia a la hora de identificar una voz co-

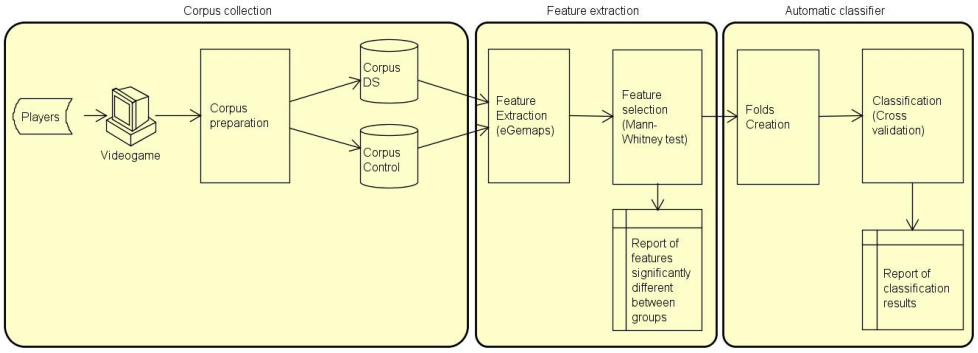

Figura 1.3: Metodología desarrollada para realizar la comparación entre las voces de personas con síndrome de Down y las voces de personas con desarrollo típico.

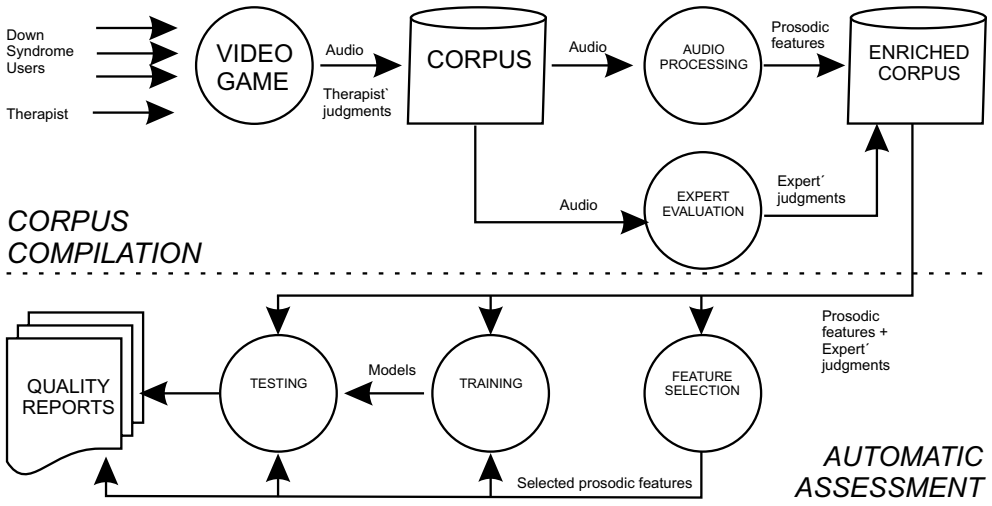

Figura 1.4: Metodología desarrollada para el análisis de una evaluación automática de las grabaciones del videojuego.

mo proveniente de una persona con síndrome de Down, y complementa el análisis de las características acústicas extraídas de las grabaciones.

Finalmente, para la consecución del objetivo O2.2, relacionado con la evaluación automática de las producciones del videojuego, se sigue la metodología definida en la figura 1.4, extraída del tercer artículo del compendio de publicaciones incluido en esta tesis, sección 2.3. Para este estudio se utilizan tres corpus de grabaciones del videojuego, pero obtenidos en momentos diferentes. El primer corpus contiene grabaciones de 5 personas con síndrome de Down realizadas en

4 sesiones con el videojuego, conteniendo 605 grabaciones en total. El segundo corpus contiene grabaciones de otras 5 personas con síndrome de Down, con un total de 168 grabaciones. El tercer corpus contiene grabaciones obtenidas con una versión previa del videojuego, con un total de 193 grabaciones procedentes de 13 personas con síndrome de Down. Estas grabaciones son evaluadas por una experta en prosodia, basándose en criterios como la entonación, el acento, el ritmo o la composición de las frases. Siguiendo estos criterios, la experta utiliza una evaluación binaria, decidiendo si la grabación tiene buena calidad o no. Además, algunas de las grabaciones del corpus también están evaluadas por terapeutas que realizaron las sesiones de juego junto a los usuarios con síndrome de Down. De las grabaciones ya evaluadas, se extraen una serie de características acústicas y prosódicas, tales como características de frecuencia, energía y temporales, utilizado el software openSMILE y el conjunto eGeMAPS. Con estas características se entrenan una serie de clasificadores automáticos (*Decision trees, multilayer perceptron y support vector machines*) con el objetivo de predecir las evaluaciones realizadas por la experta en prosodia. El proceso de entrenamiento y prueba de los clasificadores es el mismo descrito anteriormente, utilizando la técnica de validación cruzada. Para analizar las características más importantes para clasificar las grabaciones en 5 usuarios concretos, se realiza un ranking de características basándose en la importancia de cada característica aislada para clasificar las grabaciones de cada uno de estos 5 usuarios, usando para ello el *Área bajo la curva ROC*, que es una medida de precisión basada en variar el umbral utilizado por los clasificadores para decidir si una muestra pertenece a un tipo o al otro. Este análisis tiene como objetivo el mostrar como la heterogeneidad de la población con síndrome de Down influye en los resultados de clasificación. Con este mismo objetivo, se comparan las evaluaciones de las dos evaluadoras y los resultados de un clasificador para estos mismos 5 usuarios. Por último, se calcula la correla-

ción de los perfiles psicológicos (Test de vocabulario en imágenes Peabody [16], Escala de inteligencia de Wechsler para niños [7] y Test de matrices progresivas Raven [43]) y los resultados del test de habilidades prosódicas PEPS-C [34] con los resultados de las evaluaciones y de un clasificador para estos 5 usuarios.

1.5. Organización de la memoria

La presente memoria se divide en los siguientes apartados. El presente capítulo se divide en cuatro secciones. La sección 1.1 describe los problemas en el uso de la prosodia de las personas con síndrome de Down, las soluciones existentes para paliar estas dificultades y las propuestas que este trabajo de tesis plantea para abarcar las limitaciones que estas soluciones presentan. La sección 1.2 describe los objetivos iniciales de esta tesis en relación a las limitaciones encontradas previamente. La sección 1.3 detalla el marco de trabajo propuesto y la sección 1.4 especifica la metodología a desarrollar para alcanzar los objetivos definidos.

El capítulo 2 incluye los tres artículos del compendio de publicaciones que se presentan en esta tesis. Cada publicación se centra en uno de los objetivos de esta tesis. La primera publicación, incluida en la sección 2.1, contiene un análisis en profundidad del estado del arte en relación a los problemas en el lenguaje de las personas con síndrome de Down, la utilización de herramientas tecnológicas para entrenar la prosodia y el uso de videojuegos para entrenar diferentes habilidades en personas con síndrome de Down. Además, se describe en detalle el diseño del videojuego, con todos los elementos que lo componen y la justificación de su inclusión. También se describen en detalle las actividades de entrenamiento incluidas en el videojuego, tanto las centradas en la prosodia como otro tipo de actividades. Por último, se detalla la estrategia de evaluación del videojuego, los resultados obtenidos en relación a la interacción con el videojuego, la motivación

hacia su uso y la mejora en la pronunciación, y una discusión sobre los resultados obtenidos. La segunda publicación, incluida en la sección 2.2, describe el estado del arte en relación al uso de la prosodia por personas con síndrome de Down y sus diferencias con respecto al uso de la prosodia por personas con desarrollo típico. Las diferencias descritas se centran principalmente en características extraídas de la señal acústica, como la frecuencia fundamental, la energía, el ritmo o las relacionadas con el espectro de la señal. Además, se describe el proceso experimental para identificar automáticamente una voz como proveniente de una persona con síndrome de Down o de una persona con desarrollo típico utilizando ciertas características acústicas extraídas de las grabaciones del videojuego. Por último, se describen los resultados obtenidos y una discusión sobre estos resultados. La tercera publicación, incluida en la sección 2.3 describe los trabajos existentes en evaluación automática de la calidad de una grabación en voces patológicas, así como las dificultades que entraña esta tarea. También se detallan los experimentos realizados para evaluar automáticamente la calidad prosódica de las grabaciones del videojuego. Estos experimentos se basan en obtener una evaluación inicial realizada por expertas en prosodia e intentar reproducir esta evaluación utilizando algunas características acústicas extraídas de las grabaciones y entrenando clasificadores automáticos. Por último, se presentan los resultados de estos experimentos y la discusión sobre los mismos.

El capítulo 3 presenta un resumen y discusión de los resultados más importantes surgidos de la investigación realizada, estraídos de las publicaciones del compendio incluido en esta tesis. Estos resultados son consecuencia de varios experimentos realizados con el fin de cumplir con los objetivos propuestos en el capítulo 1. En la sección 3.1, se muestran los resultados de la evaluación con usuarios reales del videojuego. En esta sección se muestran algunas estadísticas de uso del videojuego para cada tipo de usuario (síndrome de Down, adultos

sin discapacidad, niños sin discapacidad) y se comparan, además de mostrar los resultados de los cuestionarios realizados a los usuarios al final de las sesiones de juego. Con estos resultados se analiza si el videojuego consigue motivar a los jugadores con síndrome de Down para realizar las tareas de entrenamiento de la prosodia y se observa si las actividades y el diseño propuesto son apropiados para esta población. En la sección 3.2 se muestra la comparación entre algunas características acústicas relacionadas con la prosodia y extraídas de las grabaciones almacenadas por el videojuego para dos tipos de usuarios, jugadores con síndrome de Down y jugadores con desarrollo típico. En la sección 3.3 se muestran los resultados de la evaluación automática del uso de la prosodia de las grabaciones del videojuego centrándose en los cinco usuarios de los que se tienen el perfil psicológico y sus resultados en un test de evaluación de la prosodia. En la sección de resultados se incluye también una discusión sobre los mismos en relación a los objetivos definidos.

Por último, en el capítulo 4 se incluyen tres secciones. La sección 4.1 describe las conclusiones del trabajo realizado basándose en los objetivos definidos y en los resultados obtenidos en los diferentes experimentos realizados. En la sección 4.2 se describen algunos aspectos que quedan pendientes de explorar en relación con el trabajo realizado y otros que se han estudiado pero en los que se puede profundizar más. Por último, la sección 4.3 enumera las publicaciones y comunicaciones en congresos surgidos del trabajo de esta tesis, el corpus final recopilado, una descripción del videojuego utilizado para grabar el corpus y realizar la evaluación de la interacción con los jugadores y los proyectos de donde deriva el trabajo presentado en esta memoria de tesis.

Capítulo 2

Compendio de publicaciones

2. COMPENDIO DE PUBLICACIONES

En este capítulo se muestran las publicaciones del compendio incluido en esta tesis en su formato original, según fueron publicados en las tres revistas donde se incluyen estas publicaciones. El primer artículo, titulado *Engaging Adolescents with Down Syndrome in an Educational Video Game* fue publicado en la revista *International Journal of Human Computer Interaction*, en 2017 (ver sección 2.1). En este artículo, el autor de esta tesis participó desarrollando el software utilizado en la investigación, realizando las evaluaciones del videojuego con los usuarios de los colegios de educación especial, procesando los datos almacenados durante las sesiones de juego, redactando algunas secciones del propio artículo y revisando el contenido final. El segundo artículo, titulado *Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome* fue publicado en la revista *Speech Communication*, en 2018 (ver sección 2.2). El tercer artículo, titulado *Automatic Assessment of Prosodic Quality in Down Syndrome: Analysis of the Impact of Speaker Heterogeneity*, fue publicado en la revista *Applied Sciences*, en 2019 (ver sección 2.3).

2.1. Engaging Adolescents with Down Syndrome in an Educational Video Game

C. González-Ferreras[1], D. Escudero-Mancebo[1], M. Corrales-Astorgano[1], L. Aguilar-Cuevas[2], and V. Flores-Lucas[3], "Engaging adolescents with Down syndrome in an educational video game". International Journal of Human–Computer Interaction, vol. 33, no. 9, pp. 693–712, 2017.

[1]*Departamento de Informática, Universidad de Valladolid, Valladolid, Spain*

[2]*Departamento de Filología Española, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain*

[3]*Departamento de Psicología, Universidad de Valladolid, Valladolid, Spain*

https://doi.org/10.1080/10447318.2017.1278895

2. COMPENDIO DE PUBLICACIONES

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
http://dx.doi.org/10.1080/10447318.2017.1278895

Taylor & Francis
Taylor & Francis Group

Engaging Adolescents with Down Syndrome in an Educational Video Game

César González-Ferreras[a], David Escudero-Mancebo[a], Mario Corrales-Astorgano[a], Lourdes Aguilar-Cuevas[b], and Valle Flores-Lucas[c]

[a]Departamento de Informática, Universidad de Valladolid, Valladolid, Spain; [b]Departamento de Filología Española, Universitat Autònoma de Barcelona, Barcelona, Spain; [c]Departamento de Psicología, Universidad de Valladolid, Valladolid, Spain

ABSTRACT
This article describes the design, implementation and evaluation of an educational video game that helps individuals with Down syndrome to improve their speech skills, specifically those related to prosody. Special attention has been paid to the design of the user interface, taking into account the cognitive, learning, and attentional limitations of people with Down syndrome. The learning content is conveyed by activities of production and perception of prosodic phenomena, aimed at increasing their communicative competence. These activities are introduced within the narrative of a video game so that the players do not conceive the tool as a mere succession of learning activities, but so that they learn and improve their speech while playing. The evaluation strategy that has been followed involves real users and combines different evaluation activities. Results show a high level of acceptance by participants and also by professionals, speech therapists, and special education teachers.

1. Introduction

Children and adolescents with Down syndrome (DS) are digital natives that use the information and communications technologies (ICT) with relative ease. They enjoy interacting with computers and mobile devices, browsing the web and playing with video games, just like any other people (Feng, Lazar, Kumin, & Ozok, 2010). There have been experiences that seek to exploit this fact to test ICT educational tools with this group of users (Ng, Bakri, & Rahman, 2014). Although in many cases they can use the same tools as the rest of the users without any kind of adaptation, people with DS present characteristics that determine the effectiveness with which they can use the ICT educational tools (Pazos González, Raposo-Rivas, & Martínez-Figueira, 2015). Among the characteristics that limit the effectiveness of the ICT educational tools is the fact that many people with DS present attention deficit (Hernández Martínez, Pastor Duran, & Navarro Navarro, 2011), low tolerance to frustration, delayed language development (Chapman, Schwartz, & Bird, 1991), difficulty to process information that comes from different channels simultaneously (Lanfranchi, Cornoldi, Vianello, & Conners, 2004), lack of motivation (Wuang, Chiang, Su, & Wang, 2011), problems with the short term memory (Chapman & Hesketh, 2001), and difficulties in understanding the meaning of iconic symbols (Yussof & Zaman, 2011). All these limitations are an important challenge, taking into account the fact that our objective was to build a tool for speech training, thus requiring the development of a multimodal interaction system with both audio and visual input and output.

There is a well developed literature on the potential of games to improve motivation and engagement in education (McFarlane, Sparrowhawk, & Heald, 2002), but limited information on the advantages of using them to improve the speech production and prosodic skills of people with Down syndrome (Kent & Vorperian, 2013). Although some tools do exist (Saz et al., 2009), they are not effectively used by people with DS, since a high degree of motivation is required, which is not easily achievable by this type of users.

In this article, we describe the design and evaluation of an educational video game for speech training that has been designed taking into account the specific limitations of people with DS, in order to increase their chances of success in the formative tasks. A combination of design decisions that affect human–computer interaction (HCI) allowed us to achieve an increase in user motivation, and thus, an improvement of the effectiveness of the training sessions. The main idea is to apply typical elements of computer games, such as feedback and guidance (Kapp, 2012), in the development of an educational video game for speech training for people with DS called "The Magic Stone".

"The Magic Stone" is a serious video game grounded in the genre of graphic adventure video games, in which the player assumes the role of the main character in an interactive story driven by exploration and problem-solving. Training activities are inserted into the video game and users must perform the activities in order to continue advancing in the adventure. Moreover, some activities force players to interact with game characters using their voice, which helps them to feel included in the story. The use of a narrative where the learning activities are included is what makes the game immersive. The idea is that the users feel like they are in the world that the game

CONTACT César González-Ferreras cesargf@infor.uva.es Departamento de Informática, Universidad de Valladolid, Paseo de Belén 15 47011, Valladolid, Spain.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hihc.

creates, so they do the learning activities in an unnoticed way. Besides this, well established learning objectives concerning prosodic skills allow the progress of the users to be monitored.

The evaluation strategy of the video game was developed to answer three main research questions (RQ):

- **RQ1:** Do players with DS interact differently with the video game than people without intellectual disabilities? In spite of the fact that children and adolescents with DS are digital natives who are used to interacting with computers, we assess whether there are usability aspects affected by their special characteristics that condition the training performance.

- **RQ2:** Does the video game engage players with DS and keep them training? We evaluate whether the HCI elements included in the video game sufficiently motivate the users to keep on doing the training activities.

- **RQ3:** Do players with DS improve their communication skills using the video game? We evaluate whether the activities of the video game are effective from the point of view of speech training.

The results of the pilot test show that the special characteristics of users with DS cause a different use of the software compared to users without DS. At the same time, the conducted tests highlight the fact that the abundance and type of reinforcements used allow effective training to be created for pronunciation improvement. The results permit us to discuss how the HCI techniques employed have allowed us to address the users' attention deficit, memory difficulties and low tolerance to frustration. We also discuss how the concurrent interaction between teacher and the student while using the program is a key factor in obtaining a good training performance. The system was also tested by people without intellectual disabilities, with the aim of verifying if players with DS interact differently with the video game than people without intellectual disabilities.

The structure of the article is as follows. Section 2 reviews related works of the state of the art. Section 3 describes the features of the video game. Section 4 explains in detail the activities that must be done by the users. Section 5 presents the evaluation strategy and Section 6 shows the results. We discuss in Section 7 how the HCI elements introduced in the video game help to increase the users' motivation. Finally, the conclusions are presented in Section 8.

2. Background and related work

Several sources of knowledge are needed to design an educational tool addressed to people affected by DS. It is first of all necessary to know the characteristics of the Down syndrome, since their difficulties should be considered in the development of the tool (Section 2.1). Second, since the video game focuses on prosodic skills, an overview of prosody is also necessary, as well as the existence of tools that help in the learning of prosody (Section 2.2). Third, specific software and video games for training people with DS has to be reviewed in order to know which kind of solutions have been adopted in the state of the art (Section 2.3).

2.1. Down syndrome and language difficulties

Down syndrome, according to the new diagnostic classification system DSM-5 (American Psychiatric Association, 2013), is a subtype of intellectual developmental disorder (IDD). IDD is characterized by significant limitations both in intellectual functioning (with an intelligence quotient -IQ- score equal to or below 70, with cognitive difficulties) and in adaptive behavior, showing difficulties in conceptual, social and practical adaptive skills. Besides, this condition appears before age 18 (Schalock et al., 2010).

Down syndrome is not the most common cause of IDD, but it is the most frequent of genetic origin. It is not a hereditary disorder and occurs equally in all ethnic groups, cultures and social classes. Although a clear cause is unknown, some risk factors have been established, among which the most widely agreed is the age of the mother.

People with DS present profiles of development and difficulties which are different in terms of different variables, such as IQ score, severity, etc. However, they generally present important difficulties in their language skills, with better language comprehension than production (speech). Among their most common language difficulties are: articulation, in some cases their speech is almost unintelligible; phonological discrimination; more severe difficulties with the acquisition of morphological and syntactic aspects of the language, e.g. understanding and producing some complex morphological marks; and a very poor comprehension and production of syntax, so they do not understand or produce long and syntactically complex sentences, such as compound sentences. Although they have higher levels of lexical-semantic and pragmatic development, they also have limitations in these components: they have less vocabulary and problems to understand complex semantic relations. Even though they have good general conversational skills, they have difficulties with the use of mechanisms of discourse cohesion (anaphora, ellipsis, distinction between information already given and new information, etc.), difficulties to understand and produce indirect speech acts and figurative language, etc. (Flores Lucas & Belinchón Carmona, 2010; Kent & Vorperian, 2013; Martin, Klusek, Estigarribia, & Roberts, 2009). Individuals with DS also have limitations in the perception and production of prosody (Pettinato & Verhoeven, 2009; Stojanovik, 2011). Chapman (1997) concludes that interventions including language therapy targeting the practice of expressive language can increase communicative effectiveness.

Regarding their cognitive difficulties, perhaps one of the most severe and which compromises the following of the video game, is their memory difficulties, specifically, their difficulties to retain and process auditory and serial information (Chapman & Hesketh, 2001; Jarrold & Baddeley, 2001). They also have deficits in the integration of information, as well as in the processes of reasoning and deduction, thus making it necessary to provide more explicit information to alleviate their problems in making inferences. They also show difficulties with executive control functions (Daunhauer et al., 2014),

thus, among others, they have deficits with the inhibitory response control, that is, they issue a response impulsively without having analyzed all the information previously or taking time to stop giving a wrong answer.

2.2. Training speech prosody with ICT

When people talk or read, not only sequences of phonetic segments are articulated, but also variations of the intensity, the speaking rate and the tone. In linguistics, all this extra information is referred to as prosody, which is relevant because it conveys a huge amount of information. Prosody is used in speech with a twofold purpose: 1) to indicate the segmentation of the discourse and its organization for the benefit of the interpretation of the messages and 2) to apply different intonation and stress patterns to convey the intended meanings, going from modality differences to nuances of emotions. Moreover, the information that is transmitted by prosody cannot be replaced by lexical or grammatical elements (Martínez Celdrán & Fernández Planas, 2007).

Prosody has been managed by computers for a long time, mainly due to the contribution of prosody to improving the naturalness of TTS systems (Escudero-Mancebo, González-Ferreras, & Cardeñoso-Payo, 2002; Taylor, 2009). There are algorithms that permit the acoustic signal of speech to be processed so as to code the relevant information encoded by prosody (Aguilar, Bonafonte, Campillo, & Escudero, 2009; Escudero, Cardeñoso, & Bonafonte, 2002; González-Ferreras, Escudero-Mancebo, Vivaracho-Pascual, & Cardeñoso-Payo, 2012). These types of algorithm are used by several tools, like the ones presented by Saz et al. (2009) for training speech and prosody. Video games are especially useful for studying and practising prosody with students affected by DS as they allow the incorporation of audio files to capture the differences between pronunciations (in order to work in the perception domain) and to record their own productions (in order to work in the production domain). Perception activities focus on the discrimination between meanings that cannot be attributed to lexical or grammatical elements (for instance, differences in the expression of emotions), while production activities focus on acquiring utterances pronounced by the student until a correspondence with the model is achieved.

There are tools that help in the learning of prosody. For instance, a software that augments text with visual prosodic cues to improve expressive reading (Patel, Kember, & Natale, 2014). The idea is to provide cues of the underlying prosody: pitch, duration and intensity. These authors carried out an experiment with children whose results suggest that beginning readers could benefit from explicit visual prosodic cues and thus enhance oral reading expressiveness.

Other tools incorporate speech technologies to assist speech therapy with patients that have a speech disorder. For example, Shahin et al. (2015) describe a system for automated speech therapy in childhood apraxia of speech (CAS). The system is able to identify the three main types of error commonly associated with CAS: groping errors (delay in sound production), articulation errors (incorrect pronunciation of phones) and prosodic errors (inconsistent lexical stress). Rodríguez, Saz, and Lleida (2012) present another computer program for providing speech therapy to individuals with speech disorders. The tool is used to train such speech skills as voice production, intensity, blow, vocal onset, phonation time, tone and vocalic articulation. An experimental study of 27 subjects, the majority with cognitive disabilities, showed improvements in the voice capabilities of a remarkable number of users. Also interesting was the finding that the use of the program motivated the subjects, because they were attracted by the game-like applications. The tool consists of a set of game-like applications that are independent from each other. The main difference with our work is that "The Magic Stone" is designed as a graphic adventure game, and thus, all the activities are integrated into a story. This allows the different linguistic activities to be contextualized and increases the player's sense of immersion in the video game.

2.3. ICT and video games for training people with Down syndrome

Children with DS are able to use computers in their daily lives to perform many tasks. A study about how children and young adults with DS use computers (Feng, Lazar, Kumin, & Ozok, 2008) found that 83% of the 561 participants started using computers at six years old. The majority of the children use the computer for learning (80%) and entertainment (95%). Educational software, video games and programs for the Internet were the most used applications. They spent an average of 3.5 hours a week using the computer at school and 4.94 hours a week using the computer at home.

The potential of games to improve motivation and engagement in education has been examined (McFarlane et al., 2002). The idea is to apply typical elements of computer games, such as feedback, guidance, time pressure, and rewards (Kapp, 2012). Moreover, these elements can be included in non-gaming systems to improve user experience and engagement (Darejeh & Salim, 2016; Deterding, Sicart, Nacke, O'Hara, & Dixon, 2011). Games allow learners the chance to practice in environments close to real world situations. Rather than applying learning in a vacuum, games provide a context for learning, allowing players to apply what they have learned to solve real life problems (Prensky, 2007). Modern theories of effective learning suggest that learning is most effective when it is active, experiential, situated, problem-based and provides immediate feedback (Boyle, Connolly, & Hainey, 2011). Video games can include activities which have these features.

Graphic adventures are a video game gender that use graphics to convey the environment in which a story is developed during the game. Learners remember information more easily when they learn that information in the form of a story rather than from a bulleted list (Adams, 2013). Stories evoke emotions and provide a context for placing information. Involving a learner in a story can make the learning more powerful. A well-crafted story focuses on helping learners to solve problems, educates the learners, and is easily recalled when the actual situation arises or when a learner is in a similar situation.

In addition, there are some studies that show the efficiency of ICT and video games in the cognitive

2.1 Engaging Adolescents with Down Syndrome in an Educational Video Game

rehabilitation and teaching of people with intellectual disability: improvement in choice reaction time (Standen, Anderton, Karsandas, Battersby, & Brown, 2009), stimulating cognitive abilities of children (Brandão et al., 2010), independent decision making (Standen, Rees, & Brown, 2009) and improvement in mathematical skills (Brown, Ley, Evett, & Standen, 2011; Bruno et al., 2003; González, Noda, Bruno, Moreno, & Muñoz, 2015; Ortega-Tudela & Gómez-Ariza, 2006; Shafie et al., 2013; Wan Ahmad, Muddin, & Shafie, 2014). The program Phonics Alive! (Phonics Alive!, 1996) was used in (Gallaher, van Kraayenoord, Jobling, & Moni, 2002) to teach reading to a young adult with DS. Tangible interfaces are used in (Santana & Muro, 2011) to implement the reading method described in (Troncoso & Del Cerro, 1999), which is specific for children with DS. Torrente, Del Blanco, Moreno-Ger, and Fernández-Manjón (2012) developed two games whose main aim is to promote independent living for people with intellectual disability, such as interacting with others at work or the use of public transportation. A similar idea is presented in (Lopez-Basterretxea, Mendez-Zorrilla, Garcia-Zapirain, Madariaga-Ortuzar, & Lazcano-Quintana, 2014), where players learn how to go shopping. Furthermore, other games are focused on therapy for children with speech disorders (Cagatay, Ege, Tokdemir, & Cagiltay, 2012). Reviews of educational software for people with DS can be found in (Black, 2006; Ng et al., 2014; Pazos González et al., 2015).

On the other hand, the evaluation of the user interface is important to guarantee the usability of the educational tool. However, there is little research in human–computer interaction literature related to people with DS. Lazar, Kumin, and Feng (2011) reported an ethnographic observation that examined computer skills of expert users with DS and Kumin, Lazar, Feng, Wentz, and Ekedebe (2012) carried out a usability evaluation of multi-touch tablet devices by adults with DS. Both studies report the great diversity in cognition, communication, skills, capabilities and computer experience of people with DS.

The direct involvement of people with DS during the design process is very challenging, due to their communication, cognitive and behavioral difficulties. One solution would be for designers to work with parents and carers as proxies for the user. Designers could also observe the interaction of the users with prototypes. There are some examples of this approach in the literature. A system that helps users with DS to learn addition and subtraction is described in (González et al., 2015). Different types of experts were involved in the design process (teachers, experts in usability, and experts in Down syndrome). Several prototypes were evaluated by the experts and redesigned with the proposed modifications. The final system was evaluated both by experts and by users with DS. Carmien et al. (2005) presented a system for supporting people with cognitive disabilities using public transportation. The interaction of people with cognitive disabilities was observed and interviews with specialists who help them were carried out. In the testing stages, there was a direct involvement of the users.

3. System description

"The Magic Stone" is a video game whose main aim is to help people with Down syndrome to improve communication skills that have been affected due to their disability, especially those related with prosody. To do this, players will have to do some activities related with prosody and others not related with language, such as puzzle-solving, that have been introduced to add variety to the game. Another aim of the video game is that it can be used by teachers or speech therapists in special education centers as part of their educational activities.

The video game has the structure of a graphic adventure game, including conversations with characters, getting and using items and navigating through different scenes. Players have to use the mouse to interact with the elements of the game. Players go through different scenes where they have to do some actions, like doing an activity or using an item. The main innovation is that players have to record sentences using their voice in some activities. The speech therapist or teacher decides if the player has done it correctly or not. In other activities, an audio is played in the context of a conversation and the player has to choose between some options to continue this conversation.

At the beginning of the game, the configuration screen is shown, where players can introduce a user name and configure their game profile. They can select an avatar, which represents the player in the game, the difficulty level and reader profile (reader/non reader). Then, a video is played to introduce the story of the game.

The plot of the adventure is a mysterious story in which the player takes the role of a hero who is responsible for saving the city from environmental destruction. To do this, the player has to retrieve the magic stone which has been stolen and which will restore order in the city. The search for the magic stone requires the user to carry out several activities in order to progress in the game. Four activity types are defined to practice speech, communication and prosodic skills. The activities will be described in Section 4.

The development of the video game is based on multimedia learning, presenting visual and verbal materials together simultaneously (Mayer, 2002). In this way, the users do not depend only on the textual channel to receive important information, as it can be completed by information with images, because this information modality has lower difficulties for people with intellectual disabilities (Chapman & Hesketh, 2001). In fact, it has been shown that using images to support and complement verbal information is a better educational strategy than using only verbal information (Buckley & Sacks, 2002). Other studies using multimedia learning systems have shown that it is an effective method for this population (Khan, 2010).

3.1. Learning objectives

The learning objectives of the video game are to:

- Perceive and discriminate the different sentence types: declarative, interrogative and exclamatory.

- Identify the correct sentence type (declarative, interrogative or exclamatory) to use in a particular communicative exchange.
- Associate the different sentence types with the corresponding prosodic patterns.
- Produce the appropriate prosodic pattern according to the type of sentence.
- Produce sentences keeping the rhythm of the utterance, respecting the location and duration of pauses and the form of intonation.
- Control the volume and intensity.

3.2. System architecture

Figure 1. shows the system architecture. Two users interact with the system: the player and the trainer. The player is normally a person with language deficits, specifically in prosodic comprehension and production. The trainer is typically a helper (teacher, speech therapist, family) that helps the player during game sessions. When trainer and player are working together on a game activity, the trainer will help the player in the correct use of voice and also to configure the system. The trainer evaluates the player's recordings in production and prosodic activities, making the player repeat the exercise when the result is not correct. To evaluate the player's recordings, the trainer uses the keyboard of the same computer that the player is using. Therefore, the trainer has to sit next to the player. 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 and helps them to solve such difficulties as understanding the story and the activities.

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.

The application has a multimodal interface, as much for input as for output. Input is performed using voice and mouse. On voice training exercises, players will have to use their voice. Output is performed using visual and sound channels. The sound output channel is used for the voiceover that narrates the story, to play character voices and to play the voice of the virtual assistant. In any case, a recorded voice is used instead of TTS, as it allows emotions to be included that synthetic voices cannot yet express.

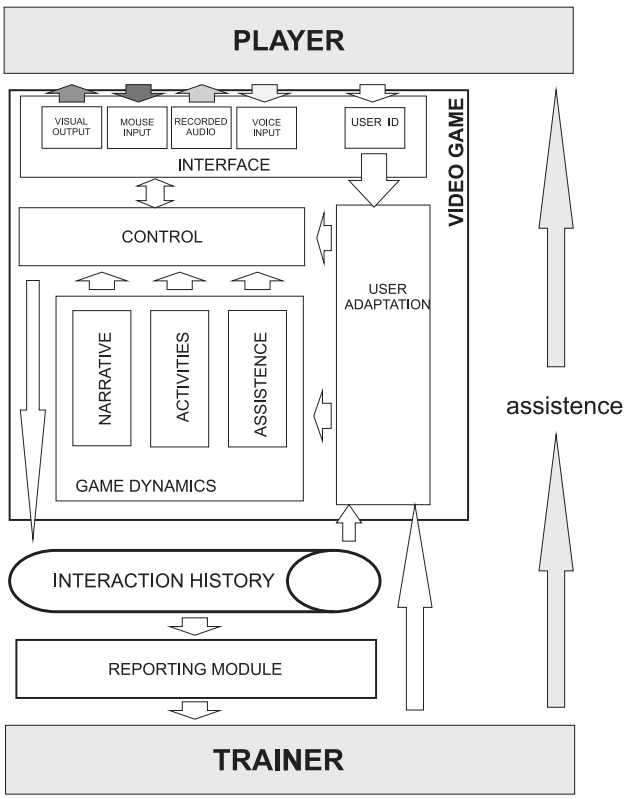

Figure 1. System architecture.

3.3. User profile

The video game encourages the improvement of skills in a gradual way. Moreover, due to the variety of users that Down syndrome includes and their respective cognitive abilities, it is necessary for the video game to be able to adapt to the particular needs of each person. Therefore, the video game includes a number of levels and reinforcements adapted to the individuality of each player.

The video game allows the trainer to define the user profile of the player. First, the trainer chooses between "reader" and "non reader", as the reading difficulty is an important aspect. The activities have more visual and audio instructions for non reader users. Second, the trainer selects the difficulty level (beginner, intermediate, advanced) which affects the complexity of the activities and the instructions that players receive.

To improve the feeling of immersion within the game, players can choose a personal avatar at the beginning of the game. It represents the players' image in the game, allowing players to identify themselves with the character of the story. This is important because promoting the presence of the player in the game is necessary for the completion of the maximum number of activities.

3.4. Virtual assistant

A virtual assistant is used to guide the player during the game. This assistant is a parrot, as shown in Figure 2. This assistant tells players to which place they must go or which object they must find or use. It also reminds players of the current goal.

3.5. Interface design

The development of the scenes, items and characters has a uniform design, close to cartoons, but without making them too childish. Bright colors are used in accordance with the scenes represented in the game. A simple text font is used, with a larger size than usual to make it easier to read (Arial Rounded MT Bold, size 19 pt).

The instructions of the activities are formulated in a way easy to understand for the users. We use simple sentences and high frequency words, so they can understand all the words used. An expert on developmental and language disorders and on intellectual disability revised the sentences to guarantee that they are in accordance with the cognitive level of people with DS. Furthermore, in all the activities, they are provided with visual clues to facilitate the task (like the traffic lights in the prosodic production tasks or the microphone in the other speech tasks). In addition, the provided feedback is both visual and auditory.

3.6. Narrative and game immersion

The scenes, characters and items that the players find in the game are representations of real world elements. The objective of this is that players can identify daily situations and be able to transfer the game lessons to real world situations. For instance, some scenes are the player's home, a library, a bus and a bookshop. The player will meet different characters, such as the bus driver, the librarian or the bookshop assistant. Some items that the player will find are: a wallet, a book, a magnifying glass and a map. The story also includes some imaginary elements to motivate the player.

3.7. Feedback

Each activity offers the users feedback according to the results obtained. However, due to the difficulties presented by the target users, it is important not to cause frustration that can produce an abandonment of the game. For this reason, errors are dealt with in a positive way. Thus, they are always allowed to progress regardless of the results. A negative feedback is shown when they go wrong, to show them that they committed an error. However, this feedback is complemented by a positive message that helps them to keep playing and not get

Figure 2. Main game screen, in which the player's avatar (the girl) and the virtual assistant (the parrot) are visible.

demoralized, showing them that making mistakes is totally normal. Moreover, in order to not stress the users, we avoided the use of scores of any kind. For the same reason, the available time is not shown to the users.

3.8. Learner model

An important feature of educational tools is the possibility of monitoring the learning of the users. For this reason, the video game stores the interaction history of each player. Then, the reporting module analyses the sequence of operations performed in a session and provides useful information to the instructors in order to see the evolution of the students' skills along the different sessions of the game. The following information is provided:

- Total time taken to finish the game.
- All the recorded audios of the players.
- For each activity:
- Time spent to complete the activity.
- Number of errors.
- Number of additional listenings of the phrase.
- Number of speech aids.
- Number of incorrect clicks.
- Number of phrase repetitions.

4. Activities

This section describes the activities included in the video game. These activities are included in the general context of the game and players need to do them in order to progress in the game. Four activity types are defined to practice speech, communication and prosodic skills. All the activities have been planned according to the principles of learning that have proven most effective for teaching and presenting information to people with DS (Buckley & Bird, 2000, 2001). Furthermore, we use visual clues to support the verbal information, because this type of combination has proven to be the best way for people with DS.

4.1. Comprehension activities

These activities are focused on lexical-semantic comprehension and the improvement of prosodic perception in a specific context, such as asking a question or asking for something politely.

These activities introduce players to different conversations with game characters. They have to choose between a series of sentences to continue the conversation, of which only one is correct in the context (as shown in Figure 3). These sentences are played at the beginning of the activity to complement the textual information with audios, so users can get the information from two channels to prevent possible reading problems. Players have the option of replaying the sentences again. If players choose the correct sentence, the activity ends, showing a positive feedback. If players choose the wrong one, the activity plays an audio and they can choose again. If, after a given number of attempts (4 or 5 attempts, depending on the activity), players are unable to choose the correct one, the activity ends showing a negative feedback.

During the activity, the virtual assistant plays a series of audio clues to guide the player. These clues are played to explain the activity to players, when players choose an incorrect option, or when they do not do any action over a period of time (the timeout is set to 20 seconds).

The player cannot perform actions in the activity when a sound is played. This is to make audio instructions as effective as possible, to prevent possible distractions (because of their attentional deficits (Hernández Martínez et al., 2011)) and to avoid an impulsive answer without hearing or understanding the instructions (because of their deficits in executive control functions (Daunhauer et al., 2014)).

The difficulty level chosen modifies the complexity of the sentences and their structure, but the mechanics of the activity is the same. It also modifies the number of available options.

Figure 3. Comprehension activity. The player has to choose the right option to continue the conversation.

4.2. Production activities

These activities are focused on voice production, so players have to practice with their voice on such prosodic aspects as the tone of voice, the expression of emotions or the syllabic emphasis. Players engage in conversations with game characters, where they have to repeat different sentences related with the context of the conversation (as shown in Figure 4).

As in the comprehension activities, at the beginning of the activity, the virtual assistant introduces the activity, in order to relate it to the conversation in progress. Later, the game plays the sentence to be recorded. The player must record the sentence shown on the screen in order to continue the conversation with the character. To do this, the player must click on the microphone button and start recording. The trainer must decide if the player has repeated the sentence correctly or incorrectly. If the recording is correct, the game will show a positive feedback on the screen. If it is incorrect, the player can repeat the recording a certain number of times (2 or 3 times, depending on the activity). At the end, if the player cannot make the recording correctly, the game will show a negative feedback and the game continues.

These activities also have sound instructions during their development, in case of error and in case the player does not execute any action. As in the comprehension activities, the player cannot execute any action when a sound instruction is being played.

In these activities, the difficulty level also modifies the complexity of the sentence that the player has to record.

4.3. Prosodic activities

The objective of these activities is to mimic the prosodic curve of a sentence. To do this, players have to move an object (a bus) on the screen using their tone of voice, as shown in Figure 5. In addition, the text of the sentence is shown below the curve, highlighting the ascent or descent of the curve, with the goal that the players can connect the curve with the sentence and learn to modify their tone of voice depending on the type of sentence (interrogative or affirmative). To indicate when the player has to make a break, the game uses a traffic light. Moreover, players are asked to reproduce the sentence in the normal way and humming. The objective is that players understand the intonation changes in both ways.

Before players start the activity, an example is reproduced: the game plays a recorded voice of the sentence and an object moves automatically, following the tone of speech. In these activities, the trainer also has to decide if the recording is correct or incorrect.

In order to translate the speech into object motion, we use a pitch tracker based on the fast lifting wavelet transform (FLWT) to extract the fundamental frequency of the user's voice in real time (Larson & Maddox, 2005). Production activities previous to the prosodic activities are used to calibrate the pitch tracker and to adapt it to the current player.

4.4. Visual activities

The visual activities are more variable in their development. They are introduced to increase the playability of the video game.

- Hidden pieces: the player has to search, using a magnifying glass, for several pieces hidden between the pages of a book. These pieces are only seen when the player passes the magnifying glass over them.
- Puzzle: the player has to form a drawing of the moon using the pieces found in the previous activity. According to the difficulty level, the player will have more or less clues on how to build the shape. Also, a sound stimulus will be played when the player puts a piece incorrectly, in order to tell the player that an error has been made.
- Cookies: the objective of this activity is that the player understands the difference between an offer (question) and an affirmation (declarative), from the point of view of the character of the game. Two options (a declarative and a question) are presented and the player is asked to select the option that can be used to offer cookies to another person.

Figure 4. Production activity. The player has to say the sentence to continue the conversation.

Figure 5. Prosodic activity. The user has to reproduce the intonation contour and to make the breaks at the traffic lights.

- Dusting a bookshelf: the player has to clean the dust from a bookshelf with a cloth handled by the mouse.
- Hidden map: this activity is similar to the hidden pieces activity. This time, the player has to find a map hidden on the bookshelf.

5. Evaluation strategy

Figure 6 shows the strategy that has been followed for the design, implementation and evaluation of the video game. The goal of the strategy is to answer the three research questions (RQ1, RQ2, RQ3) presented in the introduction of the article. This fact determines the relative weight of the different usability aspects (represented by the size of the circles, according to (Quesenbery, 2003)) and the evaluation activities that have been performed. We used user centered design, in which the users are the main focal point and have a significant impact on the design decisions. In this approach, the use of both usability inspection methods and testing with real users is recommended, combining qualitative and quantitative methods (Lewis, 2014).

The development of the video game started with a script that specifies in detail the narrative of the graphic adventure

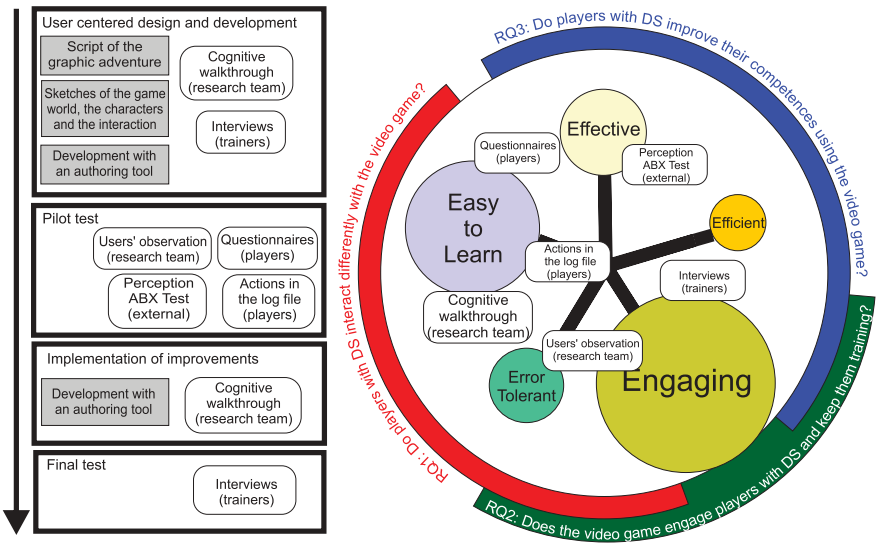

Figure 6. Evaluation strategy. On the left, temporal sequence of the design, implementation and evaluation activities. On the right the five dimensions that describe the different aspects of usability according to (Quesenbery, 2003) and their relation with the three research questions and the evaluation activities. The circles represent the usability aspects (the size reflects the importance of the aspect), the rectangles represent the development and evaluation phases, and the rounded rectangles are the evaluation activities with the participants in brackets.

and the activities. A set of sketches was also used to describe the different scenes, the characters and the interaction of the video game. Cognitive walkthroughs (Mahatody, Sagar, & Kolski, 2010; Nielsen, 1994; Polson, Lewis, Rieman, & Wharton, 1992) were applied by the experts in DS of the research team to ensure that the narrative and the activities were adapted to the way users with DS process information and to their cognitive and linguistic abilities. The goal was to make an attractive interface that engages the users. Another objective was to ensure an easy to learn interface that takes into account the special characteristics of the users and that is tolerant to the players' possible mistakes. As a result of these cognitive walkthroughs, a series of design principles, presented in the discussion section, were collected. The authoring tool (Adobe Flash) permitted us to adapt the interface to the experts' suggestions. In this way, the different prototypes were reviewed by the experts under the perspective of the users with DS and their special characteristics.

A pilot test was performed with real users. The goals were to check the degree of engagement of the players, detect the difficulties of use, observe the reaction of the users to the video game feedback and check whether the system is effective for practicing pronunciation. Users did not have any training before the pilot test. During the pilot test, one of the researchers took the role of the trainer. One of the staff members of the education center played the role of observer during the test. In this test, objective and subjective evaluation methods were combined. On the one hand, the game itself recorded data about the interaction between the player and the game. On the other hand, at the end of the game session, evaluators gave a questionnaire on general aspects of usability to the player; this was complemented by observations that evaluators collected during the test. In addition, at the end of the questionnaire, the opinions of speech therapists or teachers who are dedicated to special education were collected. As the video game records the voice of the players, we used these recordings in a perception test, to check whether the players' pronunciation improves as they use the video game (see Section 6.3).

The pilot test was performed by 14 participants with Down syndrome: 10 boys and 4 girls (chronological ages: 13, 16, 16, 17, 18, 20, 21, 22, 23, 24, 25, 25, 26 and 39). All of them have a moderate intellectual disability. All tests were done face to face at the place where the participants performed their educational activities for people with intellectual disabilities. In particular, these tests were conducted at the Niu School (Barcelona), Aura Foundation (Barcelona), and the School of Special Education "El Pino de Obregón" (Valladolid).

In addition, tests were performed on 10 children aged between 6 and 9 years and on 10 adults. These tests were done with the purpose of comparing the results obtained with the results of the players with DS. We can thus see if there are significant differences in the use of the video game depending on the different player profiles, in order to shed light on the research question RQ1.

The questionnaire was prepared to obtain information from the users relating to various aspects of the game, such as the story, the assistant or the difficulty. We collected information from children and users with DS. We ignored the information that adults could provide for being not relevant to drawing conclusions about the game. The questionnaire contains a series of yes/no questions and some with two possible answers. Moreover, in some questions, users were asked to explain why they gave that answer or to give more details of things that they would change in order to improve the game. Children filled in the questionnaire on their own. In the case of the users with DS, the questionnaire was used to guide a short interview in which the evaluation team, with the help of the teachers, retrieved the information while trying to avoid the use of directed questions. The questionnaire is included in Appendix A.

After the pilot test, an evaluation report was prepared which is summarized in Section 6. A number of small changes were done in the interface concerning the aspect and position of the recording button, the simplification of the prosodic activity, and some improvements on the audio instructions and help messages.

Finally, the School of Special Education "El Pino de Obregón" made more field tests with real users (both players and trainers). After two months' work (two hours per week in the language classes), they completed three sessions with each of the children with DS from the school that participated in the pilot test. We carried out a semi-structured interview with the teachers of the school in order to know their opinions of the video game and to get some feedback. We prepared a set of open-ended questions (see Appendix B) and allowed the teachers to respond in an open way. This interview triangulates the information obtained in the pilot test to shed light on the research questions RQ2 and RQ3, concerning the degree of engagement of the players, and the effectiveness of the use of the software for training pronunciation.

6. Results

In this section we present the results obtained in the different tests. The results are grouped according to the research question to which they relate.

6.1. Analysis of interaction difficulties of players with DS

The main source of information to address this topic is the log file, which records the actions performed by the users during the pilot test. The number of players that participated in the pilot test and the total playing time are shown in Table 1. There are significant differences in time between the three user groups ($p < 0.001$ Kruskal-Wallis test). The game does not offer a playable challenge to adults and children, so the game is carried out without interruption. Adults play the game faster than children ($p = 0.005$ Mann-Whitney test).

Table 1. Number of players and total time taken to finish the game in seconds (mean and standard deviation for each type of player).

		Time	
	Number of Users	Mean	SD
Children	10	942.2	130.3
Adults	10	824.8	72.7
Down syndrome	14	1342.1	301.4

Moreover, the game time for people with DS is significantly higher than for adults ($p < 0.001$ Mann-Whitney test) and children ($p < 0.001$ Mann-Whitney test). There is also a high standard deviation in people with DS, mainly due to their heterogeneity. They have specific characteristics which could produce varied response times. Therefore, people with DS have more difficulty solving the challenges offered by the game than adults and children.

Table 2 presents the data obtained in the comprehension activities. This table contains average data on resolution time in seconds, errors (wrong choice of sentence in conversation), clicks on the speaker button (to hear a phrase again), sound aids (those reproduced when the user takes more time than expected to perform some action), and clicks that are made on the interface when mouse clicks are disabled (e.g. when an audio is playing). The Kruskal-Wallis test showed differences between the three user groups in time ($p < 0.001$), errors ($p = 0.003$), and clicks ($p < 0.001$). Resolution times are higher for people with DS than for adults (166.0 vs. 105.9 with $p < 0.001$ Mann-Whitney test) and for children (166.0 vs. 118.8 with $p = 0.007$ Mann-Whitney test). Besides, people with DS made more errors than adults (1.0 vs. 0.1 with $p = 0.007$ Mann-Whitney test) and children (1.0 vs. 0.1 with $p = 0.007$ Mann-Whitney test). Finally, the number of incorrect clicks made by children is worth noting. It is higher than the number of incorrect clicks made by adults (1.6 vs. 0.0 with $p < 0.001$ Mann-Whitney test) and by people with DS (1.6 vs. 0.1 with $p < 0.001$ Mann-Whitney test). This is because the audio instructions are numerous and repetitive, something that is not harmful to people with DS, but which slows the game down and makes children get impatient because they want to move forward but they have to wait until the audio ends.

Table 3 shows the data obtained in production activities (the variables represented are the same as those shown in Table 2). There are significant differences in time ($p < 0.001$), number of errors ($p < 0.001$), phrase repetitions ($p < 0.001$), and clicks ($p < 0.027$) between the three user groups using the Kruskal-Wallis test. Again, observed times are higher for people with DS (335.9 vs. 173.1 compared with adults, with $p < 0.001$; 335.9 vs. 189.5 compared with children, with $p < 0.001$; Mann-Whitney test), but with greater differences with respect to the comprehension activities. The number of errors is much higher for people with DS, due to the difficulties in speech production of this population (4.4 vs. 0.1 compared with adults, with $p < 0.001$; 4.4 vs. 0.3 compared with children, with $p < 0.001$; Mann-Whitney test). Also significant is the number of clicks on the speaker button (phrase repetitions) that people with DS made (3.7 vs. 0.0 compared with adults, with $p < 0.001$; 3.7 vs. 0.6 compared with children, with $p = 0.012$; Mann-Whitney test). Some users with DS had problems with the reading and needed to hear the sentence again in order to do the recording better. In these activities, again, the children did a lot of incorrect clicks, more than other user types (6.2 vs. 1.0 compared with adults, with $p = 0.013$; 6.2 vs. 2.2 compared with people with DS, with $p = 0.026$; Mann-Whitney test).

Table 4 presents the data of prosodic activities. This activity was difficult for all users because the melodic curve proposed was too demanding to follow in real time. It has two parts, one with lexical content and another without it, based on a humming activity. The part based on the humming of the sentence causes confusion in all players, because it is hard to remember the hummed pattern once the recording starts.

The results of the visual activities are shown in Tables 5 and 6. First, the average time spent in the hidden pieces activity is shown. The Kruskal-Wallis test shows that there are significant differences between the three user groups ($p < 0.001$). Again, players with DS employ more time to complete the activity (100.9 vs. 27.4 compared with adults, with $p < 0.001$; 100.9 vs. 50.1 compared with children, with $p = 0.002$; Mann-Whitney test). This result is symptomatic of the deficits in executive control functions, and more specifically in the inhibitory control of response, which provides the ability to not answer a question or stop giving the answer if it is not adequate (Daunhauer et al., 2014).

Second, average time spent and average number of errors is shown for the puzzle activity. There are significant differ-

Table 2. Comprehension activities: time spent in seconds, number of errors, number of additional listenings of the phrase, number of speech aids and number of incorrect clicks (mean and standard deviation for each type of user in the 5 comprehension activities).

	Time		Errors		Phrase repetitions		Speech aids		Incorrect clicks	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Children	118.8	17.2	0.1	0.3	1.2	1.3	0.0	0.0	1.6	1.5
Adults	105.9	7.4	0.1	0.3	0.3	0.5	0.0	0.0	0.0	0.0
Down syndrome	166.0	43.3	1.0	0.9	1.6	1.7	0.6	1.0	0.1	0.5

Table 3. Production activities: time spent in seconds, number of errors, number of additional listenings of the phrase, number of speech aids and number of incorrect clicks (mean and standard deviation for each type of user in the 7 production activities).

	Time		Errors		Phrase repetitions		Speech aids		Incorrect clicks	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Children	189.5	31.3	0.3	0.7	0.6	1.1	0.1	0.3	6.2	6.0
Adults	173.1	6.5	0.1	0.3	0.0	0.0	0.0	0.0	1.0	0.9
Down syndrome	335.9	102.6	4.4	2.3	3.7	3.7	1.6	2.8	2.2	3.5

Table 4. Prosodic activities: time spent in seconds, number of phrase repetitions and number of humming repetitions (mean and standard deviation for each type of user).

	Time		Phrase repetitions		Humming repetitions	
	Mean	SD	Mean	SD	Mean	SD
Children	99.5	7.1	0.1	0.3	0.0	0.0
Adults	98.6	10.0	0.1	0.3	0.1	0.3
Down syndrome	106.5	10.4	0.4	0.7	0.4	0.5

Table 5. Visual activities: hidden pieces (time spent in seconds), puzzle (time spent in seconds and number of errors) and cookies (time spent in seconds and number of errors).

	Hidden pieces		Puzzle				Cookies			
	Time		Time		Errors		Time		Errors	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Children	50.1	15.3	27.9	5.7	0.2	0.6	27.6	6.7	0.4	0.5
Adults	27.4	8.6	21.5	2.3	0.0	0.0	22.1	5.0	0.0	0.0
Down syndrome	100.9	48.3	54.0	20.5	1.8	2.8	30.2	9.6	0.3	0.5

Table 6. Visual activities: dusting a bookshelf (time spent in seconds) and hidden map (time spent in seconds).

	Dusting a bookshelf		Hidden map	
	Time		Time	
	Mean	SD	Mean	SD
Children	21.2	10.9	16.2	7.1
Adults	15.9	9.1	11.2	3.3
Down syndrome	33.3	17.7	55.9	47.5

ences between the three user groups (time: $p < 0.001$; errors: $p = 0.010$; Kruskal-Wallis test). On the one hand, players with DS take longer to perform the activity (54.0 vs. 21.5 compared with adults, with $p < 0.001$; 54.0 vs. 27.9 compared with children, with $p < 0.001$; Mann-Whitney test). On the other hand, adult and child players committed very few errors in assembling the puzzle, but players with DS made many more mistakes until they assembled the puzzle correctly (1.8 vs. 0.0 compared with adults, with $p = 0.001$; 1.8 vs. 0.2 compared with children, with $p = 0.045$; Mann-Whitney test). People with DS have difficulties processing information, because, in this population, the two types of local and global processing seem to be dissociated (i.e. the ability to attach pieces of information to a whole, or the inverse, to segment comprehensive information into parts). In addition, these people often have a bias toward global information processing; they do not process and integrate details appropriately, especially if they are not connected to a whole, as in this case (D'Souza, Booth, Connolly, Happé, & Karmiloff-Smith, 2016).

Third, the average time spent and average number of errors is shown for the cookies activity. In this activity, there is not much difference between the times of all types of users ($p = 0.051$ Kruskal-Wallis test) or between errors ($p = 0.098$ Kruskal-Wallis test). Finally, the average time spent in the dusting a bookshelf and the hidden map activities is shown. There are significant differences in both activities between the three user groups ($p = 0.001$ for dusting a bookshelf; $p < 0.001$ for hidden map; Kruskal-Wallis test). As can be seen, players with DS take longer to perform both activities, especially searching the map (dusting a bookshelf: 33.3 vs. 15.9 compared with adults, with $p = 0.001$; 33.3 vs. 21.2 compared with children, with $p = 0.016$. Hidden map: 55.9 vs. 11.2 compared with adults, with $p < 0.001$; 55.9 vs. 16.2 compared with children, with $p < 0.001$. Mann-Whitney test).

Apart from the actions recorded in the log file, the second source of information for analyzing the differences between the types of players is the pilot test questionnaire, whose results are presented in Tables 7 and 8. The most relevant differences between the answers of the children and the players with DS are related to the understanding of the story (100% vs. 64.3%), the perceived ease of use (100% vs. 78.6%), and the perception that the video game is long (33.3% vs. 71.4%).

Users' observation performed by the research team during the pilot test confirms that, in spite of the fact that players with DS can easily interact with the system, the differences reported in the previous paragraphs are clearly observed during the playing sessions: players with DS are slower and make more mistakes.

6.2. Degree of engagement of the players

Users' observation carried out during the pilot test and the opinions collected in the teachers' interview performed after the field test allow us to describe the behavior of the users with DS during the gaming sessions as attentive, taking the video game seriously; they like the video game and feel motivated to play; and they focus their attention, even students with attentional problems. Teachers reported:

"It seemed curious to us how quiet they are, even the ones who are more active. Some of them are very focused".

"They feel the responsibility of doing it well, they are super serious and super formal".

"It caught our attention how they remain staring at the screen".

"Paula is a girl that has attention deficit. [During the game session] she has been very attentive, focused and giving the answer immediately. She has done everything right. But above all, the attitude of being calm, following everything correctly, super responsible".

"What I liked most is the responsibility of the children. The video game motivates them. For instance, Paula is a hyperactive girl and with the video game she is focused … with full attention, she was delighted, answering correctly".

"The video game is useful for motivating the students in their daily learning, for focusing their attention".

Tables 7 and 8 show the results of the questionnaire for each group of players interviewed after the pilot test. Players expressed the fact that they liked to test the game and would like to continue the story later. There was also fair unanimity in relation to the assistant of the game and the opportunity to

Table 7. Questionnaire results (yes/no questions).

	Children		Down syndrome	
	Yes	No	Yes	No
The user likes video games	100.0%	0.0%	92.9%	7.1%
The user is used to playing video games	80.0%	20.0%	78.6%	21.4%
The user enjoyed trying the game	90.0%	10.0%	92.9%	7.1%
The user would like to play again in the future /the user would like to know how the story continues	90.0%	10.0%	85.7%	14.3%
The user understands the story	100.0%	0.0%	64.3%	35.7%
The user likes the story	90.0%	10.0%	78.6%	21.4%
The user likes the parrot as the adventure's companion	90.0%	10.0%	85.7%	14.3%
The user likes talking to the characters of the video game	88.9%	11.1%	92.9%	7.1%
The user would like to create his/her own avatar (face, eyes, mouth, clothes, etc)	88.9%	11.1%	64.3%	35.7%
The user manifests to have learned something with the video game	70.0%	30.0%	78.6%	21.4%
The user likes the city images and the people	100.0%	0.0%	85.7%	14.3%

Table 8. Questionnaire results (two options questions).

			Children		Down syndrome	
	Option1	Option2	Option1	Option2	Option1	Option2
The video game seemed …	Funny	Boring	88.9%	11.1%	85.7%	14.3%
The video game seemed …	Easy	Hard	100.0%	0.0%	78.6%	21.4%
The video game seemed …	Short	Long	66.7%	33.3%	28.6%	71.4%

talk to the characters of the game. Finally, the vast majority were satisfied with the character graphics and game scenes. Teachers reported that their students insisted on replaying "The magic stone" even weeks after the playing sessions. There was also much unanimity with respect to the entertainment and ease aspects of the game. Most of the players believed that the game is funny and easy (first and second row of Table 8). The fun aspect is important to improve motivation. The opinions about the ease of the game suggest that the difficulty level is adequate. The overall assessment of the game is thus positive, as described by the teachers:

"When they were finishing, we asked them for an assessment: 'what do you think?, do you like it?, don't you like it?'. All of them said that they liked it a lot. 'Is there anything that you don't like?, what would you remove?' They said no, that they liked everything".

Players with DS believe that the game is long, while the children think that it is short (third row of Table 8). We already mentioned in the previous subsection that players with DS took more time to finish according to the actions recorded in the log file (see Table 1). Although they had to work longer, all of them kept working until the end of the game session, which is also a consequence of the high degree of engagement of the players with DS. The engagement is also reflected by the fact that there is a high level of immersion in the game, as commented by the teachers:

"Some children are so imaginative that they get so much into the role that they are incorporated inside the video game, they get into the video game. They are so imaginative. Some of them have come to dream of the video game. A boy has come to say: 'I am a hero'".

6.3. Analysis of the improvement in pronunciation

In order to assess the efficiency of the activities of production, we analyzed the quality of the pronunciation of the players with DS in the sentence "Hola tío Pau, ¿sabes dónde vive la señora Luna?" (Hello uncle Pau, do you know where Mrs. Moon lives?). This sentence has been selected because users are required to repeat it several times during the video game so that we can analyze whether the last production improves with respect to the first one.

Seven evaluators judged in an ABX perception test (Pisoni & Lazarus, 1974) which one of the utterances in a pair seems more natural or appropriate. We selected people without any specific background on speech therapies in this test, as we were interested in the perception of normal people concerning the possible improvement of the players. For that reason, we decided not to make any more specific question about intonation and prosody because we didn't want to condition the opinion of the evaluators. This is not an intelligibility test (the spoken sentence was always the same) but a perception test of the overall differences between a pair of utterances of the same sentence.

The pair is composed of the first and last production of the sentence during the working session. The first of them was uttered after reading and listening to it. The second one was produced after receiving a visual stimulus. The list of pairs (one per player) of utterances are presented in a web interface. Utterance A and utterance B are assigned randomly to the first and last production of the sentence. The evaluators can listen to utterance A and utterance B as many times as they consider necessary.

Two questions are presented to the testers:

- Q1: Do you perceive any difference between both utterances?
- Q2: Do you have any preference?

The possible answers to Q1 are in a 5-point Likert scale: 1 means "I do not perceive any difference" and 5 means "I perceive the differences very clearly". We obtained 98 answers (14 subjects and 7 raters) with a percentage of consistency of 71.4% (tolerance = 3). Table 9 shows the statistics per speaker. Low values evidence small differences in all the cases. All the mean values are over 3, which indicates that the evaluators perceive differences between the utterances.

The possible answers to Q2 are: "I prefer utterance A", "I prefer utterance B", "No preference: both are good", "No preference: none is good", "No preference: I do not perceive any difference". We obtained 98 answers (14 subjects and 7 raters) with an inter-transcriber consistency Kappa Fleiss index of 0.224 which is "Fair agreement" in the Landis and Koch (1977) scale. Most of the judgments consider that the last utterance is the best one (64.2% with p-value = 0.014 binomial test). Additionally, as can be seen in Table 9, there is a clear dependence on the player. None of the evaluations observed any improvement when evaluating the players spk2 and spk10. After checking the control forms of the test, we observed that these two players were not able to pay attention to the game, which justifies the bad results. Indeed, the percentage of preference of the last utterance is 74.3%, after removing these two speakers. The speakers that improve are 8 out of 14, as can be seen in Table 9.

Apart from this perception test, we have also obtained the opinions of the teachers from the teachers' interview performed after the field test. Teachers think that the video game helps the students to improve their pronunciation. Even in the same session, they noticed pronunciation improvements. Players also improve along the different sessions. Mainly, they noticed improvements in the use of pauses and in the control of intonation of interrogative sentences. Teachers reported:

"They have improved. Within the same session it may be that they improve, but over the days they do clearly better".

Table 9. ABX test.

	Q1		Q2				
Speaker	Mean	SD	Last better	Last worse	Both are good	None is good	No difference
spk1	4.4	0.8	6	0	0	1	0
spk2	4.3	1.1	0	5	0	2	0
spk3	4.3	1.1	7	0	0	0	0
spk4	3.6	1.3	3	4	0	0	0
spk5	3.7	1.0	3	1	3	0	0
spk6	3.7	1.1	5	1	0	0	1
spk7	4.3	1.0	4	3	0	0	0
spk8	3.0	1.3	4	2	1	0	0
spk9	3.0	1.2	3	2	2	0	0
spk10	3.1	1.3	0	6	1	0	0
spk11	4.4	0.8	7	0	0	0	0
spk12	3.1	1.2	1	3	3	0	0
spk13	3.3	1.4	4	2	0	0	1
spk14	3.0	1.2	5	0	0	0	2

"Paula is a hyperactive girl … She was making very short breaks and she has corrected it a lot. She reads well, but she read super fast and with low intonation, and with the repetition she has improved a lot. Of the girls you see that change, amazing".

"The video game allows us to work on interrogative sentences and on pauses".

When teachers were asked about the reasons for the engagement and the improvement of the players, they assessed that the context and the story of the video game are appropriate, which is essential for developing the activities with proficiency. Although players miss some details about the story, the teachers assessed that the players understand the main objective of the video game (searching for a magic stone) and use of a significative context is very helpful for interpreting the pragmatics of prosody. They commented:

"They do comprehension activities quite well. We are surprised that they have been working so positively. They understand the context, even when they have to choose between three possible answers. It is not difficult for them, just the opposite".

"They do understand the story, they know that they have to search for a magic stone".

Teachers recommend the use of the video game, because of the following reasons. First, because it has interesting activities, which they think are fun. Second, because students can play independently. Finally, because it allows them to work with the students in a different way. In regular lessons, teachers are used to being in front of the students explaining the exercises. However, using the video game, they do not need to explain anything, just sit next to the players, help them if required and evaluate the activities. Comments of the teachers:

"I would recommend the video game to a colleague, to a teacher of children with Down syndrome. I would recommend the video game because it provides a dynamic of repeating in a playful way, functional and quite independent. I'm beside, not in front of them".

"I would use the video game for repetitive learning, for pauses, and for making interrogative sentences in a mechanical way. For training it is great".

"I would use it in the classroom".

Teachers think that the instructions and the feedback of the video game are suitable. Besides, they believe that the virtual assistant is useful. Teachers also like the graphics of the video game. They reported:

"The feedback provided by the video game, the applauses, is wonderful".

7. Discussion

The experimental results show that performing repetition exercises using the video game can lead to improvements in the production and use of prosody by speakers with DS. Despite the fact that players with DS use the video game with greater difficulty than other users (RQ1), they have a high motivation to use it (RQ2), which makes them strive to do the activities, with the consequent improvement (RQ3). We think that this is an excellent result because it encourages a repetition of the strategy in other scenarios to achieve engagement of users with DS in serious games. We give the credit for this to the application of a number of HCI design elements that have taken into account the special characteristics of the users such as: deficits in language development, problems with short term memory, attention deficit, lack of motivation, bad response to frustration, problems to process information coming from different channels, and difficulties in understanding the meaning of iconic symbols. We focus on these characteristics in the following subsections, referring to the HCI design elements that could have influenced the achievement of such results.

7.1. Motivating the users with a graphic adventure

People with DS usually have low motivation for difficult tasks like educational ones, because of their cognitive and learning problems. The inclusion of learning activities that must be accomplished in order to progress in the game has been one of the keys to motivating the users. The initial story (an animated cartoon of about 3 minutes) engages the players because they feel represented as the main character of the action of the video game. For the users, doing the activities in a proper way became a significant goal. As a result, almost all the players keep doing the activities until the end of the game without showing signs of fatigue or boredom.

The scenes of the video game represent real life situations (for practicing prosody in context), but the narrative is based on a fantastic story of a magic stone to be rescued. One of the teachers argued this fact as an important engaging factor, as many of the children and adolescents with DS have a remarkably imaginative character.

7.2. Attention deficit and multimodal interface

The teachers were surprised by the result of the game with students with deficit of attention. Even the most problematic students in terms of attention carried out the activities until the end. We attribute this level of engagement to the high motivation mentioned in the previous subsection, and to the combination of attractive graphics and sound stimulation.

As for the aesthetics, bright colors are used, but we avoid childish cartoons. One of the challenges of developing the video game is to find an adequate content and design presentation. People with DS are more competent in social and emotional skills than in cognitive capabilities. As they grow older, their preferences and tastes become more mature, but the cognitive capabilities lag behind (Feng et al., 2010). On the one hand, computer programs that appeal to young people are usually too advanced for adolescents with DS. On the other hand, programs to learn skills at the right level tend to be boring for them, because they do not match their age and interests.

The sound stimulation consists of the voice of the voiceover on the one hand, and on the turns of the assistant (the parrot) on the other. Both voices use a friendly tone to help and guide the user during the adventure. The attitude of the voiceover and the assistant is patient, repeating the messages as many times as necessary, with the aim that the user will focus on the activity.

7.3. Multimodal feedback and intolerance to frustration

Performing tasks that can cause experiences of failure may demotivate people with DS, because of their low tolerance to frustration. Therefore, the feedback offered to the user by the game is always positive, both when the user performs the activity correctly and incorrectly. When the player fails, a sad face appears accompanied by a message that encourages the player to try again. When the player performs the activity successfully, the sound of applause and the corresponding icon rewards the player. We have observed that, in spite of the fact that this positive message is always the same, the users respond positively to it, mostly with a smile. Moreover, after a number of incorrect attempts, the video game continues with the following activity.

We avoided the use of rankings, scores and other similar rewarding strategies, because we didn't want to stress the users. Instead, the role of the teacher was shown to be very relevant to assisting the student in improving and getting positive feedback. The possibility of giving automatic feedback was also abandoned because the risk of having an incorrect feedback in real time is high, leading to potential undesirable situations. We advise against the use of the game in autonomous mode (without the assistance of a teacher or a parent) because we predict that some players would focus on progressing in the graphic adventure without putting the required effort into doing the activities (each activity has a maximum number of tries after which the game continues).

7.4. Audio messages to help the abstraction of concepts

The icons and scenes of the video game are visually intuitive and simple, with the goal of avoiding ambiguities. Furthermore, the multimodal output helps the players as the voiceover pronounces the name of the represented objects when they appear (magnifying glass, book, map, etc.) and also informs about the current scene (library, bus stop, etc.).

All the activities provide oral instructions to the user. At the beginning of the activity, the voiceover explains what has to be done in the activity. If the user gets stuck on the activity, more instructions are provided. These instructions seem to work properly, guiding the user through the different activities. In addition, the sound auxiliary aids, which are reproduced when the player does nothing for a certain time, are an essential resource, as most players with DS need more time to analyze the activity and do it.

In some cases, we could not avoid confusions, as we observed that some players started talking before pushing the microphone button, or pushed the speaker button to record. This highlights the difficulties of users with DS to understand clearly the purpose of each interface element.

7.5. Deficits on language development

The learning goals of the video game are related with the training of a particular aspect of language such as prosody. It has been shown that the repetition activities, in combination with the help of the teachers, result in the improvement of the oral production of the players. Additionally, experimental results show that users with DS make more mistakes and need more time to finish these activities than adults and children. Therefore, comprehension activities introduce some level of difficulty for users with DS, which helps to improve both lexical-semantic comprehension and prosodic perception in specific contexts.

The use of short and simple sentences that appear in the messages of the program (both written and oral messages) has been shown to be useful in minimizing the impact of these language development deficits. The implementation of three levels of difficulty permits the teacher to adapt the game to the specific characteristics of each user.

7.6. Difficulties for processing information and poor short term memory

Visual activities are introduced to increase the playability of the game. The results obtained showed the difficulties of people with DS to do these activities, because of their cognitive difficulties. For instance, in the hidden pieces activity, the pieces that the player has to look for are too small and some players have trouble finding them. In the activity of dusting a bookshelf, some users leave areas uncleaned, because those areas are too small to be seen with the naked eye. In the hidden map activity, some players have trouble locating the map for the same reasons as in the hidden pieces activity.

Additionally, results show significant differences with respect to the time taken to do the activities by the different user groups. Users with DS are slower than the rest of the users, and again, audio stimulus has been shown to be useful to remind them of the activity's goal when the players take more time than expected to perform the task.

7.7. Problems to integrate information coming from different sources

This article presents a good example in which people with DS benefit from the use of multimodal information coming from two different channels, such as audio and image. J. C. Martin (1998) classifies the possible channel combination modes into transfer, equivalence, specialization, complementarity and redundancy. Redundancy has been shown to be useful to disambiguate the icons and scenes of the video game and to offer feedback to the user. The audio channel is specialized in giving messages that reinforce the argument of the graphic adventure and remind them of the goal of the activities, given the poor short term memory of the users.

Problems for integrating information from different channels have been observed mainly in the prosodic activities. In these activities, players have to move an object (a bus) on the screen using their tone of voice. The objective is to mimic the prosodic curve of the sentence. Players with DS find this activity difficult to understand. Users able to read generally focus on the text, ignoring the bus animation and the prosodic contours. Even for non reader users, teachers manifest serious doubts about whether the players understand that the bus movement is a result of their voice. The difficulty may also be connected with the fact that the elements of these activities are too abstract for users with DS.

7.8. Limitations

We have included three levels of cognitive and linguistic difficulty in the video game (see section 3.3), but this categorization may not be accurate enough to adjust the game dynamics to the specific profiles of all potential users. It should be noted that there is great heterogeneity in the profiles of cognitive and linguistic development among people with Down syndrome (Abbeduto, McDuffie, Thurman, & Kover, 2016; Dierssen, Ortiz-Abalia, Arqué, Martínez De Lagrán, & Fillat, 2006). During the testing, the teacher or speech therapist helped players with the obstacles they found. We think that adapting the video game to the whole possible spectrum of people with Down syndrome is a difficult challenge and may not even be possible. We assume that even though we have made a remarkable effort in adapting the interface, the role of the assistant teacher or speech therapist remains key.

The video game is a tool of educational support in the areas of prosody and pragmatics, while at the same time it serves to build a corpus of knowledge about prosodic and pragmatic limitations of the individuals affected by Down syndrome. The perceptual test performed has demonstrated that playing with the video game improves the quality of the vocal production of the users (see section 6.3). In this research, we have focused on the definition of the interface that the players use to perform the activities, but the effectiveness of the training was not measured in terms of each particular prosodic ability, but in terms of the overall perceived quality. In further steps of the project, specific tests of prosodic abilities will be conducted to obtain qualitative and quantitative measures of the learning of prosodic skills. The results of these tests will serve to refine the activities, but the structure and the elements of the interface will remain, since it has been shown that they serve to engage and motivate adolescents affected by DS.

8. Conclusions and future work

This article has described the design and implementation of a video game whose main objective is to improve communication skills, especially those related to prosody, of individuals with intellectual disabilities, in particular, those affected by Down syndrome. The use of a video game proved to be a promising research line, since it improves the motivation and engagement in the learning process. There exist tools in the field of speech therapy and audition that serve the professionals as assistance systems, but the innovative aspect of our work is that it incorporates activities of production and perception of prosodic phenomena through the combined strategies of reading aloud, imitative speech and resolution of speech acts. Players have to do activities that involve comprehension, production and prosodic tasks, together with visual activities. Nevertheless, due to the difficulties that people with DS encounter in human–computer interaction, we have paid special attention to the user interface, since it determines the interaction of the players with the video game. This task has turned out to be difficult, since there are not too many video games oriented to this type of users.

In this sense, the following eight key elements have been shown to be efficient in engaging and motivating adolescents with Down syndrome in the activities of the video game. The authors of this article encourage game designers and designers of interactive applications to use them as guidelines for the development of tools for people with cognitive impairments because they are founded in theoretical and descriptive studies about limitations of individuals with intellectual disabilities:

(1) Adapt the video game to the users' needs with a system of levels and reinforcements according to their individuality. The use of several levels of difficulty allows the complexity of the activities to be adjusted. Positive reinforcements motivate users when they succeed in the activities. In order to avoid frustration, allow the players to continue the game even if they do not succeed in the activities.
(2) Adapt the design of the scenes, items and characters to the specific difficulties of people with DS. Avoid using an excessive number of graphic elements that could distract attention from the relevant elements of the activity.
(3) Use short and simple sentences and high frequency words. The instructions and sentences used in the video game must be in accordance with the cognitive level of people with DS.
(4) Use audio messages to give instructions to the players and to assist them when their response time is high. The information provided by the audio channel helps users when it is redundant with the visual information.
(5) Include visual clues in the scenes to help players continue the game and prevent abandonment. Highlight the visual elements needed to advance in the interactive story.
(6) Consider the inclusion of a virtual assistant to guide the player during the game. The presence of this assistant ensures that the player knows at every moment what to do.
(7) Use realistic scenes to favor game immersion and to provide a context for learning. Real life situations allow individuals affected by DS to better interact with other people.
(8) Use fantastic elements to maintain the play-oriented nature of the proposed activities. These elements reinforce the players' interest in the video game.

The most relevant aspect of the results, due to the special characteristics of the people with DS, is that the majority of users kept playing throughout the video game with hardly any distractions. The players understand the main points of the narrative and are motivated by the fact that they have to solve the problem faced by the residents of a city. In general, they interact properly with the video game and succeed in doing the different activities (comprehension, production, prosodic and visual) and react well to positive reinforcement, even when they do not complete the activity perfectly. The users like to use the microphone to communicate with the

characters of the game, and the different levels of difficulty in the video game allow them to progress through the various activities without getting blocked.

Another aim of the video game is that it can be used by teachers or speech therapists in special education centers as part of their educational activities. Related to this, the evaluation carried out shows a high level of acceptance and that the final product can be really useful for different areas of disability education.

Overall, we have succeeded in developing a tool that motivates players with Down syndrome because it imposes challenges and, as a consequence, they are attentive and improve their speech in a pleasant environment. Thus, we conclude that the use of a video game to improve the oral and communication skills of people with Down syndrome is a promising research line.

Further research includes the adaptation of the video game to other collectives with intellectual disabilities that also present communicative limitations, especially in prosodic and pragmatic abilities, such as people with Williams syndrome (Martínez-Castilla, Sotillo, & Campos, 2011; Martínez-Castilla, Stojanovik, Setter, & Sotillo, 2012), fragile-X syndrome (Ferrier, Bashir, Meryash, Johnston, & Wolff, 1991) or other neurodevelopmental disorders such as autism spectrum disorders (Filipe, Frota, Castro, & Vicente, 2014; Peppé, Cleland, Gibbon, O'Hare, & Martínez-Castilla, 2011). Testing the video game with users of new profiles will allow us to determine whether the HCI elements that have succeeded with adolescents with DS are also effective in other populations with intellectual disabilities.

The current version of the video game focuses on prosodic and pragmatic limitations of adolescents with Down syndrome. It is future work to redesign the activities for the users to train other linguistic levels, such as phonetics (Rochet-Capellan & Dohen, 2015), syntax (Fortunato-Tavares et al., 2015) or vocabulary.

Currently the game is only available in Spanish, but we intend to extend the video game to users of other languages (among others, Catalan, English and French). To do this, besides the obvious aspects of translation, other modifications may also be required to adapt the video game to the specific cultural context.

Acknowledgments

This work was supported 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 soporte educativo" - PZ611683-2013ACUP00202), by Fundacion BBVA (project "Pradia: la aventura gráfica de la pragmática y la prosodia" - CF613399), by Ministerio de Economía y Competitividad y Fondos FEDER (project "Videojuegos sociales para la asistencia y mejora de la pronunciación de la lengua española" - TIN2014-59852-R) and by Junta de Castilla y León (project "Evaluación automática de la pronunciación del español como lengua extranjera para hablantes japoneses"- VA145U14). The authors would like to thank all the participants who took part in the evaluation. We also would like to thank Yurena Gutiérrez, Patricia Sinobas, Valentín Cardeñoso, Ferran Adell and Juan María Garrido for their collaboration in the project.

References

Abbeduto, L., McDuffie, A., Thurman, A. J., & Kover, S. T. (2016). Language development in individuals with intellectual and developmental disabilities: From phenotypes to treatments. In R. M. Hodapp & D. J. Fidler (Eds.), *Fifty years of research in intellectual and developmental disabilities* (vol. 50, pp. 71–118). Cambridge, MA: Academic Press.

Adams, E. (2013). *Fundamentals of game design*. Berkeley, CA: Pearson Education.

Aguilar, L., Bonafonte, A., Campillo, F., & Escudero, D. (2009). Determining intonational boundaries from the acoustic signal. In M. Uther, R. Moore, & S. Cox (Eds.), *10th Annual Conference of the International Speech Communication Association (interspeech)* (pp. 2447–2450). Brighton, UK: ISCA.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). *(DSM-5)* Arlington, VA: American Psychiatric Publishing.

Black, B. (2006). Educational software for children with Down syndrome - an update. *Down Syndrome News and Update*, *6* (2), 66–68.

Boyle, E., Connolly, T. M., & Hainey, T. (2011). The role of psychology in understanding the impact of computer games. *Entertainment Computing*, *2* (2), 69–74.

Brandão, A., Brandão, L., Nascimento, G., Moreira, B., Vasconcelos, C. N., & Clua, E. (2010). Jecripe: Stimulating cognitive abilities of children with Down syndrome in pre-scholar age using a game approach. In V. Shen, H. Duh, M. Inami, M. Haller, & Y. Kitamura (Eds.), *7th International Conference on Advances in Computer Entertainment Technology* (pp. 15–18). Taipei, Taiwan: ACM.

Brown, D. J., Ley, J., Evett, L., & Standen, P. (2011). Can participating in games based learning improve mathematic skills in students with intellectual disabilities?. In N. Dias, M. Cunha, & R. Simões (Eds.), *IEEE 1st International Conference on Serious Games and Applications for Health (SeGAH)* (pp. 1–9). Braga, Portugal: IEEE.

Bruno, A., González, C., Moreno, L., Noda, M., Aguilar, R., & Muñoz, V. (2003). Teaching mathematics to children with Down's syndrome. In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *11th International Conference on Artificial Intelligence in Education* (pp. 1–8). Sydney, Australia: AIED.

Buckley, S., & Bird, G. (2000). *Education for individuals with Down syndrome - an overview*. Portsmouth, UK: Down Syndrome Educational Trust.

Buckley, S., & Bird, G. (2001). *Memory development for individuals with Down syndrome – an overview*. Portsmouth, UK: Down Syndrome Educational Trust.

Buckley, S., & Sacks, B. (2002). *An overview of the development of teenagers with Down syndrome (11–16 years)*. Portsmouth, UK: Down Syndrome Educational Trust.

Cagatay, M., Ege, P., Tokdemir, G., & Cagiltay, N. E. (2012). A serious game for speech disorder children therapy. In Y. Aydin-Son & T. Can (Eds.), *7th international symposium on health informatics and bioinformatics (HIBIT)* (pp. 18–23). Nevsehir, Turkey: IEEE.

Carmien, S., Dawe, M., Fischer, G., Gorman, A., Kintsch, A., & Sullivan, J. F. (2005). Socio-technical environments supporting people with cognitive disabilities using public transportation. *ACM Transactions on Computer–Human Interaction*, *12* (2), 233–262.

Chapman, R. S. (1997). Language development in children and adolescents with Down syndrome. *Mental Retardation and Developmental Disabilities Research Reviews*, *3* (4), 307–312.

Chapman, R. S., & Hesketh, L. (2001). Language, cognition, and short-term memory in individuals with Down syndrome. *Down Syndrome Research and Practice*, *7* (1), 1–7.

Chapman, R. S., Schwartz, S. E., & Bird, E. K. R. (1991). Language skills of children and adolescents with Down syndrome I. comprehension. *Journal of Speech, Language, and Hearing Research*, *34* (5), 1106–1120.

D'Souza, D., Booth, R., Connolly, M., Happé, F., & Karmiloff-Smith, A. (2016). Rethinking the concepts of 'local or global processors': Evidence from Williams syndrome, Down syndrome, and autism spectrum disorders. *Developmental Science*, *19* (3), 452–468.

2.1 Engaging Adolescents with Down Syndrome in an Educational Video Game

Darejeh, A., & Salim, S. S. (2016). Gamification solutions to enhance software user engagement - a systematic review. *International Journal of Human–Computer Interaction*, *32* (8), 613–642.

Daunhauer, L., Fidler, D., Hahn, L., Will, E., Lee, N., & Hepburn, S. (2014). Profiles of everyday executive functioning in young children with Down syndrome. *American Journal on Intellectual and Developmental Disabilities*, *119* (4), 303–318.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification. using game-design elements in non-gaming contexts. In D. Tan, S. Amershi, B. Begole, W. Kellogg, & M. Tungare (Eds.), *CHI '11 extended abstracts on human factors in computing systems* (pp. 2425–2428). Vancouver, Canada: ACM.

Dierssen, M., Ortiz-Abalia, J., Arqué, G., Martínez De Lagrán, M., & Fillat, C. (2006). Pitfalls and hopes in Down syndrome therapeutic approaches: In the search for evidence-based treatments. *Behavior Genetics*, *36* (3), 454–468.

Escudero, D., Cardeñoso, V., & Bonafonte, A. (2002). Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish. In F. Taylor, J. Principe, & H. Bourlard (Eds.), *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (vol. 1, pp. I481–I484). Orlando, FL: IEEE.

Escudero-Mancebo, D., González-Ferreras, C., & Cardeñoso-Payo, V. (2002). Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis in Spanish. In J. Hansen & B. Pellom (Eds.), *Seventh International Conference on Spoken Language Processing (ICSLP)* (pp. 1165–1168). Denver, CO: ISCA.

Feng, J., Lazar, J., Kumin, L., & Ozok, A. (2008). Computer usage by young individuals with Down syndrome: An exploratory study. In S. Harper & A. Barreto (Eds.), *10th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 35–42). Halifax, Canada: ACM.

Feng, J., Lazar, J., Kumin, L., & Ozok, A. (2010). Computer usage by children with Down syndrome: Challenges and future research. *ACM Transactions on Accessible Computing*, *2* (3), 1–44.

Ferrier, L. J., Bashir, A. S., Meryash, D. L., Johnston, J., & Wolff, P. (1991). Conversational skills of individuals with fragile-X syndrome: A comparison with autism and Down syndrome. *Developmental Medicine & Child Neurology*, *33* (9), 776–788.

Filipe, M. G., Frota, S., Castro, S. L., & Vicente, S. G. (2014). Atypical prosody in Asperger syndrome: Perceptual and acoustic measurements. *Journal of Autism and Developmental Disorders*, *44* (8), 1972–1981.

Flores Lucas, V., & Belinchón Carmona, M. (2010). Dificultades en la comprensión de la ironía en personal con TEA, e implicaciones para la hipótesis de un déficit de habilidades de Teoría de la mente. In M. Belinchón Carmona (Ed.), *Investigaciones sobre autismo en español: Problemas y perspectivas* (pp. 139–154). Madrid, Spain: Centro de Psicología Aplicada.

Fortunato-Tavares, T., Andrade, C. R. F., Befi-Lopes, D., Limongi, S. O., Fernandes, F. D. M., & Schwartz, R. G. (2015). Syntactic comprehension and working memory in children with specific language impairment, autism or Down syndrome. *Clinical Linguistics & Phonetics*, *29* (7), 499–522.

Gallaher, K., van Kraayenoord, C., Jobling, A., & Moni, K. (2002). Reading with Abby: A case study of individual tutoring with a young adult with Down syndrome. *Down Syndrome Research and Practice*, *8* (2), 59–66.

González, C., Noda, A., Bruno, A., Moreno, L., & Muñoz, V. (2015). Learning subtraction and addition through digital boards: A Down syndrome case. *Universal Access in the Information Society*, *14* (1), 29–44.

González-Ferreras, C., Escudero-Mancebo, D., Vivaracho-Pascual, C., & Cardeñoso-Payo, V. (2012). Improving automatic classification of prosodic events by pairwise coupling. *IEEE Transactions on Audio, Speech, and Language Processing*, *20* (7), 2045–2058.

Hernández Martínez, M., Pastor Duran, X., & Navarro Navarro, J. (2011). Attention deficit disorder with or without hyperactivity or impulsivity in children with Down's syndrome. *International Medical Review on Down Syndrome*, *15* (2), 18–22.

Jarrold, C., & Baddeley, A. (2001). Short-term memory in Down syndrome: Applying the working memory model. *Down Syndrome Research and Practice*, *7* (1), 17–23.

Kapp, K. M. (2012). *The gamification of learning and instruction: Game-based methods and strategies for training and education*. San Francisco, CA: Pfeiffer.

Kent, R. D., & Vorperian, H. K. (2013). Speech impairment in Down syndrome: A review. *Journal of Speech, Language, and Hearing Research*, *56* (1), 178–210.

Khan, T. M. (2010). The effects of multimedia learning on children with different special education needs. *Procedia – Social and Behavioral Sciences*, *2* (2), 4341–4345.

Kumin, L., Lazar, J., Feng, J. H., Wentz, B., & Ekedebe, N. (2012). A usability evaluation of workplace-related tasks on a multi-touch tablet computer by adults with Down syndrome. *Journal of Usability Studies*, *7* (4), 118–142.

Landis, J., & Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33* (1), 159–174.

Lanfranchi, S., Cornoldi, C., Vianello, R., & Conners, F. (2004). Verbal and visuospatial working memory deficits in children with Down syndrome. *American Journal on Mental Retardation*, *109* (6), 456–466.

Larson, E., & Maddox, R. (2005). Real-time time-domain pitch tracking using wavelets. In S. Errede (Ed.), *Proceedings of the University of Illinois at Urbana Champaign Research Experience for Undergraduates Program* (pp. 1–12). Urbana-Champaign, IL: University of Illinois.

Lazar, J., Kumin, L., & Feng, J. H. (2011). Understanding the computer skills of adult expert users with Down syndrome: An exploratory study. In K. McCoy & Y. Yesilada (Eds.), *13th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 51–58). Dundee, UK: ACM.

Lewis, J. R. (2014). Usability: lessons learned … and yet to be learned. *International Journal of Human–Computer Interaction*, *30* (9), 663–684.

Lopez-Basterretxea, A., Mendez-Zorrilla, A., Garcia-Zapirain, B., Madariaga-Ortuzar, A., & Lazcano-Quintana, I. (2014). Serious games to promote independent living for intellectually disabled people: Starting with shopping. In Q. Mehdi, A. Elmaghraby, I. Marshall, A. Lauf, J. Jaromczk, R. Ragade…R. Yampolskiy (Eds.), *Computer games: AI, animation, mobile, multimedia, educational and serious games (CGAMES)* (pp. 1–4). Louisville, KY: IEEE.

Mahatody, T., Sagar, M., & Kolski, C. (2010). State of the art on the cognitive walkthrough method, its variants and evolutions. *International Journal of Human–Computer Interaction*, *26* (8), 741–785.

Martin, G. E., Klusek, J., Estigarribia, B., & Roberts, J. E. (2009). Language characteristics of individuals with Down syndrome. *Topics in Language Disorders*, *29* (2), 112–132.

Martin, J. C. (1998). TYCOON: Theoretical framework and software tools for multimodal interfaces. In J. Lee (Ed.), *Intelligence and multimodality in multimedia interfaces* (pp. 1–25). Menlo Park, CA: AAAI Press.

Martínez Celdrán, E., & Fernández Planas, A. M. (2007). *Manual de fonética española: Articulaciones y sonidos de español*. Barcelona, Spain: Editorial Ariel.

Martínez-Castilla, P., Sotillo, M., & Campos, R. (2011). Prosodic abilities of Spanish-speaking adolescents and adults with Williams syndrome. *Language and Cognitive Processes*, *26* (8), 1055–1082.

Martínez-Castilla, P., Stojanovik, V., Setter, J., & Sotillo, M. (2012). Prosodic abilities in Spanish and English children with Williams syndrome: A cross-linguistic study. *Applied Psycholinguistics*, *33* (1), 1–22.

Mayer, R. E. (2002). Multimedia learning. *Psychology of Learning and Motivation*, *41*, 85–139.

McFarlane, A., Sparrowhawk, A., & Heald, Y. (2002). *Report on the educational use of games*. Cambridge, UK: TEEM (Teachers evaluating educational multimedia).

Ng, K. H., Bakri, A., & Rahman, A. A. (2014). A review on courseware for Down syndrome children. *Journal of Information Systems Research and Innovation*, *8*, 56–65.

Nielsen, J. (1994). Usability inspection methods. In C. Plaisant (Ed.), *Conference companion on human factors in computing systems* (pp. 413–414). Boston, MA: ACM.

Ortega-Tudela, J. M., & Gómez-Ariza, C. J. (2006). Computer-assisted teaching and mathematical learning in Down syndrome children. *Journal of Computer Assisted Learning*, *22* (4), 298–307.

Patel, R., Kember, H., & Natale, S. (2014). Feasibility of augmenting text with visual prosodic cues to enhance oral reading. *Speech Communication*, *65*, 109–118.

Pazos González, M., Raposo-Rivas, M., & Martínez-Figueira, M. E. (2015). Las TIC en la educación de las personas con Síndrome de Down: Un estudio bibliométrico. *Virtualidad, Educación Y Ciencia*, *6* (11), 20–39.

Peppé, S., Cleland, J., Gibbon, F., O'Hare, A., & Martínez-Castilla, P. (2011). Expressive prosody in children with autism spectrum conditions. *Journal of Neurolinguistics*, *24* (1), 41–53.

Pettinato, M., & Verhoeven, J. (2009). Production and perception of word stress in children and adolescents with Down syndrome. *Down Syndrome Research & Practice*, *13*, 48–61.

Phonics Alive! (1996). *Advanced software*. Retrieved from http://www.phonicsalive.com.au.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *The Journal of the Acoustical Society of America*, *55* (2), 328–333.

Polson, P. G., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*, *36* (5), 741–773.

Prensky, M. (2007). *Digital game-based learning*. Saint Paul, MN: Paragon House Publishers.

Quesenbery, W. (2003). The five dimensions of usability. In M. Albers & B. Mazur (Eds.), *Content and complexity: Information design in technical communication* (pp. 81–102). Mahwah, NJ: Lawrence Erlbaum Associates.

Rochet-Capellan, A., & Dohen, M. (2015). Acoustic characterisation of vowel production by young adults with Down syndrome. In M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith, & J. Scobbie (Eds.), *18th International Congress of Phonetic Sciences (ICPhS)* (pp. 1–5). Glasgow, UK: IPA.

Rodríguez, W. R., Saz, O., & Lleida, E. (2012). A prelingual tool for the education of altered voices. *Speech Communication*, *54* (5), 583–600.

Santana, P. C., & Muro, B. P. (2011). Tangible interfaces to support the teaching of reading and writing to children with Down syndrome. *IEEE Learning Technology Newsletter*, *13* (2), 9–12.

Saz, O., Yin, S., Lleida, E., Rose, R., Vaquero, C., & Rodríguez, W. R. (2009). Tools and technologies for computer-aided speech and language therapy. *Speech Communication*, *51* (10), 948–967.

Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntix, W. H. E., Coulter, D. L., Craig, E. M., … Yeager, M. H. (2010). *Intellectual disability. Definition, classification, and systems of supports* (11th ed.). Washington, DC: American Association on Intellectual and Developmental Disabilities.

Shafie, A., Wan Ahmad, W. F., Mohd, N., Barnachea, J. J., Taha, M. F., & Yusuff, R. L. (2013). "SynMax": A mathematics application tool for Down syndrome children. In H. B. Zaman, P. Robinson, P. Olivier, T. K. Shih, & S. Velastin (Eds.), *Advances in visual informatics: Third international visual informatics conference* (pp. 615–626). Selangor, Malaysia: Springer.

Shahin, M., Ahmed, B., Parnandi, A., Karappa, V., McKechnie, J., Ballard, K. J., & Gutierrez- Osuna, R. (2015). Tabby talks: An automated tool for the assessment of childhood apraxia of speech. *Speech Communication*, *70*, 49–64.

Standen, P., Anderton, N., Karsandas, R., Battersby, S., & Brown, D. (2009). An evaluation of the use of a computer game in improving the choice reaction time of adults with intellectual disabilities. *Journal of Assistive Technologies*, *3* (4), 4–11.

Standen, P., Rees, F., & Brown, D. (2009). Effect of playing computer games on decision making in people with intellectual disabilities. *Journal of Assistive Technologies*, *3* (2), 4–12.

Stojanovik, V. (2011). Prosodic deficits in children with Down syndrome. *Journal of Neurolinguistics*, *24* (2), 145–155.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge, UK: Cambridge University Press.

Torrente, J., Del Blanco, Á., Moreno-Ger, P., & Fernández-Manjón, B. (2012). Designing serious games for adult students with cognitive disabilities. In T. Huang, Z. Zeng, C. Li, & C. Leung (Eds.), *Neural Information Processing: 19th International Conference ICONIP* (pp. 603–610). Doha, Qatar: Springer.

Troncoso, M. V., & Del Cerro, M. M. (1999). *Síndrome de Down: Lectura y escritura*. Barcelona, Spain: Masson.

Wan Ahmad, W. F., Muddin, H. N. B. I., & Shafie, A. (2014). Number skills mobile application for Down syndrome children. In J. Jaafar, A. Ahmad, A. Said, & N. Arshad (Eds.), *2014 International Conference on Computer and Information Sciences (ICCOINS)* (pp. 1–6). Kuala Lumpur, Malaysia: IEEE.

Wuang, Y.-P., Chiang, C.-S., Su, C.-Y., & Wang, -C.-C. (2011). Effectiveness of virtual reality using Wii gaming technology in children with Down syndrome. *Research in Developmental Disabilities*, *32* (1), 312–321.

Yussof, R. L., & Zaman, H. B. (2011). Scaffolding in early reading activities for Down syndrome. In H. B. Zaman, et al. (Eds.), *Visual informatics: Sustaining research and innovations* (pp. 180–192). Selangor, Malaysia: Springer.

About the Authors

César González-Ferreras is an Assistant Professor at the Department of Computer Science at the University of Valladolid, Spain. His research interests include human–computer interaction, spoken language processing and prosody recognition.

David Escudero-Mancebo is an Associate Professor at the Department of Computer Science at the University of Valladolid, Spain. He is co-author of several publications in the field of computational prosody, both concerning modeling of prosody for text-to-speech systems and prosodic labeling of corpora. His research interests also include human–computer interaction.

Mario Corrales-Astorgano is working on a Ph.D. in Computer Science at the University of Valladolid, Spain. His research interests mainly focus on human–computer interaction, especially in the design and evaluation of user interfaces for people with disabilities.

Lourdes Aguilar-Cuevas is an Associate Professor at the Department of Spanish Philology at the Autonomous University of Barcelona, Spain. She is a specialist in experimental phonetics and phonology of intonation in Spanish and Catalan. She has extensive experience in the field of prosody and automatic modeling of prosody.

Valle Flores-Lucas is an Associate Professor at the Department of Psychology at the University of Valladolid, Spain. Her research interests focus on the study of the problems of language development and communication, mainly in autism and Down syndrome, and their relationship with the theory of mind.

Appendix A. Script of the Interview with the Player

- Do you like video games?
- Have you ever played video games before?
- In general, did you like trying it?
- Would you like to play again in the future? Would you like to know how the story continues?
- Did you understand the story?
- Did you like the story?
- Did you like the parrot as the adventure's companion?
- Did you like talking to the characters of the video game?
- Would you like to create your own avatar (face, eyes, mouth, clothes, etc)?
- Did you learn something with the video game? What?
- Did you like the city images and the people?
- The video game seemed funny or boring.
- The video game seemed easy or hard.
- The video game seemed short or long.

Appendix B. Script of the Interview with the Teachers

- How have you used the video game?
- Do you think that the video game helped them to improve their pronunciation?
- Do you think that the context and the story of the video game are useful to perform the activities?
- Are the instructions suitable? Is the feedback suitable?
- Is it useful that the teacher can decide if the activity is right or wrong?
- What is the reaction of the students during the game sessions?
- Do you think that the virtual assistant is useful?
- What do you think of the graphics?
- Would you recommend the video game?

2.2. Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome

M. Corrales-Astorgano[1], D. Escudero-Mancebo[1], and C. González-Ferreras[1]. "Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome". Speech Communication, vol. 99, pp. 90–100, 2018.

[1]Departamento de Informática, Universidad de Valladolid, Valladolid, Spain

https://doi.org/10.1016/j.specom.2018.03.006

2.2 Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome

Speech Communication 99 (2018) 90–100

Contents lists available at ScienceDirect

Speech Communication

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

Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome

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

Departamento de Informática, Universidad de Valladolid, Valladolid, Spain

ARTICLE INFO

Keywords:
Speech characterization
Prosody
Down syndrome
Intellectual disabilities
Automatic classification
Perceptual test

ABSTRACT

There are many studies that identify important deficits in the voice production of people with Down syndrome. These deficits affect not only the spectral domain, but also the intonation, accent, rhythm and speech rate. The main aim of this work is the identification of the acoustic features that characterize the speech of people with Down syndrome, taking into account the different frequency, energy, temporal and spectral domains. The comparison of the relative weight of these features for the characterization of Down syndrome people's speech is another aim of this study. The openSmile toolkit with the GeMAPS feature set was used to extract acoustic features from a speech corpus of utterances from typically developing individuals and individuals with Down syndrome. Then, the most discriminant features were identified using statistical tests. Moreover, three binary classifiers were trained using these features. The best classification rate, using only spectral features, is 87.33%, and using frequency, energy and temporal features, it is 91.83%. Finally, a perception test has been performed using recordings created with a prosody transfer algorithm: the prosody of utterances from one group of speakers was transferred to utterances of another group. The results of this test show the importance of intonation and rhythm in the identification of a voice as non typical. As conclusion, the results obtained point to the training of prosody in order to improve the quality of the speech production of those with Down syndrome.

1. Introduction

Individuals with Down syndrome (DS) have problems in their language development that make their social relationships and their developmental ability more problematic (Cleland et al., 2010; Martin et al., 2009; Chapman, 1997). Many DS individuals have some physiological peculiarities that affect their voice production, such as a smaller vocal tract with respect to the tongue size or soft palatal shape, among others Guimaraes et al. (2008). Muscular hypotonia also affects their capabilities for performing a correct articulation, degrading the quality of the spectral characteristics of sounds (Markaki and Stylianou, 2011). In addition, hearing loss during childhood (Shott et al., 2001) and fluency deficits (Devenny and Silverman, 1990) influence the frequency, energy and temporal domains of the voice signal.

Although problems derived from physiological peculiarities are permanent (even if surgery Leshin, 2000 or prostheses Bhagyalakshmi et al., 2007 could ameliorate them), intonation and fluency deficits can be improved by speech therapy and training. There are tools available for this goal (Saz et al., 2009b; González-Ferreras et al., 2017) based on perception and production activities to be performed with the assistance of therapists who help patients to properly manage their breathing and intonation patterns. Although there is general consensus about the importance of improving prosody by training (see Kent and Vorperian, 2013 for a complete state of art revision), there are very few works that provide empirical evidence of the importance of the prosody related features (those belonging to fundamental frequency, energy and duration domains) with respect to other acoustic features belonging to the spectral domain.

The use of the video game described by González-Ferreras et al. (2017) has allowed the formation of a speech corpus, which has been used in this work to analyze and characterize the speech of people with Down syndrome. This corpus, described in Section 3.1, contains recordings of people with Down syndrome and typically developing people. Both groups recorded the same sentences, so statistical and perceptual tests have been used to compare the acoustic features of the two groups of speakers, so that the most relevant differences could be identified.

This work aims to find the best acoustic features to characterize the speech of people with Down syndrome. To do this, features of frequency, energy, temporal and spectral domains have been extracted from the recordings of the gathered corpus. In addition, the relative

* 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).

https://doi.org/10.1016/j.specom.2018.03.006
Received 18 December 2017; Received in revised form 21 February 2018; Accepted 13 March 2018
Available online 14 March 2018

2. COMPENDIO DE PUBLICACIONES

weight of each domain in the characterization of people with Down syndrome has been included in this paper, especially the comparison between the spectral and the other domains.

The methodology described above was developed to answer two main research questions (RQ):

- RQ1: Which are the most discriminative acoustic features between the recordings of speakers with Down syndrome and typically developing speakers?
  - **Issue 1.1:** Are there statistical differences between these features?
    Issue 1.2: Are these differences in accordance with what is expected or described in the state of the art?
- RQ2: What is the relative weight of the spectral features in comparison with the rest of the domains?
  Issue 2.1: What is the relative weight of the different features when identifying atypical speech using automatic classifiers?
  Issue 2.2: What is the relative weight of the different domains when identifying atypical speech in a perceptual test?

The structure of the article is as follows. Section 2 reviews related works from the state of the art and presents the innovation of our proposal. Section 3 describes the experimental procedure, including the corpus description, the features extraction process, the automatic classification experiment and the perceptual test. Section 4 shows the statistical test results of the different domain features, the automatic classification results and the perceptual test results. Finally, Section 5 describes the discussion and Section 6 the conclusions.

2. Background and related work

The age of the population selected for the study seems to be important for the results obtained, due to the physiological differences between children and adults. Concerning adults, Lee et al. (2009), Rochet-Capellan and Dohen (2015), Albertini et al. (2010) and Corrales-Astorgano et al. (2016) found significantly higher F0 values in adults with Down syndrome as compared to adults without intellectual disabilities. In addition, Lee et al. (2009) and Seifpanahi et al. (2011) found lower jitter (frequency perturbations) in adult speakers with Down syndrome. As for energy, Albertini et al. (2010) found significantly lower energy values in adults with Down syndrome. Moreover, Saz et al. (2009a) concluded that adults with Down syndrome had poor control over energy in stressed versus unstressed vowels. Albertini et al. (2010) found lower shimmer (amplitude perturbations) in male adults with Down syndrome than in adults without intellectual disabilities. Finally, temporal domain results depend on the unit of analysis employed. Saz et al. (2009a) found that people with cognitive disorders presented an excessive variability in vowel duration, while Rochet-Capellan and Dohen (2015) and Bunton and Leddy (2011) reported longer durations of vowels in adults with Down syndrome. Albertini et al. (2010) discovered a lower duration of words in male adults with Down syndrome. Moreover, people with Down syndrome present some disfluency problems. Although disfluency (stuttering or cluttering) has not been demonstrated as a universal characteristic of Down syndrome, it is a common problem of this population (Van Borsel and Vandermeulen, 2008; Devenny and Silverman, 1990; Eggers and Van Eerdenbrugh, 2017). These disfluencies can affect the speech rhythm of people with Down syndrome.

On the other hand, Zampini et al. (2016) indicated that children with Down syndrome had lower F0 than children without intellectual disabilities. Moura et al. (2008) found higher jitter in children with Down syndrome than children without intellectual disabilities. In terms of energy, Moura et al. (2008) indicated higher shimmer in children with Down syndrome than in children without intellectual disabilities.

The unit of analysis and the phonation tasks used by the researchers are different. Rochet-Capellan and Dohen (2015) used Vowel–Consonant–Vowel by syllabes, Saz et al. (2009a) and Albertini et al. (2010) recorded words, Rodger (2009) and Zampini et al. (2016) built these corpora using semi-spontaneous speech and Corrales-Astorgano et al. (2016) analyzed sentences. Lee et al. (2009) combined words, reading and natural speech. The majority of the studies are focused on the English language (Kent and Vorperian, 2013), but there are others focused on Italian (Zampini et al., 2016; Albertini et al., 2010), Spanish (Corrales-Astorgano et al., 2016; Saz et al., 2009a), French (Rochet-Capellan and Dohen, 2015) or Farsi (Seifpanahi et al., 2011).

The use of spectral features to assess pathological voice has frequently been applied in the literature. Dibazar et al. (2006) used MFCCs and pitch frequency with a hidden Markov model (HMM) classifier for the assessment of normal versus pathological voice using one vowel as the unit of analysis. Markaki and Stylianou (2011) suggested the use of modulation spectra for the detection and classification of voice pathologies. Markaki and Stylianou (2010) created a method for the objective assessment of hoarse voice quality, based on modulation spectra, using a corpus of sustained vowels. The voice quality was evaluated using the long term average spectrum (LTAS) and alpha ratio by Leino (2009). Although these works do not refer to people with Down syndrome, they do refer to some aspects that appear in this kind of speakers and we refer to them in the discussion section.

Formant frequency and amplitude have also been studied in people with Down syndrome. A larger vowel space in people with Down syndrome was found by Rochet-Capellan and Dohen (2015), while other studies denoted a reduction of the vowel space in children (Moura et al., 2008) and adults (Bunton and Leddy, 2011). Moreover, the voice of people with Down syndrome showed significantly reduced formant amplitude intensity levels (Pentz Jr, 1987).

In order to compare our study with the state of the art, a summary of other similar studies is shown in Table 1. A description of the corpus employed by these studies is shown in Table 2. To the best of our

Table 1
Results of different studies in the state of the art.

Author	Group	Frequency	Duration	Loudness
Rodger (2009)	Adults and Children	No differences		
Zampini et al. (2016)	Children	Good control for linguistics low for pragmatics. Lower F0.		
Saz et al. (2009a)	Adults and Children	Good control in pronounced vowels	Longer pronounced vowels. Dispersed mispronounced vowels	Low control of intensity in unstressed vowels
Albertini et al. (2010)	Adults	Higher F0	Lower duration (only for men)	Lower energy.Shimmer lower (only men)
Rochet-Capellan and Dohen (2015)	Adults	Higher F0	Longer vowels	
Lee et al. (2009)	Adults	Smaller pitch range. Higher F0. Lower jitter.		
Corrales-Astorgano et al. (2016)	Adults	Higher F0 excursions	More pauses to complete turns	Different range

2.2 Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome

Table 2
Description of the corpus used in the state of the art.

Author	Group	Down syndrome	Control	Type	Size	Language
Rodger (2009)	Adults and Children	22	52	Semi spontaneous	5 picture descriptions per speaker	English
Zampini et al. (2016)	Children	9	12	Semi spontaneous	20 minutes per speaker	Italian
Saz et al. (2009a)	Adults and Children	3	168	Words	9576 words (6 hours) Control. 684 words (38 minutes) Down syndrome	Spanish
Albertini et al. (2010)	Adults	30	60	Words	NA	Italian
Rochet-Capellan and Dohen (2015)	Adults	8	8	Vowel-consonant-vowel	144 per speaker	French
Lee et al. (2009)	Adults	9	9	Vowel. Reading. Natural speech	3 vowels per speaker. 1 reading per speaker. 1 minute per speaker	English
Corrales-Astorgano et al. (2016)	Adults	18	20	Sentences	479 utterances	Spanish

knowledge, our study is one of the first to analyze some features from the frequency, energy, temporal and spectral domains together. These features were extracted from the same recordings, which can help in the study of the relative importance of each domain in the characterization of the speech of people with Down syndrome. The use of a standard feature set (extended Geneva Minimalistic Acoustic Parameter Set, eGeMAPS; detailed in Section 3.2 and Appendix A) can reduce the extraction methodology dependence, which can make it easier to compare the results of different studies.

Perceptual studies show mixed results. Moura et al. (2008) described the voice of children with Down syndrome as being statistically different from the voice of children without intellectual disabilities in five speech problems: grade, roughness, breathiness, asthenic speech and strained speech. Moran and Gilbert (1982) judged the voice quality of adults with Down syndrome as hoarse. In addition, Rodger (2009) noted discrepancies between perceptual judgments of pitch level and acoustic measures of F0. In our study, we did not want to compare each acoustic measure with a perceptual judgment of the same feature. Our aim is the assessment of the domain relevance in the identification of a recording as being from a person with Down syndrome, using automatic classifiers and perceptual tests.

3. Experimental procedure

Fig. 1 shows the experimental methodology that we have followed. Firstly, the speech corpus recorded by people with Down syndrome and by typically developing people was gathered. Secondly, acoustic features were extracted from all the recordings of each corpus and a statistical test to analyze the differences between groups was carried out. Finally, the automatic classification experiment was carried out, in which the features with significant differences were used.

3.1. Corpus collection

We developed a computer video game to improve the prosodic and communication skills of people with Down syndrome (González-Ferreras et al., 2017). This video game is a graphic adventure game where users have to use the computer mouse to interact with the elements on the screen, listen to audio instructions and sentences from the characters of the game, and record utterances using a microphone in different contexts. The video game was designed using an iterative methodology in collaboration with a school of special education located in Valladolid (Spain). The feedback provided by teachers of special education was complemented by research into the difficulties of this population to use information and communication technologies. They have some difficulties, such as attention deficit (Martínez et al., 2011), lack of motivation (Wuang et al., 2011), or problems with the short term memory (Chapman and Hesketh, 2001) that had to be taken into account when developing the video game. The game was developed for the Spanish language.

Inside the narrative of the game, some learning activities were included to practice communication skills. There are three different types of activities: comprehension, production and visual. Firstly, the comprehension activities are focused on lexical-semantic comprehension and on improving prosodic perception in specific contexts. Secondly, production activities are focused on oral production, so the players are encouraged by the game to train their speech, keeping in mind such prosodic aspects as intonation, expression of emotions or syllabic emphasis. At the beginning of these activities, the video game introduces the context where the sentence has to be said. Then, the game plays the sentence and the player must utter the sentence while it is shown on the screen. The production activities include affirmative, exclamatory and interrogative sentences. Finally, visual activities include other activities designed to add variety to the game and to reduce the feeling of

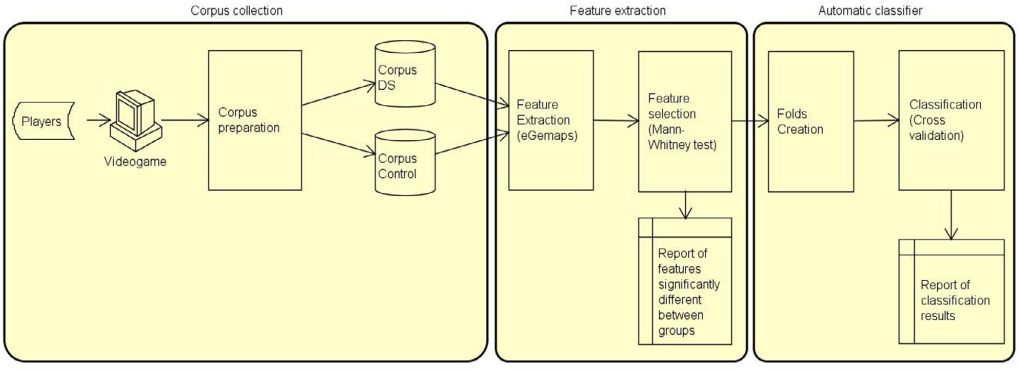

Fig. 1. Scheme of the experimental procedure which includes corpus collection, feature extraction and automatic classification.

monotony while playing.

The video game collected examples of sentences with different modalities (i.e. declarative, interrogative and exclamatory). Usually, the intonation patterns vary depending on the modality. Neutral declarative sentences usually end with a decline to a low tone, while total interrogatives end with an upgrade to a high pitch. On the other hand, partial interrogative sentences, which are characterized by an interrogative element at the beginning of the sentence, start with a high tone associated with that interrogative element and usually end with a fall. Finally, exclamatory sentences are usually a marked variation of the corresponding declarative, so the variation lies basically in such aspects as the intensity, volume and tonal range used by the speaker.

Moreover, the combination of different sentences allows the inclusion of inflections that indicate a particular segmentation in oral production. Depending on the context and speed of elocution, these inflections may correspond to a pause, which implies a silence and, normally, the end of the sentence, or a semi-pause, which implies an intonation change in the same sentence. For instance, one of the examples collected in the corpus includes the three modalities and forces the speaker to make a pause between sentences: *¡Hola! ¿Tienen lupas? Quería comprar una. (Hello! Do you have magnifiers? I wanted to buy one*). In other cases, the tonal inflection corresponds to a semi-pause involving no change of modality or silence: *¡Hasta luego, tío Pau! (See you later, uncle Pau!*). Thus, the combination of these types of inflection allows the collection of examples with different segmentation. The sentences recorded can be seen in Table 3.

The recording sessions were carried out in the same facilities of the centers where the players attended their regular classes to assure the comfort of the players. In addition, a staff member of the centers was always with the players. The players were selected by the staff members because the distinct cognitive abilities of each student limited their possibilities as potential players, as some of them were not able to follow the structured process of the game in a reliable way. Eighteen speakers with Down syndrome participated, 11 males (chronological ages: 16, 16, 18, 20, 21, 21, 23, 24, 25, 26 and 30) and 7 females (chronological ages: 16, 17, 18, 19, 21, 22, 25). All of them were native speakers of Spanish, aged 16 to 30. They were students of two special education schools located in Valladolid and Barcelona (Spain) and have a moderate or mild intellectual disability. Besides, to reduce the ambient noise in the recording process, the players used a headset with a microphone incorporated (Plantronics USB headset). In addition, players recorded a different number of sentences, depending on their performance in the video game and the number of game sessions they did. It should be noted that for the production activities, not all speakers with Down syndrome reproduced the target sentence exactly. Some of them had hearing problems, while others had reading difficulties or cluttering derived from their intellectual disability.

Table 3
Sentences included in the corpus.

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 would 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 4
Number of users and recordings of each group of the corpus.

User type	#Users	#Recordings	Length (seconds)
Control (TD)	22	250	650
Down syndrome (DS)	18	349	1442

To obtain a control sample of the recordings, twenty two adult speakers without any intellectual disability, 13 males and 9 females, were recorded. Therefore, two groups representing different populations were thus obtained: typically developing adults (TD) and people with Down syndrome (DS). Table 4 shows the number of users of each group of speakers, the number of recordings made by them and the total length in seconds of the recordings.

3.2. Feature extraction

Acoustic low-level descriptors (LLD) and temporal features were automatically extracted from each recording using the openSmile toolkit (Eyben et al., 2013). Two minimalistic feature sets were used. On the one hand, these sets provided enough features to characterize the audio recordings. On the other hand, we avoid the problem of having too many parameters relative to the number of observations. This problem can produce overfitting in the training phase, because the classifier adapts to the concrete set of inputs. This adaptation can produce good classification results for this particular set, but negatively affects the generalization capacity of the classifier. The Geneva Minimalistic Standard Parameter Set (GeMAPS) and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), described by Eyben et al. (2016), were selected. The features extracted from each recording are sorted into four groups:

- Frequency related features: fundamental frequency and jitter.
- Energy related features: loudness, shimmer and Harmonics-to-Noise Ratio.
- Spectral features: alpha ratio, Hammarberg index, spectral slope, formant 1, 2, 3 relative energy, harmonic difference H1-H2, harmonic difference H1-A3, formant 1, 2, 3 frequency and formant 1, 2, 3 bandwidth.
- Temporal features: the rate of loudness peaks per second, mean length and standard deviation of continuous voiced and unvoiced segments and the rate of voiced segments per second, approximating the pseudo syllable rate.

In total, there are 25 LLD. The arithmetic mean and the coefficient of variation are calculated on these 25 LLD. Some functionals are applied to fundamental frequency and loudness: 20-th, 50-th, and 80-th percentile, the range of 20-th to 80-th percentile, and the mean and standard deviation of the slope of rising/falling signal parts. All these functionals are computed by the openSmile toolkit. In addition, the process used by the openSmile toolkit to extract the eGeMAPS features did not differentiate between silences and unvoiced regions, which can produce errors in the functions applied to each feature. Therefore, the Praat software (Boersma, 2006) was used to extract all silences from each recording and these silences were excluded from the analysis process.

Furthermore, 4 additional temporal features were added: the silence and sounding percentages, silences per second and the mean silences. These new features were added to improve the information about the temporal characterization of the recordings. In this case, the initial and final silence of each recording were excluded from the analysis process

2.2 Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome

because their lengths were different due to the recording process. To sum up, the acoustic feature set contains 88 features from the eGeMAPS feature set and 4 new features introduced from the research team (92 features).

A statistical test was used to detect the significant differences between the features extracted from the recording of each group. The Mann-Whitney non-parametric test was used. Only the features with a p-value lower than 0.01 were selected for analysis and classification.

3.3. Automatic classification

In order to make an automatic classification of the recordings, the Weka machine learning toolkit (Hall et al., 2009) was used. This toolkit permits to a collection of machine learning algorithms to be accessed for data mining tasks. Three different classifiers were used to compare their performance: the C4.5 decision tree (DT), the multilayer perceptron (MLP) and the support vector machine (SVM).

In addition, the 10-fold cross validation technique was used to create the training and testing datasets. To avoid classifier adaptation, all folds were created by recordings of different speakers. Therefore, the recordings of each speaker were joined in the same fold and each fold was balanced in terms of the number of recordings.

To analyze the performance of the classification, we used the classification rate. The unweighted average recall (UAR) (Schuller et al., 2016) was also used. This metric is the mean of sensitivity (recall of positive instances) and specificity (recall of negative instances). UAR was chosen as the classification metric because it equally weights each class regardless of its number of samples, so it represents more precisely the accuracy of a classification test using unbalanced data.

3.4. Perception test

In order to evaluate the impact of prosody in the perception of the listeners, we used prosody transfer techniques. These techniques have previously been used in other studies of the state of the art. For instance, Luo et al. (2017) investigated the role of different prosodic features in the naturalness of English L2 speech. The prosodic modification method was applied to native and L2 learners' speech. Later, they used a perceptual test to evaluate the impact of prosody modification. A similar methodology was used by Escudero et al. (2017), where the characteristic prosodic patterns of the style of different groups of speakers was investigated. After the prosodic modification of the utterances, the characteristic prosodic patterns were validated using a perceptual test. The procedure described in Escudero et al. (2017) for transferring prosody is used in the experiments reported in this paper.

Fig. 2 shows the experimental procedure used to perform the perception test. The sentence *¡Hola tío Pau! ¿Sabes donde vive la señora Luna?* (*Hello uncle Pau! Do you know where Mrs Luna lives?*) recorded by all the speakers was selected. This sentence was selected because of its prosodic richness (combining an affirmative and an interrogative sentence), because it was used in another of our studies (González-Ferreras et al., 2017) and because it was the most recorded sentence. To obtain a phonetic segmentation of the recordings, the BAS web services (Schiel, 1999; Kisler et al., 2017) were used. This tool returns the time intervals of each phoneme using the audio file and the transcription as inputs. Manual revision of the segmentation was necessary to correct transcription errors. The sentence was recorded by 22 TD speakers and by 16 speakers with DS. However, each speaker did not have the same number of recordings. In total, there were 62 recordings.

Once the segmentation was corrected, a prosody transfer algorithm implemented in Praat (Boersma, 2006) was executed. This algorithm transfers, phoneme by phoneme, the pitch, energy and duration from one audio to another. Therefore, the new audio file contains the original utterance but with the prosody transferred from another utterance. The algorithm was executed combining the audios of each speaker with the audios of the rest of the speakers, so, in total, 3525 audio files were generated (not all the speakers had the same number of recordings). As a result, there are four types of audio files, as shown in Fig. 2. Five audio files of each type were selected randomly for the perception test, so the test included twenty audio files, balanced in terms of gender.

The perception test was performed using a web application. First, personal information of the evaluator was collected. Then, the twenty audio files selected in the previous phase were shown randomly. The evaluators have to answer the following question for each utterance: *keeping in mind the way of speaking, do you think that the person who is speaking has intellectual disabilities? Ignore the audio distortion produced by the non natural voice synthesis.* The possible answers to the question were in a 5-point Likert scale: 1 means "no way" and 5 means "very sure". Thirty evaluators judged each utterance using this scale. People without any specific background on speech therapies were selected for this test, as we were interested in the perception of normal people concerning the importance of prosody in the identification of speech from people with intellectual disability.

4. Results

4.1. Characterization results

Table 5 shows the features with statistically significant differences (Mann–Whitney test with p-value < 0.01) related to frequency, energy

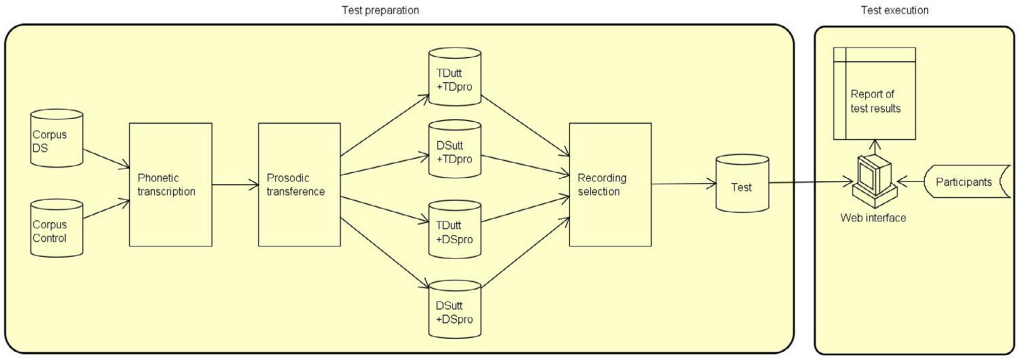

Fig. 2. Experimental procedure followed to perform the perceptual test. The utterances used in the test were: TDutt + TDpro (utterance of a TD person with prosody transferred from an utterance of another TD person), DSutt + TDpro (utterance of a person with DS with prosody transferred from an utterance of a TD person), TDutt + DSpro (utterance of a TD person with prosody transferred from an utterance of a person with DS) and DSutt + DSpro (utterance of a person with DS with prosody transferred from an utterance of another person with DS).

Table 5
List of frequency, energy and temporal features with higher statistically significant differences (Mann-Whitney test with p-value < 0.01), sorted by mean differences. The meaning of the features in column *Variable* can be seen in Appendix A. The units are reported in (Eyben et al., 2016).

Variable	Control	Control (CI 95%)	Down syndrome	Down syndrome (CI 95%)
F0 domain				
F0_stddevRisingSlope	166.17 ± 231.44	(137.35,195.01)	220.85 ± 273.67	(192.08,249.62)
jitter_stddevNorm	1.15 ± 0.39	(1.11,1.21)	1.46 ± 0.47	(1.42,1.52)
jitter_mean	0.04 ± 0.02	(0.045,0.050)	0.03 ± 0.01	(0.035,0.039)
F0_pctlrange	4.63 ± 1.9	(4.4,4.88)	3.91 ± 2.88	(3.61,4.22)
F0_percentile20	26.89 ± 4.49	(26.33,27.45)	30.32 ± 4.63	(29.84,30.81)
F0_percentile50	29.18 ± 4.22	(28.66,29.71)	32.33 ± 4.28	(31.89,32.79)
F0_mean	29.3 ± 4.11	(28.79,29.82)	32.38 ± 4.14	(31.95,32.82)
F0_stddevNorm	0.13 ± 0.07	(0.129,0.147)	0.12 ± 0.07	(0.116,0.132)
F0_percentile80	31.52 ± 4.34	(30.99,32.07)	34.24 ± 4.67	(33.75,34.73)
Energy domain				
loudness_percentile20	0.95 ± 0.38	(0.91,1.01)	1.77 ± 1.03	(1.66,1.88)
loudness_percentile50	1.93 ± 0.73	(1.84,2.02)	3.29 ± 2.22	(3.06,3.53)
loudness_mean	2.09 ± 0.78	(1.99,2.19)	3.37 ± 1.99	(3.17,3.58)
loudness_percentile80	3.15 ± 1.24	(3,3.31)	4.9 ± 2.94	(4.6,5.22)
loudness_pctlrange	2.19 ± 0.96	(2.08,2.32)	3.13 ± 2.06	(2.92,3.35)
loudness_stddevRisingSlope	15.3 ± 7.18	(14.41,16.2)	19.63 ± 14.24	(18.14,21.13)
loudness_stddevNorm	0.57 ± 0.07	(0.57,0.58)	0.49 ± 0.07	(0.48,0.5)
shimmer_mean	1.55 ± 0.38	(1.51,1.61)	1.36 ± 0.37	(1.32,1.4)
shimmer_stddevNorm	0.86 ± 0.14	(0.84,0.88)	0.78 ± 0.16	(0.77,0.8)
Temporal domain				
silencePercentage	0.1 ± 0.11	(0.09,0.12)	0.22 ± 0.19	(0.2,0.24)
silencesMean	0.16 ± 0.2	(0.14,0.19)	0.31 ± 0.3	(0.28,0.35)
StddevVoicedSegmentLengthSec	0.15 ± 0.08	(0.14,0.16)	0.25 ± 0.2	(0.23,0.27)
MeanVoicedSegmentLengthSec	0.26 ± 0.15	(0.25,0.29)	0.44 ± 0.39	(0.41,0.49)
silencesPerSecond	0.39 ± 0.38	(0.35,0.44)	0.57 ± 0.4	(0.53,0.62)
VoicedSegmentsPerSec	3.42 ± 1.06	(3.29,3.55)	2.47 ± 1.04	(2.37,2.59)
loudnessPeaksPerSec	5.76 ± 1	(5.64,5.89)	4.39 ± 0.94	(4.29,4.49)
MeanUnvoicedSegmentLength	0.05 ± 0.02	(0.05,0.06)	0.06 ± 0.03	(0.06,0.07)
soundingPercentage	0.89 ± 0.11	(0.88,0.91)	0.77 ± 0.19	(0.76,0.8)

and temporal domains, sorted by mean differences. In the case of frequency, 9 of 12 features present significant differences. The first rows (from F0_stddevRisingSlope to jitter_mean) refer to the temporal evolution of the F0 contour. In all cases, figures present a higher value for speakers with Down syndrome, both when the *stddev* value is analyzed or the *Risingslope* and *jitter* (*jitter* value is lower because it focuses on the periods, which are the inverse of the F0 values). The last rows refer to mean values, coefficient of variation, ranges and percentiles of the F0 contour (from F0_pctlrange to F0_percetile80). Speakers with Down syndrome exhibit higher values than the speakers of the control group in all the cases, with a lower coefficient of variation in the Down syndrome group. These results seem to indicate that the participants with Down syndrome use higher F0 values with more temporal changes in the F0 contours.

There are 9 of 14 energy features that present statistically significant differences (Mann-Whitney test with p-value < 0.01), as shown in Table 5. The first four rows (from loudness_percentil20 to loudness_pctlrange) refer to mean, range and percentile values. Values are higher for speakers with Down syndrome in all the cases. The last columns refer to the temporal variation of the energy values. In this case, Down syndrome speakers exhibit lower values. These results seem to indicate that participants with Down syndrome speak louder with less variation in the energy.

With respect to the temporal features displayed in Table 5, 9 of 10 features presented statistically significant differences (Mann-Whitney test with p-value < 0.01). Speakers with Down syndrome use more pauses and they are longer (higher silencePercentage, silencePerSecond and silenceMean). The length of the voiced segment is longer, indicating that participants with Down syndrome speak more slowly.

As for spectral features (Table 6), 34 of 56 features showed statistically significant differences (Mann–Whitney test with p-value < 0.01). Results show that the LTAS could be a useful instrument to detect differences, as clear differences appear when the features related with slope, Hammarberg and alpha index are taken into account. Formant 1 and Formant 3 (to a lower degree) also allow differences to be identified. As expected, MFCC values (the four analyzed) permit both groups to be separated. With respect to the variables related with the harmonic differences, only two variables appear in the list: logRelF0H1A3_stddevNorm and logRelF0H1A3_mean.

4.2. Classification results

Table 7 shows the classification results in the task of identifying the group of the speaker (TD or SD) of each utterance. The classifiers explained in Section 3.3 and the selected features presented in the previous section were used. Only the features with significant differences between TD and DS groups are used. DT shows the lower classification results in all feature groups. MLP shows a better performance using frequency (UAR 0.64), temporal (UAR 0.78), frequency + energy + temporal (UAR 0.91) and all (UAR 0.95) feature groups. SVM works better with energy features (UAR 0.78). The results using spectral features are the same in MLP and SVM classifiers (UAR 0.87).

In addition, the best classification results are obtained using all features, independently of which classifier is used. Frequency features show the worst performance when they are used alone. Energy and temporal features have similar results, with only 9 features per group.

When frequency, energy and temporal features are used together, the performance is noticeably better than using each group separately. Finally, spectral features show a slightly worse performance than all and frequency + energy + temporal features.

4.3. Perception test results

Table 8 shows the results of the perception test and Fig. 3 visually presents the differences between the groups. When the prosody of TD speakers was transferred to utterances of TD speakers, 84% of the answers identified the audios as TD speakers (answer 1 of row TDutt + TDpro). In this case, the doubts in the identification of the audio files

2.2 Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome

Table 6
List of spectral features with higher statistically significant differences (Mann–Whitney test with p-value < 0.01), sorted by mean differences. The meaning of the features in column *Variable* can be seen in Appendix A. The units are reported in (Eyben et al., 2016).

Variable	Control	Control (CI 95%)	Down syndrome	Down syndrome (CI 95%)
LTAS related features				
slopeV0500_mean	0 ± 0.03	(0,0.01)	0.05 ± 0.03	(0.056,0.063)
slopeUV0500_mean	−0.06 ± 0.04	(−0.07,−0.06)	0.05 ± 0.03	(0.02,0.03)
slopeV0500_stddevNorm	−1.12 ± 13.82	(−2.85,0.6)	0.69 ± 2.64	(0.41,0.97)
alphaRatioUV_mean	−12.06 ± 11.37	(−13.48,−10.65)	1.07 ± 6.37	(0.41,1.75)
hammarbergIndexUV_mean	20.79 ± 13.51	(19.11,22.48)	5.4 ± 7.24	(4.64,6.16)
alphaRatioV_mean	−11.79 ± 5.52	(−12.49,−11.11)	−8.46 ± 5.55	(−9.05,−7.88)
hammarbergIndexV_mean	20.8 ± 7.06	(19.93,21.69)	16.35 ± 7.14	(15.61,17.11)
hammarbergIndexV_stddevNorm	0.48 ± 0.67	(0.4,0.57)	0.57 ± 1.01	(0.47,0.68)
slopeV5001500_mean	−0.02 ± 0	(−0.03,−0.02)	−0.02 ± 0	(−0.021,−0.020)
spectralFlux_mean	1.96 ± 1.09	(1.83,2.1)	2.94 ± 2.32	(2.7,3.19)
spectralFluxUV_mean	1.4 ± 1.35	(1.23,1.57)	2.1 ± 2.11	(1.88,2.32)
spectralFluxV_mean	2.11 ± 1.12	(1.98,2.26)	3.13 ± 2.53	(2.87,3.4)
spectralFlux_stddevNorm	0.72 ± 0.19	(0.7,0.75)	0.67 ± 0.12	(0.66,0.69)
MFCC related features				
mfcc3_stddevNorm	0.25 ± 24.92	(−2.85,3.36)	−54.35 ± 1039.94	(−163.68,54.98)
mfcc2V_mean	1.49 ± 7.41	(0.58,2.42)	−2.45 ± 6.88	(−3.17,−1.73)
mfcc4_stddevNorm	1.54 ± 44.52	(−4.01,7.09)	−2 ± 19.36	(−4.04,0.03)
mfcc2_stddevNorm	1.97 ± 26.17	(−1.29,5.23)	−1.16 ± 27.11	(−4.01,1.69)
mfcc2_mean	4.05 ± 7.08	(3.18,4.94)	−2.32 ± 6.45	(−3,−1.64)
mfcc4V_stddevNorm	−1.23 ± 9.51	(−2.42,−0.05)	−0.45 ± 4.73	(−0.96,0.04)
mfcc4_mean	−11.17 ± 7.74	(−12.14,−10.21)	−17.34 ± 9.91	(−18.39,−16.3)
mfcc3V_stddevNorm	−0.78 ± 71.43	(−9.68,8.11)	−0.28 ± 21.18	(−2.51,1.94)
mfcc4V_mean	−14.75 ± 8.58	(−15.82,−13.68)	−18.3 ± 10.83	(−19.44,−17.17)
mfcc1V_mean	26.42 ± 7.31	(25.51,27.34)	20.93 ± 9.61	(19.93,21.95)
mfcc1_mean	22.52 ± 7.73	(21.56,23.49)	18.16 ± 9.95	(17.11,19.21)
Formants related features				
F3amplitudeLogRelF0_stddevNorm	−1.18 ± 0.25	(−1.22,−1.16)	−1.36 ± 0.41	(−1.41,−1.32)
F2amplitudeLogRelF0_mean	−49.47 ± 17.65	(−51.68,−47.28)	−42.63 ± 20.55	(−44.79,−40.47)
F2amplitudeLogRelF0_stddevNorm	−1.35 ± 0.26	(−1.39,−1.32)	−1.54 ± 0.61	(−1.61,−1.48)
F1bandwidth_stddevNorm	0.2 ± 0.08	(0.19,0.21)	0.23 ± 0.09	(0.22,0.24)
F1frequency_stddevNorm	0.35 ± 0.09	(0.34,0.37)	0.4 ± 0.09	(0.39,0.41)
F3frequency_stddevNorm	0.09 ± 0.02	(0.095,0.102)	0.1 ± 0.02	(0.1,0.11)
F3frequency_mean	2665.98 ± 145.97	(2647.81,2684.17)	2643.51 ± 203.27	(2622.15,2664.89)
F3amplitudeLogRelF0_mean	−53.64 ± 17.44	(−55.82,−51.47)	−45.02 ± 19.5	(−47.08,−42.98)
Harmonic differences features				
logRelF0H1A3_stddevNorm	1.6 ± 16.02	(−0.39,3.6)	0.18 ± 7.44	(−0.6,0.97)
logRelF0H1A3_mean	18.91 ± 6.26	(18.13,19.69)	15.86 ± 7.09	(15.12,16.61)

Table 7
Classification results for identifying the group of the speaker. Classification rate (c. rate) and UAR using different feature sets and different classifiers are reported. The features used are those with significant differences between TD and DS groups. The classifiers are decision tree (DT), support vector machine (SVM) and multilayer perceptron (MLP). # is the number of input features in each set.

		SVM		MLP		DT	
Set	#	C. Rate	UAR	C. Rate	UAR	C. Rate	UAR
Frequency	9	62.67	0.61	64.33	0.64	60.17	0.60
Energy	9	79.33	0.78	76	0.76	72.5	0.71
Temporal	9	76.83	0.76	77.83	0.78	74.33	0.75
Frequency + Energy + Temporal	27	90	0.9	91.83	0.91	82	0.82
Spectral	34	87.33	0.87	87.33	0.87	84.33	0.84
All	61	94.17	0.94	95.17	0.95	86.5	0.87

Table 8
Number of responses of the perception tests for each type of audio file. A response of 1 means "no way" and 5 means "very sure" in the identification of the audio file as a speaker with Down syndrome. NR means no response. TDutt + TDpro means utterance of a TD person with prosody transferred from an utterance of another TD person; DSutt + TDpro means utterance of a person with DS with prosody transferred from an utterance of a TD person; TDutt + DSpro means utterance of a TD person with prosody transferred from an utterance of a person with DS; and DSutt + DSpro means utterance of a person with DS with prosody transferred from an utterance of another person with DS.

Type	1	2	3	4	5	NR	Total
TDutt + TDpro	124	15	3	1	4	3	150
DSutt + TDpro	42	42	31	18	11	6	150
TDutt + DSpro	17	21	34	43	31	4	150
DSutt + DSpro	1	11	26	49	56	7	150

as TD or DS represent only 2% of the answers (answer 3 of row TDutt + TDpro). On the other hand, when the prosody of DS speakers was transferred to utterances of DS speakers, 73% of the answers identified the audios as DS speakers (answers 4 and 5 of row DSutt + DSpro). In this case, the doubts in the identification of the audio files as TD or DS represent 18% of the answers (answer 3 of row DSutt + DSpro), and the identifications as TD are only 8% (answers 1 and 2 of row DSutt + DSpro).

The answers given about the audio files that combined utterances of one group with prosody of the other group present much more variability. However, prosody had more influence in the identification process than the original utterance. When the prosody of TD speakers was transferred to utterances of speakers with DS, 58% of the answers identified the audios as TD speakers (answers 1 and 2) versus only 20% of DS identifications (answers 4 and 5). On the other hand, 51% of the answers identified the audios as speakers with DS (answers 4 and 5) when the prosody of speakers with DS was transferred to an utterance of TD speakers, versus only 26% of TD identifications (4 and 5 answers). In both cases, the number of answers 3 is relevant (22% and 23% of answers 3, respectively).

Moreover, two statistical tests were used to compare the answers obtained. The results of the Kruskal–Wallis non-parametric test showed significant differences (with a p-value < 0.001) between the answers given to the four groups (TDutt + TDpro, DSutt + TDpro, TDutt + DSpro

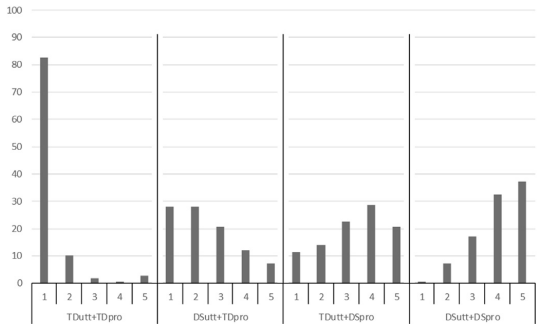

Fig. 3. Results of the perception tests for each type of audio file. TDutt + TDpro means utterance of a TD person with prosody transferred from an utterance of another TD person; DSutt + TDpro means utterance of a person with DS with prosody transferred from an utterance of a TD person; TDutt + DSpro means utterance of a TD person with prosody transferred from an utterance of a person with DS; and DSutt + DSpro means utterance of a person with DS with prosody transferred from an utterance of another person with DS.

and DSutt + DSpro). Furthermore, the Mann–Whitney non-parametric test was used to compare each group with the others, in groups of two. All the comparisons showed significant differences (p-value $<$ 0.001).

5. Discussion

5.1. Characterization of the speech of people with Down syndrome

Fundamental frequency is significantly higher in speakers with Down syndrome. The same results were found by Albertini et al. (2010), Rochet-Capellan and Dohen (2015) and Lee et al. (2009). In addition, the F0 range is lower in speakers with Down syndrome, which can be explained by a less melodious intonation. Continuing with frequency, jitter is significantly lower in the DS group, as found by Lee et al. (2009) and by Seifpanahi et al. (2011).

Concerning temporal features, on the one hand, the number of continuous voiced regions per second is lower in the speakers with Down syndrome, which means that the oral production of speakers with Down syndrome was slower than that of control speakers. Reading difficulties that some people with Down syndrome present can have influenced these results. On the other hand, Van Borsel and Vandermeulen (2008) found disfluencies in Down syndrome speaking, such as cluttering and stuttering. These disfluencies can produce the insertion of more silences and the presence of more temporal variety in the speech of people with Down syndrome, as found in this study.

In terms of energy, loudness features were found to be significantly higher in the speakers with Down syndrome and its range was higher. This result contradicts that reported by Albertini et al. (2010), which showed lower energy values in speakers with Down syndrome. Another study focused on vowels (Saz et al., 2009a) found an increase in the energy of unstressed vowels in Down syndrome speakers. Energy is always a difficult variable in the analysis of prosody, as its values are very dependent on the recording conditions: the dynamic range of the microphone and the distance between the speaker and the microphone. On the other hand, some of the participants have slight hearing problems, which may be another possible explanation for the higher energy values.

Our corpus also permitted the detection of differences related with the spectral features. Table 6 highlights the fact that LTAS has been proposed in Gauffin and Sundberg (1977) for the identification of breathy and hypokinetic voice. The relative amplitude of the first harmonic was also related with breathy voices by Hillenbrand and Houde (1996). The speech of people with DS is described as breathy by Wold DC (1979) and dysphonic by Moran (1986). MFCC features are commonly used in speaker recognition applications (Martinez et al., 2012), as they are representative of the vocal tract shape (Dusan and Deng, 1998). The relative importance of the MFCC features on the characterization of the speech of people with DS (as shown in Table 6) could thus be justified by the special anatomy of the tongue, palate, jaw, etc. of this type of speaker (Rodger, 2009). MFCC has also been used to identify nasality by Yuan and Liberman (2011) which is another aspect that has been related with the speech of people with DS in many works (Kent and Vorperian, 2013). The relative position of the formants has been associated with the degree of nasality in many works (House and Stevens, 1956; Huffman, 1989) which was also highlighted in our results table.

Finally, people with DS present hypotonia of muscles and difficulties in motor control, which affect the movement of the lips, tongue and jaw, with the consequent impact on spectral features already mentioned. The lack of muscular strength could also be another reason justifying the slower speech. As hypotonia could also affect the diaphragm, the energy values should have been lower. We hypothesize that the reason why higher values of energy were obtained could be due to the extra effort made by students to correctly complete the activities.

5.2. Relative impact of prosody

The experimental results obtained show that the features concerning the frequency, energy and temporal domains have the same or a greater impact than the spectral domain features to identify the speech of people with Down syndrome:

- There are a high number of features out of the spectral domain that present significant differences between speakers with Down syndrome and speakers without intellectual disabilities.
- Spectral features achieve high classification rates (up to 87%), but classification rates of frequency, energy and temporal features together are higher than spectral features (up to 91.83%).
- Utterances of control speakers with transferred frequency, energy and phoneme duration from speakers with Down syndrome are mostly perceived as anomalous voice. In the same way, utterances of speakers with Down syndrome with transferred frequency, energy and phoneme duration from control speakers are mostly perceived as typical speech.

To the best of our knowledge, there are few studies that assess, in an experimental way, the relative weight of prosody in the perception of speech of people with Down syndrome as a non typical voice. The differences between speakers with Down syndrome and control speakers in the spectral domain can be derived from physiological peculiarities in their phonological system. Some could be corrected by surgery, but others are impossible to be corrected. However, frequency, energy and temporal characteristics can be trained using speech therapy techniques focusing on breathing and repetition of activities. The results obtained in this paper show the potential benefits of prosody training.

The distance between the prosodic features of speakers with Down syndrome and those of control speakers can be used to devise a quality metric to be included in computer assisted pronunciation training applications. Our future work on the implementation of an automatic evaluation module of voice quality is expected to benefit from the results of this paper. This module is to be included in our speech training tools (González-Ferreras et al., 2017), so spectral features will be useful to identify a recording as a non typical speech, while prosody analysis will be necessary for the evaluation of the players' improvement over the different game sessions.

5.3. Limitations

The corpus size in speech analysis studies is very important to

2.2 Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome

achieve representative results. The recording of a corpus of speech of people with Down syndrome is always challenging because of the special characteristics of these speakers (attention deficit and problems with short term memory, among others). Our video game has allowed the recording of a speech corpus whose size is bigger than other speech corpora used in other studies (see Table 2). Although the corpus size could be larger, the statistical tests carried out guarantee that the corpus has the necessary size to obtain significant results. In addition, new recordings are currently being obtained due to the use of the video game in a school of special education.

The heterogeneity of the population with Down syndrome can have an influence on the correct generalization of the results. However, the methodology presented in this paper can be applied to individuals with the aim of identifying the concrete features that they are using wrongly. Moreover, the relative impact of these features in the identification of their speech as pathological can be analyzed.

6. Conclusions

The speech characterization experiment presented in this article has allowed us to find significant differences between the speech of individuals with Down syndrome and those of the control group that affect the use of a set of acoustic variables related to frequency, energy, temporal and spectral domains. The use of these variables in an experiment of automatic identification allows very high classification rates (above 95%) to be obtained. If these variables are used independently, the classification rates decrease, the highest being those obtained using the spectral features. However, the importance of the rest of the variables becomes clear, because when only the variables related to frequency, energy and temporal domains are used, the classification rate can be higher than that obtained using the spectral features.

A perception experiment, based on prosody transfer, allowed us to verify the high relative importance of the prosodic variables of frequency, energy and temporal domains regarding the perception of atypical speech. An adequate control of these variables in utterances of speakers with Down syndrome allows us to change the perception of them, even though the voice quality is not modified. Besides, transferring the prosody from speakers with Down syndrome to speakers of the control group means the utterances will be perceived, to a large degree, as if they were from speakers with Down syndrome. This result encourages the use of methodologies for training prosody as a means for improving the overall quality of the oral production of Down syndrome speakers.

Acknowledgments

The work described in this paper was supported (1/2016-12/2017) by the Fundacion BBVA (project “Pradia: la aventura gráfica de la pragmática y la prosodia” - CF613399). The activities of Down syndrome speech analysis continue (1/2018-12/2020) in the project funded by the Ministerio de Economía, Industria y Competitividad (MINECO) and the European Regional Development Fund FEDER (project “Incorporación de un Módulo de Predicción Automática de la Calidad de la Comunicación Oral de Personas con Síndrome de Down en un Videojuego Educativo” - TIN2017-88858-C2-1-R). The authors would like to thank all the participants who took part in the recording of the corpus. We would also like to thank Lourdes Aguilar, Valle Flores, Yolanda Martín and Ferran Adell. Special thanks to the students of the Fundación Personas (http://www.fundacionpersonas.org) for their motivation during the training sessions.

Appendix A. Description of the features

The tables included in this appendix describe the features used in each of the domains. Frequency features are presented in Table A.9. Energy features are described in Table A.10. Temporal features are explained in Table A.11. Spectral features are presented in Tables A.12 and A.13.

Table A.9
Frequency features explained. All functionals are applied to voiced regions only. Text in brackets shows the original name of the eGeMAPS features .

Feature	Description
F0_stddevRisingSlope (F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope)	Standard deviation of the slope of rising signal parts of F0
jitter_stddevNorm (jitterLocal_sma3nz_stddevNorm)	Coefficient of variation of the deviations in individual consecutive F0 period lengths
jitter_mean (jitterLocal_sma3nz_amean)	Mean of the deviations in individual consecutive F0 period lengths
F0_pctlrange (F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2)	Range of 20-th to 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile20 (F0semitoneFrom27.5Hz_sma3nz_percentile20.0)	Percentile 20-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile50 (F0semitoneFrom27.5Hz_sma3nz_percentile50.0)	Percentile 50-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_mean (F0semitoneFrom27.5Hz_sma3nz_amean)	Mean of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_stddevNorm (F0semitoneFrom27.5Hz_sma3nz_stddevNorm)	Coefficient of variation of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile80 (F0semitoneFrom27.5Hz_sma3nz_percentile80.0)	Percentile 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz

Table A.10
Energy features explained. All functionals are applied to voiced and unvoiced regions together. Text in brackets shows the original name of the eGeMAPS featurese.

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

2. COMPENDIO DE PUBLICACIONES

Table A.11
Temporal features explained.

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

Table A.12
Spectral features explained (part1). If nothing is said, the features are applied to voiced and unvoiced regions together. Text in brackets shows the original name of the eGeMAPS features .

Feature	Description
mfcc3_stddevNorm (mfcc3_sma3_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 3
slopeV0500_mean (slopeV0-500_sma3nz_amean)	Mean of linear regression slope of the logarithmic power spectrum within 0–500 Hz band in voiced regions
mfcc2V_mean (mfcc2V_sma3nz_amean)	Mean of Mel-Frequency Cepstral Coefficient 2 in voiced regions
mfcc4_stddevNorm (mfcc4_sma3_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 4
slopeUV0500_mean (slopeUV0-500_sma3nz_amean)	Mean of linear regression slope of the logarithmic power spectrum within 0–500 Hz band in unvoiced regions
slopeV0500_stddevNorm (slopeV0-500_sma3nz_stddevNorm)	Coefficient of variation of linear regression slope of the logarithmic power spectrum within 0–500 Hz band in voiced regions
mfcc2_stddevNorm (mfcc2_sma3_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 2
mfcc2_mean (mfcc2_sma3_amean)	Mean of Mel-Frequency Cepstral Coefficient 2
alphaRatioUV_mean (alphaRatioUV_sma3nz_amean)	Mean of the ratio of the summed energy from 50 to 1000 Hz and 1–5 kHz in unvoiced regions
logRelF0H1A3_stddevNorm (logRelF0-H1-A3_sma3nz_stddevNorm)	Coefficient of variation of the ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3) in voiced regions
hammarbergIndexUV_mean (hammarbergIndexUV_sma3nz_amean)	Mean of the ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region in unvoiced regions
mfcc3V_stddevNorm (mfcc3V_sma3nz_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 3 in voiced regions
mfcc4V_stddevNorm (mfcc4V_sma3nz_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 4 in voiced regions
mfcc4_mean (mfcc4_sma3_amean)	Mean of Mel-Frequency Cepstral Coefficient 4
spectralFlux_mean (spectralFlux_sma3nz_amean)	Mean of the difference of the spectra of two consecutive frames
spectralFluxUV_mean (spectralFluxUV_sma3nz_amean)	Mean of the difference of the spectra of two consecutive frames in unvoiced regions
spectralFluxV_mean (spectralFluxV_sma3nz_amean)	Mean of the difference of the spectra of two consecutive frames in voiced regions

Table A.13
Spectral features explained (part2). If nothing is said, the features are applied to voiced and unvoiced regions together. Text in brackets shows the original name of the eGeMAPS features .

Feature	Description
alphaRatioV_mean (alphaRatioV_sma3nz_amean)	Mean of the ratio of the summed energy from 50 to 1000 Hz and 1–5 kHz in voiced regions
mfcc4V_mean (mfcc4V_sma3nz_amean)	Mean of Mel-Frequency Cepstral Coefficient 4 in voiced regions
hammarbergIndexV_mean (hammarbergIndexV_sma3nz_amean)	Mean of the ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region in voiced regions
mfcc1V_mean (mfcc1V_sma3nz_amean)	Mean of Mel-Frequency Cepstral Coefficient 1 in voiced regions
hammarbergIndexV_stddevNorm (hammarbergIndexV_sma3nz_stddevNorm)	Coefficient of variation of the ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region in voiced regions
mfcc1_mean (mfcc1_sma3_amean)	Mean of Mel-Frequency Cepstral Coefficient 1
logRelF0H1A3_mean (logRelF0-H1-A3_sma3nz_amean)	Mean of the ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3)in voiced regions
F3amplitudeLogRelF0_mean (F3amplitudeLogRelF0_sma3nz_amean)	Mean of the ratio of the energy of the spectral harmonic peak at the third formant's centre frequency to the energy of the spectral peak at F0 in voiced regions
F3amplitudeLogRelF0_stddevNorm (F3amplitudeLogRelF0_sma3nz_stddevNorm)	Coefficient of variation of the ratio of the energy of the spectral harmonic peak at the third formant's centre frequency to the energy of the spectral peak at F0 in voiced regions
slopeV5001500_mean (slopeV500-1500_sma3nz_amean)	Mean of linear regression slope of the logarithmic power spectrum within 500–1500 Hz band in voiced regions
F2amplitudeLogRelF0_mean (F2amplitudeLogRelF0_sma3nz_amean)	Mean of the ratio of the energy of the spectral harmonic peak at the second formant's centre frequency to the energy of the spectral peak at F0 in voiced regions
F2amplitudeLogRelF0_stddevNorm (F2amplitudeLogRelF0_sma3nz_stddevNorm)	Coefficient of variation of the ratio of the energy of the spectral harmonic peak at the second formant's centre frequency to the energy of the spectral peak at F0 in voiced regions
F1bandwidth_stddevNorm (F1bandwidth_sma3nz_stddevNorm)	Coefficient of variation of the bandwidth of first formant in voiced regions
F1frequency_stddevNorm (F1frequency_sma3nz_stddevNorm)	Coefficient of variation of the centre frequency of first formant in voiced regions
F3frequency_stddevNorm (F3frequency_sma3nz_stddevNorm)	Coefficient of variation of the centre frequency of third formant in voiced regions
spectralFlux_stddevNorm (spectralFlux_sma3_stddevNorm)	Coefficient of variation of the difference of the spectra of two consecutive frames
F3frequency_mean (F3frequency_sma3nz_amean)	Mean of the centre frequency of third formant in voiced regions

2.2 Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome

References

Albertini, G., Bonassi, S., Dall'Armi, V., Giachetti, I., Giaquinto, S., Mignano, M., 2010. Spectral analysis of the voice in down syndrome. Res. Dev. Disabil. 31 (5), 995–1001.

Bhagyalakshmi, G., Renukarya, A., Rajangam, S., 2007. Metric analysis of the hard palate in children with Down syndrome-a comparative study. Down Syndrome Res. Pract. 12 (1), 55–59.

Boersma, P., 2006. Praat: doing phonetics by computer. http://www.praat.org/.

Bunton, K., Leddy, M., 2011. An evaluation of articulatory working space area in vowel production of adults with down syndrome. Clinical Ling. Phonetics 25 (4), 321–334.

Chapman, R., Hesketh, L., 2001. Language, cognition, and short-term memory in individuals with Down syndrome. Down Syndrome Res. Pract. 7 (1), 1–7.

Chapman, R.S., 1997. Language development in children and adolescents with Down syndrome. Ment. Retard. Dev. Disabil. Res. Rev. 3 (4), 307–312.

Cleland, J., Wood, S., Hardcastle, W., Wishart, J., Timmins, C., 2010. Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome. Int. J. Lang. Commun. Disord. 45 (1), 83–95.

Corrales-Astorgano, M., Escudero-Mancebo, D., González-Ferreras, C., 2016. Acoustic analysis of anomalous use of prosodic features in a corpus of people with intellectual disability. Advances in Speech and Language Technologies for Iberian Languages: Third International Conference IberSPEECH. Springer, pp. 151–161.

Devenny, D., Silverman, W., 1990. Speech dysfluency and manual specialization in Down's syndrome. J. Intellect. Disabil. Res. 34 (3), 253–260.

Dibazar, A.A., Berger, T.W., Narayanan, S.S., 2006. Pathological voice assessment. Engineering in Medicine and Biology Society (EMBS). IEEE, pp. 1669–1673.

Dusan, S., Deng, L., 1998. Recovering vocal tract shapes from mfcc parameters. ICSLP.

Eggers, K., Van Eerdenbrugh, S., 2017. Speech disfluencies in children with Down syndrome. J. Commun. Disord.

Escudero, D., González, C., Gutiérrez, Y., Rodero, E., 2017. Identifying characteristic prosodic patterns through the analysis of the information of sp_tobi label sequences. Comput. Speech Lang. 45, 39–57.

Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., et al., 2016. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7 (2), 190–202.

Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in opensmile, the Munich open-source multimedia feature extractor. Proceedings of the 21st ACM international conference on Multimedia. ACM, pp. 835–838.

Gauffin, J., Sundberg, J., 1977. Clinical applications of acoustic voice analysis. part II: acoustical analysis, results, and discussion. Speech Transmission Laboratory, Quarterly Progress and Status Report.

González-Ferreras, C., Escudero-Mancebo, D., Corrales-Astorgano, M., Aguilar-Cuevas, L., Flores-Lucas, V., 2017. Engaging adolescents with Down syndrome in an educational video game. Int. J. Human–Comput. Interact. 1–20.

Guimaraes, C.V., Donnelly, L.F., Shott, S.R., Amin, R.S., Kalra, M., 2008. Relative rather than absolute macroglossia in patients with Down syndrome: implications for treatment of obstructive sleep apnea. Pediatr. Radiol. 38 (10), 1062.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The weka data mining software: an update. ACM SIGKDD Explorations Newslett. 11 (1), 10–18.

Hillenbrand, J., Houde, R.A., 1996. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. J. Speech Lang. Hearing Res. 39 (2), 311–321.

House, A.S., Stevens, K.N., 1956. Analog studies of the nasalization of vowels. J. Speech Hearing Disord. 21 (2), 218–232.

Huffman, M.K., 1989. Implementation of nasal: timing and articulatory landmarks. University of California, Los Angeles.

Kent, R.D., Vorperian, H.K., 2013. Speech impairment in Down syndrome: a review. J. Speech Lang. Hearing Res. 56 (1), 178–210.

Kisler, T., Reichel, U., Schiel, F., 2017. Multilingual processing of speech via web services. Comput. Speech Lang. 45, 326–347.

Lee, M.T., Thorpe, J., Verhoeven, J., 2009. Intonation and phonation in young adults with Down syndrome. J. Voice 23 (1), 82–87.

Leino, T., 2009. Long-term average spectrum in screening of voice quality in speech: untrained male university students. J. Voice 23 (6), 671–676.

Leshin, L., 2000. Plastic Surgery in Children with Down Syndrome. Down syndrome: Health issues: News and information for parents and professionals.

Luo, D., Luo, R., Wang, L., 2017. Prosody analysis of L2 English for naturalness evaluation through speech modification. Proc. Interspeech. pp. 1775–1778.

Markaki, M., Stylianou, Y., 2010. Modulation spectral features for objective voice quality assessment. Communications, Control and Signal Processing (ISCCSP), 2010 4th International Symposium on. IEEE, pp. 1–4.

Markaki, M., Stylianou, Y., 2011. Voice pathology detection and discrimination based on modulation spectral features. IEEE Trans. Audio Speech Lang. Process. 19 (7), 1938–1948.

Martin, G.E., Klusek, J., Estigarribia, B., Roberts, J.E., 2009. Language characteristics of individuals with Down syndrome. Top. Lang. Disord. 29 (2), 112.

Martinez, J., Perez, H., Escamilla, E., Suzuki, M.M., 2012. Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques. Electrical Communications and Computers (CONIELECOMP). IEEE, pp. 248–251.

Martínez, M.H., Duran, X.P., Navarro, J.N., 2011. Attention deficit disorder with or without hyperactivity or impulsivity in children with Down's syndrome. Int. Med. Rev. Down Syndrome 15 (2), 18–22.

Moran, M.J., 1986. Identification of Down's syndrome adults from prolonged vowel samples. J. Commun. Disord. 19 (5), 387–394.

Moran, M.J., Gilbert, H.R., 1982. Selected acoustic characteristics and listener judgments of the voice of Down syndrome adults. Am. J. Ment. Defic.

Moura, C.P., Cunha, L.M., Vilarinho, H., Cunha, M.J., Freitas, D., Palha, M., Pueschel, S.M., Pais-Clemente, M., 2008. Voice parameters in children with Down syndrome. J. Voice 22 (1), 34–42.

Pentz Jr, A.L., 1987. Formant amplitude of children with Down syndrome. Am. J. Ment. Defic. 92 (2), 230–233.

Rochet-Capellan, A., Dohen, M., 2015. Acoustic characterisation of vowel production by young adults with Down syndrome. 18th International Congress of Phonetic Sciences (ICPhS 2015).

Rodger, R., 2009. Voice quality of children and young people with Down's Syndrome and its impact on listener judgement. Queen Margaret University.

Saz, O., Simón, J., Rodríguez, W., Lleida, E., Vaquero, C., et al., 2009. Analysis of acoustic features in speakers with cognitive disorders and speech impairments. EURASIP J. Adv. Signal Process. 2009, 1.

Saz, O., Yin, S.C., Lleida, E., Rose, R., Vaquero, C., Rodríguez, W.R., 2009. Tools and technologies for computer-aided speech and language therapy. Speech Communication 51 (10), 948–967.

Schiel, F., 1999. Automatic phonetic transcription of non-prompted speech. International Congress of Phonetic Sciences (ICPhS). pp. 607–610.

Schuller, B.W., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J.K., Baird, A., Elkins, A.C., Zhang, Y., Coutinho, E., Evanini, K., 2016. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. INTERSPEECH. pp. 2001–2005.

Seifpanahi, S., Bakhtiar, M., Salmalian, T., 2011. Objective vocal parameters in farsi-speaking adults with down syndrome. Folia Phoniatrica et Logopaedica 63 (2), 72–76.

Shott, S.R., Joseph, A., Heithaus, D., 2001. Hearing loss in children with Down syndrome. Int. J. Pediatr. Otorhinolaryngol. 61 (3), 199–205.

Van Borsel, J., Vandermeulen, A., 2008. Cluttering in Down syndrome. Folia Phoniatrica et Logopaedica 60 (6), 312–317.

Wold DC, M.J., 1979. Preliminary perceived voice deviations and hearing disorders of adults with Down's syndrome. Percept. Mot. Skills 49, 564-564.

Wuang, Y.-P., Chiang, C.-S., Su, C.-Y., Wang, C.-C., 2011. Effectiveness of virtual reality using Wii gaming technology in children with Down syndrome. Res. Dev. Disabil. 32 (1), 312–321.

Yuan, J., Liberman, M., 2011. Automatic measurement and comparison of vowel nasalization across languages. Proceedings of ICPhS.

Zampini, L., Fasolo, M., Spinelli, M., Zanchi, P., Suttora, C., Salerni, N., 2016. Prosodic skills in children with down syndrome and in typically developing children. Int. J. Lang. Commun. Disord. 51 (1), 74–83.

2.3. Automatic Assessment of Prosodic Quality in Down Syndrome: Analysis of the Impact of Speaker Heterogeneity

M. Corrales-Astorgano[1], P. Martínez-Castilla[3], D. Escudero-Mancebo[1], L. Aguilar[2], C. González-Ferreras[1], and V. Cardeñoso-Payo[1]. "Automatic assessment of prosodic quality in Down syndrome: Analysis of the impact of speaker heterogeneity". Applied Sciences, vol. 9, no. 7, p. 1440, 2019.

[1]*Departamento de Informática, Universidad de Valladolid, Valladolid, Spain*

[2]*Departamento de Filología Española, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain*

[3]*Departamento de Psicología Evolutiva y de la Educación, UNED, 28040 Madrid, Spain*

https://doi.org/10.3390/app9071440

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

applied sciences

MDPI

Article

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

Mario Corrales-Astorgano [1,*], Pastora Martínez-Castilla [2,*], David Escudero-Mancebo [1,*], Lourdes Aguilar [3,*], César González-Ferreras [1,*] and Valentín Cardeñoso-Payo [1,*]

1 Department of Computer Science, University of Valladolid, 47002 Valladolid, Spain
2 Department of Developmental and Educational Psychology, UNED, 28040 Madrid, Spain
3 Department of Hispanic Philology, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
* Correspondence: mcorrales@infor.uva.es (M.C.-A.); pastora.martinez@psi.uned.es (P.M.-C.); descuder@infor.uva.es (D.E.-M.); lourdes.aguilar@uab.cat (L.A.); cesargf@infor.uva.es (C.G.-F.); valen@infor.uva.es (V.C.-P.)
† This paper is an extended version of our paper published in the conference IberSPEECH2018.

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

check for updates

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

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

1. Introduction

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

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

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

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

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

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

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

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

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

2. Methodology

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

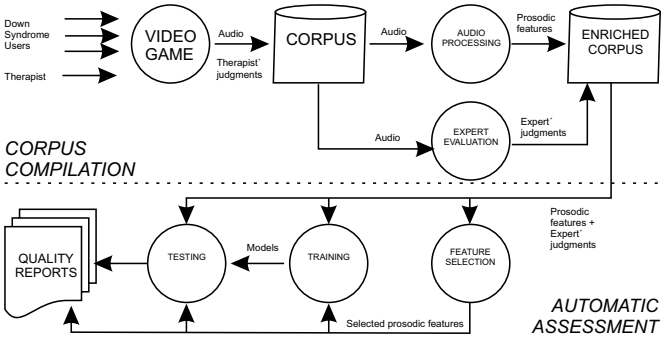

Figure 1. Experimental procedure scheme.

2.1. Game Description

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

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

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

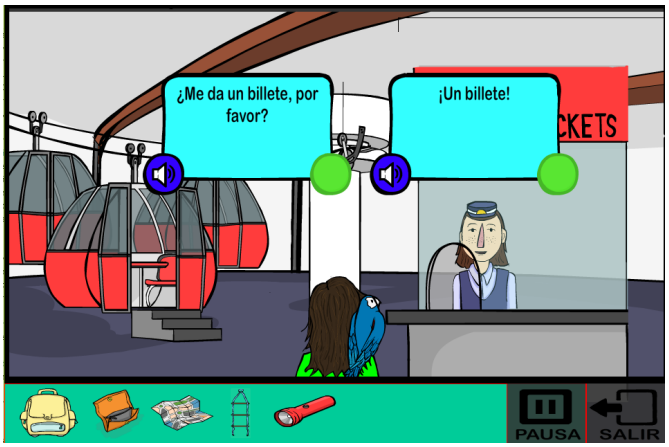

Figure 2. Example of a type of perception activity in the video game: the player must choose which of the two productions (Could I have a ticket, please? or A ticket!) is pragmatically adequate according to the communicative situation.

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

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

2.2. Corpus Description

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

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

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

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

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

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

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

2.3. Corpus Evaluation

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

2.3.1. Evaluation Criteria

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

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

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

2.3.2. Therapist Evaluations

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

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

2.3.3. Expert Judgments

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

2.4. Feature Extraction and Selection

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

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

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

2.5. Automatic Classification

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

3. Results

3.1. Classification Results

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

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

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

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

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

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

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

3.2. Speakers Variability Results

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

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

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

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

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

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

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

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

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

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

4. Discussion

4.1. Analysis of the Classification Results

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

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

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

4.2. Impact of Variability on Assessment

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

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

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

The high variability of the speech of individuals with Down syndrome has also been shown in our experimental results regarding the use of prosodic features. Table 4 shows the differences in the correlated features with the expert evaluation per speaker. Some rhythm and frequency features appear above in the ranking of all the speakers (from *silencesMean* to *F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope* on Table 4). However, intensity features seem to be very important to the S03 and S05 speakers, but present lower importance to the others. Speaker S04 presents higher ROC area values than the other speakers in his features, so the lower importance of these features to the other speakers can affect the performance of the automatic classifiers. This heterogeneity complicates the automatic assessment of prosody quality, because automatic classifiers show poorer generalization power.

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

4.3. Limitations and Future Work

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

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

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

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

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

5. Conclusions

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

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

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

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

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

Appendix A. Description of the Features

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

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

Feature	Description
F0_mean (F0semitoneFrom27.5Hz_sma3nz_amean)	Mean of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_stddevNorm (F0semitoneFrom27.5Hz_sma3nz_stddevNorm)	Coefficient of variation of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile20 (F0semitoneFrom27.5Hz_sma3nz_percentile20.0)	Percentile 20-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile50 (F0semitoneFrom27.5Hz_sma3nz_percentile50.0)	Percentile 50-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile80 (F0semitoneFrom27.5Hz_sma3nz_percentile80.0)	Percentile 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_pctlrange (F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2)	Range of 20-th to 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_meanRisingSlope (F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope)	Mean of the slope of rising signal parts of F0
F0_stddevRisingSlope (F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope)	Standard deviation of the slope of rising signal parts of F0
F0_meanFallingSlope (F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope)	Mean of the slope of falling signal parts of F0
F0_stddevFallingSlope (F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope)	Standard deviation of the slope of falling signal parts of F0
jitter_mean (jitterLocal_sma3nz_amean)	Mean of the deviations in individual consecutive F0 period lengths
jitter_stddevNorm (jitterLocal_sma3nz_stddevNorm)	Coefficient of variation of the deviations in individual consecutive F0 period lengths

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

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

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

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

References

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

4. Grieco, J.; Pulsifer, M.; Seligsohn, K.; Skotko, B.; Schwartz, A. Down syndrome: Cognitive and behavioral functioning across the lifespan. *Am. J. Med. Genet. Part C Semin. Med. Genet.* **2015**, *169*, 135–149. [CrossRef] [PubMed]
5. Martin, G.E.; Klusek, J.; Estigarribia, B.; Roberts, J.E. Language characteristics of individuals with Down syndrome. *Top. Lang. Disord.* **2009**, *29*, 112. [CrossRef]
6. Eadie, P.A.; Fey, M.; Douglas, J.; Parsons, C. Profiles of grammatical morphology and sentence imitation in children with specific language impairment and Down syndrome. *J. Speech Lang. Hear. Res.* **2002**, *45*, 720–732. [CrossRef]
7. Smith, E.; Næss, K.A.B.; Jarrold, C. Assessing pragmatic communication in children with Down syndrome. *J. Commun. Disord.* **2017**, *68*, 10–23. [CrossRef]
8. Laws, G.; Bishop, D.V. Verbal deficits in Down's syndrome and specific language impairment: A comparison. *Int. J. Lang. Commun. Disord.* **2004**, *39*, 423–451. [CrossRef]
9. Kent, R.D.; Vorperian, H.K. Speech impairment in Down syndrome: A review. *J. Speech Lang. Hear. Res.* **2013**, *56*, 178–210. [CrossRef]
10. Heselwood, B.; Bray, M.; Crookston, I. Juncture, rhythm and planning in the speech of an adult with Down's syndrome. *Clin. Linguist. Phon.* **1995**, *9*, 121–137. [CrossRef]
11. Peppé, S.J. Why is prosody in speech-language pathology so difficult? *Int. J. Speech-Lang. Pathol.* **2009**, *11*, 258–271. [CrossRef]
12. Martínez-Castilla, P.; Sotillo, M.; Campos, R. Prosodic abilities of Spanish-speaking adolescents and adults with Williams syndrome. *Lang. Cogn. Process.* **2011**, *26*, 1055–1082. [CrossRef]
13. Peppé, S.; McCann, J.; Gibbon, F.; O'Hare, A.; Rutherford, M. Receptive and expressive prosodic ability in children with high-functioning autism. *J. Speech Lang. Hear. Res.* **2007**, *50*, 1015–1028. [CrossRef]
14. Peppé, S.; McCann, J. Assessing intonation and prosody in children with atypical language development: The PEPS-C test and the revised version. *Clin. Linguist. Phon.* **2003**, *17*, 345–354. [CrossRef]
15. Stojanovik, V. Prosodic deficits in children with Down syndrome. *J. Neurolinguist.* **2011**, *24*, 145–155. [CrossRef]
16. Saz, O.; Yin, S.C.; Lleida, E.; Rose, R.; Vaquero, C.; Rodríguez, W.R. Tools and technologies for computer-aided speech and language therapy. *Speech Commun.* **2009**, *51*, 948–967. [CrossRef]
17. Rodríguez, W.R.; Saz, O.; Lleida, E. A prelingual tool for the education of altered voices. *Speech Commun.* **2012**, *54*, 583–600. [CrossRef]
18. Shahin, M.; Ahmed, B.; Parnandi, A.; Karappa, V.; McKechnie, J.; Ballard, K.J.; Gutierrez-Osuna, R. Tabby Talks: An automated tool for the assessment of childhood apraxia of speech. *Speech Commun.* **2015**, *70*, 49–64. [CrossRef]
19. Öster, A.M.; House, D.; Protopapas, A.; Hatzis, A. Presentation of a new EU project for speech therapy: OLP (Ortho-Logo-Paedia). In Proceedings of the XV Swedish Phonetics Conference (Fonetik 2002), Stockholm, Sweden, 29–31 May 2002; pp. 29–31.
20. Tan, T.S.; Ariff, A.; Ting, C.M.; Salleh, S.H. Application of Malay speech technology in Malay speech therapy assistance tools. In Proceedings of the IEEE 2007 International Conference on Intelligent and Advanced Systems, Kuala Lumpur, Malaysia, 25–28 November 2007; pp. 330–334.
21. PRADIA, misterio en la ciudad. Available online: http://www.pradia.net (accessed on 18 July 2018).
22. Adell, F.; Aguilar, L.; Corrales-Astorgano, M.; Escudero-Mancebo, D. Proceso de innovación educativa en educación especial: Enseñanza de la prosodia con fines comunicativos con el apoyo de un videojuego educativo. In Proceedings of the I Congreso Internacional en Humanidades Digitales, Valladolid, Spain, 17–19 April 2018.
23. Le, D.; Licata, K.; Persad, C.; Provost, E.M. Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2187–2199. [CrossRef]
24. Qin, Y.; Lee, T.; Feng, S.; Kong, A.P.H. Automatic Speech Assessment for People with Aphasia Using TDNN-BLSTM with Multi-Task Learning. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3418–3422.
25. Maier, A.; Haderlein, T.; Eysholdt, U.; Rosanowski, F.; Batliner, A.; Schuster, M.; Nöth, E. PEAKS–A system for the automatic evaluation of voice and speech disorders. *Speech Commun.* **2009**, *51*, 425–437. [CrossRef]
26. Kim, J.; Kumar, N.; Tsiartas, A.; Li, M.; Narayanan, S.S. Automatic intelligibility classification of sentence-level pathological speech. *Comput. Speech Lang.* **2015**, *29*, 132–144. [CrossRef]

27. Maier, A.; Hönig, F.; Hacker, C.; Schuster, M.; Nöth, E. Automatic evaluation of characteristic speech disorders in children with cleft lip and palate. In Proceedings of the Ninth Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008.
28. Laaridh, I.; Kheder, W.B.; Fredouille, C.; Meunier, C. Automatic prediction of speech evaluation metrics for dysarthric speech. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1834–1838.
29. Martínez, D.; Lleida, E.; Green, P.; Christensen, H.; Ortega, A.; Miguel, A. Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Trans. Access. Comput. (TACCESS)* **2015**, *6*, 10. [CrossRef]
30. Lee, H.Y.; Hu, T.Y.; Jing, H.; Chang, Y.F.; Tsao, Y.; Kao, Y.C.; Pao, T.L. Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 215–219.
31. Dunn, L.; Dunn, L.; Arribas, D. *Test de vocabulario en imágenes Peabody*; TEA: Madrid, Spain, 2006.
32. Corral, S.; Arribas, D.; Santamaría, P.; Sueiro, M.; Pereña, J. *Escala de Inteligencia de Wechsler para niños-IV*; TEA Ediciones: Madrid, Spain, 2005.
33. Raven, J.; Raven, J.C.; Court, J. *Test de matrices progresivas: Manual/Manual for Raven's progessive matrices and vocabulary scalesTest de matrices progresivas*; Number 159.9. 072; J C Raven Ltd.: Buenos Aires, Argentina, 1993.
34. Martínez-Castilla, P.; Peppé, S. Developing a test of prosodic ability for speakers of Iberian Spanish. *Speech Commun.* **2008**, *50*, 900–915. [CrossRef]
35. González-Ferreras, C.; Escudero-Mancebo, D.; Corrales-Astorgano, M.; Aguilar-Cuevas, L.; Flores-Lucas, V. Engaging adolescents with Down syndrome in an educational video game. *Int. J. Human–Comput. Interact.* **2017**, *33*, 693–712. [CrossRef]
36. Ladd, D.R. *Intonational Phonology*; Cambridge University Press: Cambridge, UK, 2008.
37. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B. Recent developments in opensmile, the Munich open-source multimedia feature extractor. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; ACM: New York, NY, USA, 2013; pp. 835–838.
38. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
39. Corrales-Astorgano, M.; Escudero-Mancebo, D.; González-Ferreras, C. Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome. *Speech Commun.* **2018**, *99*, 90–100. [CrossRef]
40. Kuhn, M.Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
41. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* **1997**, *30*, 1145–1159. [CrossRef]
42. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newslett.* **2009**, *11*, 10–18. [CrossRef]
43. Pardo, A.; Ruiz, M.Á. *SPSS 11: Guia para el analisis de datos.*; Mc Graw Hill: Madrid, Spain, 2002.
44. Le, D.; Provost, E.M. Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
45. Tu, M.; Berisha, V.; Liss, J. Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1849–1853.
46. Kreiman, J.; Gerratt, B.R.; Precoda, K.; Berke, G.S. Individual differences in voice quality perception. *J. Speech Lang. Hear. Res.* **1992**, *35*, 512–520. [CrossRef]
47. Li, M.; Tang, D.; Zeng, J.; Zhou, T.; Zhu, H.; Chen, B.; Zou, X. An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. *Comput. Speech Lang.* **2019**, *56*, 80–94. [CrossRef]

Capítulo 3

Discusión de resultados

En este capítulo se realiza un resumen de los principales resultados obtenidos en la investigación realizada en esta tesis y una discusión sobre los mismos. Una descripción más detallada de los resultados puede obtenerse en los artículos que forman el compendio de artículos de este trabajo de tesis (ver capítulo 2). La sección 3.1 describe los resultados relacionados con la evaluación del diseño del videojuego y de la interacción de los jugadores con el mismo, así como una discusión sobre los resultados obtenidos y sus implicaciones. La sección 3.2 detalla los resultados del análisis realizado de las características acústicas extraídas de las grabaciones del videojuego y la comparación realizada entre las grabaciones de personas con síndrome de Down y las grabaciones de personas con desarrollo típico. Por último, la sección 3.3 describe los resultados obtenidos de la evaluación automática de la calidad prosódica de las grabaciones del videojuego y cómo afecta la heterogeneidad de las personas con síndrome de Down a dichos resultados.

3.1. Evaluación del diseño e interacción

Con el objetivo de evaluar si las decisiones de diseño y de interacción del videojuego han servido para motivar a los jugadores con síndrome de Down a usar el mismo (objetivo O1.1) y de analizar la adecuación de las actividades y el diseño del videojuego con respecto a las limitaciones de las personas con síndrome de Down (objetivo O1.2), se han extraído una serie de estadísticas de la información almacenada automáticamente por el videojuego. Esta información se extrae de sesiones de juego realizadas por niños, adultos y personas con síndrome de Down, centrándose en las actividades principales del videojuego, las actividades de comprensión y producción prosódica. Además, se analizan las respuestas a los cuestionarios realizados a los participantes después de las sesiones de juego, con el objetivo de comprobar si el videojuego es capaz de motivar a los jugadores a realizar las actividades de entrenamiento propuestas.

La tabla 3.1 muestra los resultados obtenidos en las actividades de comprensión por los tres grupos que realizaron las sesiones de juego. Estos datos se calculan utilizando

Tabla 3.1: Resultados de las actividades de comprensión: tiempo necesario para completar la actividad en segundos, número de errores, número de escuchas adicionales de la frase, número de ayudas sonoras y número de clicks incorrectos (media y desviación típica para cada tipo de usuario en 5 actividades de comprensión).

	Tiempo		Errores		Repeticiones de frase		Ayudas sonoras		Clicks incorrectos	
	Media	DT	Media	DT	Media	DT	Media	DT	Media	DT
Niños	118,8	17,2	0,1	0,3	1,2	1,3	0,0	0,0	1,6	1,5
Adultos	105,9	7,4	0,1	0,3	0,3	0,5	0,0	0,0	0,0	0,0
Síndrome de Down	166,0	43,3	1,0	0,9	1,6	1,7	0,6	1,0	0,1	0,5

la información que el videojuego almacena automáticamente durante el transcurso del mismo. El test no paramétrico Kruskal-Wallis muestra diferencias significativas entre los tres grupos en el tiempo requerido para completar la actividad ($p < 0,001$), en el número de errores ($p = 0,003$) y en el número de clicks incorrectos ($p < 0,001$). Además, utilizando el test no paramétrico Mann-Whitney, se observan diferencias significativas entre las personas con síndrome de Down y los adultos en el tiempo necesario para completar la actividad (166,0 vs 105,9 con $p < 0,001$) y en el número de errores (1,0 vs 0,1 con p=0,007). Lo mismo ocurre al comparar personas con síndrome de Down con niños (166,0 vs 118,8 con p=0,007 en tiempo necesario para completar la actividad y 1,0 vs 0,1 con p=0,007 en número de errores). Tanto el tiempo necesario para completar la actividad como el número de errores es más alto en el grupo de personas con síndrome de Down en comparación con los otros grupos. Además, la desviación típica en estas mismas variables es alta en el caso de las personas con síndrome de Down.

La tabla 3.2 muestra los resultados obtenidos en las actividades de producción por los tres grupos que realizaron las sesiones de juego. Existen diferencias significativas en el tiempo empleado ($p < 0,001$), en el número de errores ($p < 0,001$), en las repeticiones de la frase ($p < 0,001$) y en el número de clicks incorrectos ($p = 0,027$) entre los tres grupos usando de nuevo el test Kruskal-Wallis. El tiempo necesario para completar la actividad es significativamente mayor en las personas con síndrome de Down comparando con adultos (335,9 vs 173,1 con $p < 0,001$) y niños (335,9 vs 189,5 con $p < 0,001$). El número de errores también es significativamente mayor en las

Tabla 3.2: Resultados de las actividades de producción: tiempo necesario para completar la actividad en segundos, número de errores, número de escuchas adicionales de la frase, número de ayudas sonoras y número de clicks incorrectos (media y desviación típica para cada tipo de usuario en 7 actividades de comprensión).

	Tiempo		Errores		Repeticiones de frase		Ayudas sonoras		Clicks incorrectos	
	Media	DT	Media	DT	Media	DT	Media	DT	Media	DT
Niños	189,5	31,3	0,3	0,7	0,6	1,1	0,1	0,3	6,2	6,0
Adultos	173,1	6,5	0,1	0,3	0,0	0,0	0,0	0,0	1,0	0,9
Síndrome de Down	335,9	102,6	4,4	2,3	3,7	3,7	1,6	2,8	2,2	3,5

personas con síndrome de Down con respecto a los adultos (4,4 vs 0,1 con $p < 0,001$) y niños (4,4 vs 0,3 con $p < 0,001$). Por último, el número de repeticiones de la frase a producir es significativamente mayor en personas con síndrome de Down con respecto a los adultos (3,7 vs 0,0 con $p < 0,001$) y a los niños (3,7 vs 0,6 con p=0,012).

Los resultados presentados en las tablas 3.1 y 3.2 muestran que los usuarios con síndrome de Down cometen más errores y necesitan más tiempo para completar dichas actividades que los usuarios con desarrollo típico. Estos resultados muestran las dificultades de las personas con síndrome de Down a la hora de completar tanto las actividades de comprensión como las actividades de producción. Estos datos pueden ser útiles para detectar los problemas específicos de cada jugador relacionados con el uso de la prosodia, de manera que el terapeuta pueda enfocarse en dichos problemas y entrenar las habilidades relacionadas con cada actividad con el objetivo de mejorarlas.

La tabla 3.3 muestra las respuestas a los cuestionarios que se realizaron después de la sesión de juego. Este cuestionario sólo se realizó a los niños y a las personas con síndrome de Down. En general, a las personas con síndrome de Down que jugaron al videojuego les gustó el mismo y querrían volver a jugar en el futuro. El acompañante elegido parece adecuado y los gráficos introducidos son satisfactorios. Existe menos unanimidad cuando se trata de entender la historia contada en el videojuego, ya que las limitaciones cognitivas de estas personas dificultan retener detalles concretos de la historia.

Los resultados relacionados con la evaluación del videojuego muestran una alta mo-

Tabla 3.3: Resultados del cuestionario realizado a los jugadores con síndrome de Down y a los niños (posibles respuestas sí/no).

	Niños		Síndrome de Down	
	Sí	No	Sí	No
Al usuario le gustan los videojuegos	100,0 %	0,0 %	92,9 %	7,1 %
El usuario suele jugar a videojuegos	80,0 %	20,0 %	78,6 %	21,4 %
El usuario disfrutó jugando	90,0 %	10,0 %	92,9 %	7,1 %
El usuario quiere volver a jugar en el futuro / el usuario quiere saber como continua la historia	90,0 %	10,0 %	85,7 %	14,3 %
El usuario entiende la historia contada en el videojuego	100,0 %	0,0 %	64,3 %	35,7 %
Al usuario le gusta la historia contada en el videojuego	90,0 %	10,0 %	78,6 %	21,4 %
Al usuario le gusta que le acompañe un loro durante el juego	90,0 %	10,0 %	85,7 %	14,3 %
Al usuario le gusta comunicarse con los personas que aparecen en el juego	88,9 %	11,1 %	92,9 %	7,1 %
Al usuario le gustaría crearse su propio avatar (cara, ojos, boca, ropa, etc)	88,9 %	11,1 %	64,3 %	35,7 %
El usuario manifiesta haber aprendido algo durante el videojuego	70,0 %	30,0 %	78,6 %	21,4 %
Al usuario le gusta las imágenes del juego	100,0 %	0,0 %	85,7 %	14,3 %

tivación por parte de los usuarios con síndrome de Down para completar las actividades del videojuego. Las personas con síndrome de Down tienen ciertas limitaciones cognitivas que les producen dificultades para realizar tareas que requieran gran concentración, como las tareas de entrenamiento. La inclusión de las actividades dentro de una narrativa es una de las principales claves para conseguir esta motivación. A pesar de que algunos usuarios indicaron que no entendían ciertas partes de la historia propuesta, el objetivo del videojuego es que realicen las actividades de entrenamiento, por lo que esta falta de entendimiento no es problemática. Los comentarios obtenidos por parte de las terapeutas de los centros de educación especial mostraron que, en general, los jugadores con síndrome de Down estuvieron concentrados durante toda la sesión de juego y que estaban motivados a realizar las actividades de entrenamiento que se les proponían.

Además, la herramienta ha servido también para recopilar un corpus de voz centrado en la prosodia de personas con síndrome de Down. Este resultado es importante ya que no existen apenas corpus de este tipo, principalmente debido a las limitaciones cognitivas de estas personas que dificultan la realización de tareas que requiere

concentración durante un largo periodo de tiempo. La motivación que el videojuego proporciona para realizar las actividades de entrenamiento también es útil para realizar la grabación del corpus de voz que permite realizar los experimentos detallados en los siguientes apartados.

3.2. Análisis de las características acústicas

Uno de los objetivos de esta tesis es caracterizar la voz de las personas con síndrome de Down en relación al uso de la prosodia comparando sus voces con las de personas con desarrollo típico (objetivo O2.1). Para realizar este análisis se han utilizado las muestras de voz recopiladas en un corpus mediante el uso del videojuego, que contiene grabaciones de 18 personas con síndrome de Down y un total de 349 grabaciones. Para poder realizar la comparativa, se han utilizado 250 grabaciones procedentes de 22 personas con desarrollo típico. El objetivo es extraer una serie de características acústicas de las grabaciones de ambos grupos y compararlas utilizando métodos estadísticos y de aprendizaje automático.

La tabla 3.4 muestra los resultados de clasificación de la tarea de clasificar una grabación como proveniente de una persona con síndrome de Down o de una persona con desarrollo típico. Se han utilizado tres clasificadores distintos (SVM, MLP y DT) para comprobar con cual se obtienen mejores resultados. Sólo se han utilizado las características en las que se han encontrado diferencias significativas entre las grabaciones de personas con síndrome de Down y personas con desarrollo típico, utilizando el test no paramétrico Mann-Whitney para realizar esta comparación. Con estas características se han formado varios subconjuntos dependiendo del tipo de característica extraída (frecuencia, energía, temporal o espectral), así como combinaciones de características de los diferentes tipos. El DT muestra los peores resultados de clasificación para todos los subconjuntos. MLP obtiene los mejores resultados en los subconjuntos de frecuencia (UAR 0,64), temporal (UAR 0,78), la unión de frecuencia, energía y temporal (UAR 0,91) y con todas (UAR 0,95). El SVM obtiene mejores resultados con el subconjunto de

Tabla 3.4: Resultados de clasificación para la tarea de identificar el grupo al que pertenece un grabación (síndrome de Down o sin discapacidad). Para cada subconjunto y clasificador, se muestra la tasa de clasificación (TC) y el UAR. Las características utilizadas en cada subconjunto son las que muestran diferencias significativas entre los grupos de hablantes. Los clasificadores son árboles de decisión (DT), máquinas de vectores de soporte (SVM) y perceptrón multicapa (MLP). # es el número de características utilizadas en cada subconjunto.

		SVM		MLP		DT	
Subconjunto	**#**	**TC**	**UAR**	**TC**	**UAR**	**TC**	**UAR**
Frecuencia	9	62,67	0,61	64,33	0,64	60,17	0,60
Energía	9	79,33	0,78	76	0,76	72,5	0,71
Temporal	9	76,83	0,76	77,83	0,78	74,33	0,75
Frequencia+Energía+Temporal	27	90	0,9	91,83	0,91	82	0,82
Espectral	34	87,33	0,87	87,33	0,87	84,33	0,84
Todas	61	94,17	0,94	95,17	0,95	86,5	0,87

energía (UAR 0,78). Los resultados utilizando el subconjunto de espectrales son iguales al utilizar MLP y SVM (UAR 0,87). Además, los mejores resultados se obtienen cuando se utilizan todas las características juntas, independientemente del clasificador utilizado. El subconjunto de frecuencia obtiene los peores resultados con cualquiera de los tres clasificadores. Cuando se unen los subconjuntos de frecuencia, energía y temporales, los resultados son mejores que los obtenidos con cada subconjunto por separado.

Para analizar estos resultados, tomamos como referencia de un buen uso de la prosodia las grabaciones de personas con desarrollo típico. Las características acústicas relacionas con la frecuencia fundamental, con la energía y con el ritmo seleccionadas por separado muestran resultados más bajos que las espectrales, pero la combinación de los tres tipos mencionados obtiene mejores resultados que las espectrales solas. Las características relacionadas con el dominio espectral suelen estar relacionadas con particularidades específicas del sistema fonológico de las personas con síndrome de Down, por lo que es complicado o imposible mejorarlas con entrenamiento. Sin embargo, las características relacionadas con los dominios de la frecuencia, la energía y temporal pueden mejorar mediante el entrenamiento, como por ejemplo practicando la respiración para controlar el ritmo en el habla. La posible mejora en estas características se puede ver reflejada en un mejor uso de la entonación, el tono, el acento o el ritmo en

Tabla 3.5: Número de respuestas al test de percepción para cada tipo de audio con la prosodia transferida. El 1 significa poca seguridad y el 5 mucha seguridad en la identificación del audio como proveniente de una persona con síndrome de Down. NR significa No sabe/No contesta. TDutt+TDpro es una grabacion proveniente de una persona con desarrollo típico con la prosodia transferida de otra grabacion de persona con desarrollo típico. DSutt+TDpro es una grabación proveniente de una persona con síndrome de Down con la prosodia transferida de una persona con desarrollo típico. TDutt+DSpro es una grabación proveniente de una persona con desarrollo típico con la prosodia transferida de una persona con síndrome de Down. DSutt+DSpro es una grabación proveniente de una persona con síndrome de Down con la prosodia transferida de una persona con síndrome de Down.

Tipo	**1**	**2**	**3**	**4**	**5**	**NR**	**Total**
TDutt+TDpro	124	15	3	1	4	3	150
DSutt+TDpro	42	42	31	18	11	6	150
TDutt+DSpro	17	21	34	43	31	4	150
DSutt+DSpro	1	11	26	49	56	7	150

el habla de las personas con síndrome de Down.

La tabla 3.5 muestra el número de respuestas al test de percepción para cada tipo de audio con la prosodia transferida. Los dos grupos de hablantes existentes son las personas con síndrome de Down (SD) y las personas con desarrollo típico (TD). Cuando la prosodia de una grabación proveniente de una persona TD es transferida a una grabación proveniente de una persona TD, el 84 % de las respuestas de los evaluadores identifican la grabación como proveniente de una persona TD (respuesta 1 de TDutt+TDpro). En este caso, no existen apenas respuestas que indiquen duda en esta afirmación (2 % de las respuestas, respuesta 3 de TDutt+TDpro). Cuando la prosodia de una grabación proveniente de una persona SD es transferida a una grabación proveniente de una persona SD, el 73 % de las respuestas de los evaluadores identifican la grabación como perteneciente a una persona SD (respuesta 4 y 5 de DSutt+DSpro). La identificación de estas grabaciones como TD tiene un porcentaje muy bajo, el 8 % (respuestas 1 y 2 de DSutt+DSpro). Al observar los resultados para las grabaciones que combinan la señal acústica y la prosodia de grupos opuestos, los resultados son más variables. Cuando la prosodia de una grabación proveniente de una persona TD es transferida a una grabación proveniente de una persona SD, el 58 % de las respuestas de los evaluadores identifican la grabación como perteneciente a una persona TD (respues-

tas 1 y 2 de DSutt+TDpro) y sólo el 20 % identifican la grabación como perteneciente a una persona SD (respuestas 4 y 5 de DSutt+TDpro). Por el contrario, cuando la prosodia de una grabación proveniente de una persona SD es transferida a una grabación proveniente de una persona TD, el 51 % de las respuestas de los evaluadores identifican la grabación como perteneciente a una persona SD (respuestas 4 y 5 de TDutt+DSpro) y sólo el 26 % identifican la grabación como perteneciente a una persona TD (respuestas 1 y 2 de TDutt+DSpro). Además, los resultados de realizar el test no paramétrico de Kruskal-Wallis entre las cuatro combinaciones muestran diferencias significativas entre ellos ($p < 0,001$). Al realizar el test no paramétrico de Mann-Whitney para comparar las combinaciones en grupos de dos, los resultados muestran diferencias significativas entre todos los pares contrastados ($p < 0,001$).

Estos resultados muestran la importancia de la prosodia en la identificación de una voz como atípica. Las grabaciones de personas con desarrollo típico con la frecuencia, energía y duración de fonemas de las grabaciones de personas con síndrome de Down son identificadas en su mayoría como voz atípica. Al mismo tiempo, las grabaciones de personas con síndrome de Down con la frecuencia, energía y duración de fonemas de las grabaciones de personas con desarrollo típico son identificadas en su mayoría como voz típica.

3.3. Evaluación automática de la calidad prosódica

Para analizar qué características son más relevantes a la hora de evaluar automáticamente una grabación como correcta o incorrecta basándose en el uso de la prosodia (objetivo O2.2), se han realizado varios experimentos utilizando las grabaciones almacenadas por el videojuego. En este caso, el corpus contiene 966 grabaciones de 23 usuarios con síndrome de Down. Las grabaciones están previamente evaluadas por una experta en prosodia, y para algunos usuarios, también por la terapeuta de uno de los centros de

Tabla 3.6: Porcentaje de concordancia entre la evaluación de la terapeuta, el clasificador (SVM) y la experta en prosodia por usuario. R representa las grabaciones clasificadas como correctas por el clasificador y W representa las grabaciones clasificadas como incorrectas por el clasificador. Cont.R., Cont. y Rep. representan las grabaciones clasificadas como buenas, regulares y malas por la terapeuta, respectivamente. Cada porcentaje en cada fila es relativo al número de grabaciones de cada tipo clasificadas por la experta en prosodia.

		Evaluación experto		**Clasificado como**		**Evaluación del terapeuta**		
Usuario	**#Total frases**	**Tipo**	**#frases**	**R**	**W**	**Cont.R**	**Cont.**	**Rep.**
S01	120	R	87	83,9 %	16,1 %	69 %	24,1 %	6,9 %
		W	33	57,6 %	42,4 %	30,3 %	36,7 %	33,3 %
S02	106	R	81	87,7 %	12,4 %	85,2 %	14,8 %	0,0 %
		W	25	28,0 %	72,0 %	84,0 %	16,0 %	0,0 %
S03	97	R	78	97,4 %	2,6 %	94,9 %	3,9 %	1,3 %
		W	19	73,7 %	26,3 %	100,0 %	0,0 %	0,0 %
S04	131	R	75	94,6 %	5,3 %	21,3 %	44,0 %	34,7 %
		W	56	41,1 %	58,9 %	5,4 %	32,1 %	62,5 %
S05	151	R	77	87 %	13 %	20,8 %	50,7 %	28,6 %
		W	74	29,7 %	70,3 %	6,8 %	20,3 %	73 %
Total	605	R	398	89,9 %	10,1 %	80,2 %	68,8 %	35,5 %
		W	207	41,1 %	58,9 %	19,8 %	31,2 %	64,5 %

educación especial donde se ha utilizado el videojuego. El objetivo de estos experimentos es reproducir las evaluaciones de la experta en prosodia utilizando las características acústicas utilizadas en el experimento previo y clasificadores automáticos entrenados con las evaluaciones de la experta. Además, se analiza qué características son las más relevantes para realizar esta evaluación automática y cómo influye la heterogeneidad de los usuarios con síndrome de Down en esta evaluación.

La tabla 3.6 muestra la comparativa de las evaluaciones de la experta en prosodia, de la terapeuta y del clasificador que obtiene la mejor tasa de clasificación. Esta comparativa se muestra para los cinco usuarios de los que se tienen los perfiles psicológicos y de los que sus grabaciones están evaluadas por la experta en prosodia y por la terapeuta. Comparando las evaluaciones de la experta en prosodia y del clasificador, se puede observar una alta coincidencia en el caso R-R para todos los usuarios (S01 83,9 %, S02 87,7 %, S03 97,4 %, S04 94,6 %, S05 87 %). La coincidencia en el caso W-W es menor: mientras que en S02 y S05 se observa una coincidencia alta (72 % y 70,3 % respectivamente), la coincidencia para el usuario S03 baja hasta el 26,3 %. Comparando

las evaluaciones de la experta en prosodia con las evaluaciones de la terapeuta, existe una alta coincidencia en el caso R-Cont.R. en los usuarios S01 (69 %), S02 (85,2 %) y S03 (94,9 %). Estos usuarios son los que tienen los niveles de desarrollo cognitivo más altos (ver sección 2.2, tabla 2). En el caso W-Rep., los mayores valores de coincidencia se dan en los usuarios S04 (62,5 %) y S05 (73 %), que son los usuarios con el nivel de desarrollo cognitivo más bajo. Cuando se tienen en cuenta todas las grabaciones de los cinco usuarios, el porcentaje de coincidencia es alto en el caso R-R (89,9 %) y R-Cont.R (80,2 %).

Estos resultados muestran la importancia de la heterogeneidad de los usuarios con síndrome de Down a la hora de evaluar su prosodia. La concordancia en la evaluación realizada por la experta en prosodia y la terapeuta depende del nivel de desarrollo cognitivo de cada usuario y del tipo de evaluación emitida (correcta o incorrecta), según muestra la tabla 3.6. Los diferentes contextos de evaluación pueden influir también en esta concordancia, ya que la experta en prosodia evaluó las grabaciones basándose sólo en los criterios prosódicos definidos (entonación, acento y ritmo) mientras que la terapeuta evaluó en tiempo real y durante el desarrollo del videojuego, por lo que no solo tuvo en cuenta estos criterios sino que también debía evaluar el progreso del jugador, de manera que pudiera evitar producir frustración por el número de evaluaciones incorrectas realizadas. Dentro del contexto del videojuego, es muy importante que el porcentaje de producciones realizadas correctamente por los usuarios pero evaluadas incorrectamente por el clasificador automático sea lo más bajo posible, con el objetivo de no producir frustración en los jugadores. Los resultados de la tabla 3.6 muestran que solo el 10,1 % de las grabaciones evaluadas como correctas por la experta en prosodia fueron evaluadas como incorrectas por el clasificador automático.

La tabla 3.7 muestra los resultados de correlación entre las medidas del perfil de los cinco usuarios analizados anteriormente y las evaluaciones realizadas por la experta en prosodia, la terapeuta y el clasificador automático. La realización del test estadístico Kolmogorov-Smirnov ha mostrado que se cumple la asunción de normalidad necesaria para la utilización del coeficiente de correlación de Pearson para todos los casos

Tabla 3.7: Valores de correlación (Coeficiente de correlación de Pearson) entre los perfiles de los usuarios S01-S05 y la evaluación de la experta en prosodia, de la terapeuta y del clasificador automático. CA representa la edad cronológica de los usuarios. VA representa la edad verbal. NVCL representa el nivel cognitivo no verbal. MPercT y MProdT representan los resultados del test PEPS-C en percepción y producción prosódica, respectivamente. RRate significa el procentaje de evaluaciones correctas de la experta en prosodia. ContRRate, ContRate, RepRate significan el porcentaje de Cont.R, Cont. y Rep. evaluados por la terapeuta, respectivamente. CR significa la tasa de clasificación del clasificador SVM. Los valores en negrita representan diferencias estadísticamente significativas en los valores de correlación con $p < 0{,}05$.

	CA	**VA**	**NVCL**	**MPercT**	**MProdT**	**RRate**	**ContRRate**	**ContRate**	**RepRate**	**CR**
CA	1,0	−0,29	−0,36	−0,53	−0,49	−0,64	−0,52	0,53	0,5	−0,16
VA		1,0	**0,96**	**0,93**	0,84	**0,91**	**0,97**	**−0,91**	**−0,96**	0,42
NVCL			1,0	0,87	0,76	**0,9**	**0,95**	**−0,92**	**−0,93**	0,29
MPercT				1,0	0,83	**0,98**	**0,96**	−0,86	**−0,99**	0,36
MProdT					1,0	0,81	**0,89**	**−0,93**	−0,83	0,80

($p > 0,05$). La edad mental (CA) está altamente y significativamente correlada con el nivel cognitivo no verbal (NVCL), con la percepción prosódica (MPercT), con el porcentaje de evaluaciones correctas de la experta en prosodia (RRate) y con todas las evaluaciones de la terapeuta (ContRRate, ContRate y RepRate). La correlación es positiva para NVCL, MPercT, RRate y ConRRate, mientras que la correlación es negativa para ContRate y RepRate. La NVCL está significativamente correlada con RRate y ContRRate (de forma positiva) y con ContRate y RepRate (de forma negativa). Además, MPercT está significativamente correlada con RRate y ContRRate (de forma positiva) y con RepRate (de forma negativa). MProdT está significativamente correlada con ContRRate (de forma positiva) y con ContRate (de forma negativa). Finalmente, la tasa de clasificación (CR) está altamente correlada con MProdT, pero no de forma significativa.

Estos resultados muestran que existe una dependencia entre el perfil psicológico de los usuarios y las evaluaciones realizadas. Existe más acuerdo tanto entre la experta y el clasificador automático como entre la experta y la terapeuta en los usuarios con un nivel cognitivo más alto y, a su vez, existe menos acuerdo tanto entre la experta y el clasificador automático como entre la experta y la terapeuta en los usuarios con un nivel cognitivo más bajo.

Capítulo 4

Conclusiones y trabajo futuro

4.1. Conclusiones

En esta memoria de tesis se presenta la investigación realizada en el contexto de un marco de trabajo para el desarrollo de videojuegos educativos centrados en el entrenamiento del habla de personas con síndrome de Down, específicamente orientados a la práctica de la prosodia. Concretamente, la investigación se ha centrado en los aspectos de usabilidad e interacción entre los usuarios y el videojuego y en la evaluación de la calidad prosódica de las actividades de producción y percepción oral incluidas en el mismo.

Con respecto al diseño del videojuego, tanto el diseño de las actividades como de la narrativa desarrollada alrededor de las mismas han sido realizados en colaboración con expertos en lenguaje y discapacidad. Además, se ha seguido un diseño centrado en el usuario, implicando en todo el proceso a los usuarios y terapeutas de centros especializados en educación especial. Este proceso ha sido de vital importancia a la hora detectar posibles errores en los diseños y contenidos de las diferentes actividades y en la propia interfaz del videojuego. La evaluación realizada con usuarios con síndrome de Down, utilizando los datos recopilados automáticamente por el videojuego y las observaciones realizadas por los terapeutas, ha permitido la extracción de una serie de aspectos clave a tener en cuenta cuando se desarrollan herramientas enfocadas a personas con síndrome de Down. En primer lugar, es importante incluir refuerzos positivos cuando se resuelvan las actividades correctamente, al mismo tiempo que hay que suavizar los refuerzos negativos cuando la actividad no se supera. Esto es fundamental a la hora de motivar a los jugadores con síndrome de Down a continuar jugando. Además, es necesario incluir un máximo de errores por actividad con el objetivo de no producir frustración en los jugadores que pueda producir bloqueos y el abandono en el uso del videojuego. En segundo lugar, hay que evitar añadir demasiados elementos gráficos que puedan distraer la atención de los elementos importantes que se quieren destacar. La inclusión de pistas visuales puede ayudar a un desarrollo fluido de la mecánica del juego evitando de nuevo posibles bloqueos. Es importante destacar que, a diferencia de la

mayoría de los videojuegos comerciales, en un videojuego educativo la atención debe de estar enfocada principalmente en las actividades de entrenamiento, usando los demás elementos incluidos como soporte de esas actividades. Esta idea cobra especialmente importancia cuando los jugadores objetivo tienen algún tipo de déficit de atención. El uso de instrucciones sonoras que complementen la información visual recibida por los jugadores es otro aspecto clave a destacar. Es necesario dar este tipo de instrucciones cuando el jugador se quede bloqueado en alguna actividad así como durante el desarrollo de la narrativa propuesta, recordándole cada cierto tiempo la información necesaria para continuar con el transcurso del juego. En esta tesis se ha experimentado con la inclusión de un asistente virtual que introduzca estas instrucciones cuando sea necesario. Además, el lenguaje utilizado tanto en los textos como en los audios debe ser sencillo y básico en el uso del vocabulario, de manera que sean comprendidas fácilmente por las personas limitaciones cognitivas, como es el caso de las personas con síndrome de Down. Por último, la inclusión de elementos y escenarios de la vida real es fundamental para que se pueda producir la transferencia de los conocimientos adquiridos durante las sesiones de juego a la vida cotidiana de estas personas.

La heterogeneidad inherente a la población con síndrome de Down se pone de manifiesto cuando se analizan las interacciones de estas personas con el videojuego. Comparando algunos datos almacenados por el videojuego entre los propios jugadores con síndrome de Down, se observa mucha variabilidad en los datos. Además, utilizando métodos estadísticos para comparar los resultados entre poblaciones con síndrome de Down y poblaciones con desarrollo típico (adultos y niños), se observa que los jugadores con síndrome de Down obtienen resultados significativamente diferentes a los obtenidos por personas con desarrollo típico, principalmente en el tiempo requerido para completar las actividades y en el número de errores. Asimismo, analizando los datos recopilados en el videojuego se ha comprobado que las actividades incluidas en el videojuego tienen un nivel de dificultad adecuado a la población con síndrome de Down. Estos datos pueden servir a su vez para detectar problemas particulares de cada jugador analizando las actividades donde el jugador comete más errores o tarda más

tiempo en completarlas, de manera que los terapeutas puedan entrenar estos problemas de manera específica. Las observaciones realizadas por las terapeutas durante las sesiones de juego destacan la motivación hacia el uso del videojuego mostrada por los usuarios con síndrome de Down. Este resultado es destacable ya que las personas con síndrome de Down tienen una serie de limitaciones cognitivas que les dificultan realizar una tarea compleja durante un largo periodo de tiempo, por lo que es positivo tener una herramienta de entrenamiento que pueda mantener la atención de los jugadores durante toda la sesión de entrenamiento.

El uso del videojuego para el entrenamiento permite que, durante las sesiones de juego, las producciones de los jugadores con síndrome de Down sean almacenadas automáticamente. Esto permite la creación de un corpus específico centrado en prosodia de personas con síndrome de Down. Las frases incluidas en las actividades del videojuego incluyen diferentes modalidades y estructuras sintácticas, lo que aporta variedad prosódica al contenido del corpus. La posibilidad de grabar un corpus de voz utilizando el videojuego es especialmente relevante cuando la población objetivo tiene alguna limitación cognitiva, como es el caso de las personas con síndrome de Down. La grabación de este corpus ha permitido realizar una comparación entre la prosodia de personas con síndrome de Down y de personas con desarrollo típico, poniendo el foco en las características acústicas extraídas de las propias grabaciones. Los experimentos realizados usando clasificadores automáticos con el objetivo de clasificar las grabaciones como procedentes de personas con síndrome de Down o de personas con desarrollo típico muestran altas tasas de clasificación, utilizando juntas las características acústicas de los cuatro dominios estudiados (frecuencia, energía, temporal y espectral). Si se utilizan las características de cada dominio por separado, los mejores resultados se obtienen usando las características espectrales. Este resultado es más bajo que cuando se utilizan las características de los dominios de la frecuencia, energía y temporal juntos. Estos resultados parecen indicar que, aunque existen diferencias específicas en cada dominio entre los grupos estudiados, la combinación de las características de los diferentes dominios es la que aporta más discriminación entre las voces de personas con

síndrome de Down y las voces de personas con desarrollo típico. Además, las características espectrales suelen estar relacionadas con el sistema fonológico utilizado para producir sonidos, por lo que los problemas fisiológicos que tienen algunas personas con síndrome de Down determinan el valor de estas características, por lo que no es posible mejorarlas con entrenamiento. Sin embargo, las características relacionadas con la frecuencia fundamental, la energía o el dominio temporal si pueden entrenarse o mejorarse con la práctica, lo que produciría un mejor uso de la entonación, el tono, el acento o el ritmo a la hora de hablar. Por otro lado, el experimento perceptual realizado con las grabaciones modificadas utilizando el algoritmo de transferencia de prosodia entre los dos grupos revela la importancia de la prosodia en la percepción de una voz como atípica, lo que refuerza la idea de que el entrenamiento de la prosodia puede ayudar a estas personas a integrarse mejor en la sociedad.

La heterogeneidad de la población con síndrome de Down afecta también a la evaluación automática de las grabaciones y a la propia evaluación perceptual realizada por expertos. Se ha observado que existe una alta correlación entre medidas del perfil cognitivo de los participantes calculadas utilizando diferentes test de evaluación lingüística (edad verbal, nivel cognitivo no verbal y resultados de habilidades prosódicas) y los resultados de las evaluaciones de los expertos. Los resultados del análisis de la concordancia entre la experta en prosodia y el clasificador automático muestran tasas de coincidencia por encima del 80 % en los casos en los que la experta evaluó una grabación como correcta y el clasificador automático también las evaluó como correctas. En los casos en los que la experta evaluó las grabaciones como incorrectas y el clasificador también las evaluó como incorrectas, la tasa de concordancia es más baja y varía dependiendo del usuario observado. Además, el porcentaje de producciones realizadas correctamente por los usuarios (según el criterio de la experta en prosodia) pero evaluadas incorrectamente por el clasificador automático es del 10,1 %. Este resultado es destacable dentro del contexto del videojuego, ya que una alta tasa de evaluaciones erróneas y negativas de las grabaciones puede provocar frustración a los jugadores y reducir la motivación hacia el videojuego. También se han observado diferentes valo-

res de concordancia entre la experta en prosodia y la terapeuta dependiendo del nivel cognitivo de los usuarios, con mayores tasas de concordancia en los usuarios con los niveles cognitivos más altos y con menores tasas de concordancia en los usuarios con los niveles cognitivos más bajos. Estos resultados sugieren que la utilización de un clasificador automático de propósito general puede no ser la solución óptima, sino que es preferible tener en cuenta esta heterogeneidad a la hora de clasificar las grabaciones del videojuego.

4.2. Trabajo futuro

El marco de trabajo planteado en esta tesis (Figura 1.1) muestra dos aspectos adicionales que no han sido abarcados en este trabajo: la adaptación de las actividades del juego al perfil concreto del jugador y la generación de informes sobre las sesiones de juego. La adaptación al usuario se plantea como un aspecto fundamental debido a la heterogeneidad de la población con síndrome de Down. Dependiendo del perfil psicológico y lingüístico de cada jugador, las actividades del videojuego pueden ser adaptadas para ajustar la dificultad de las mismas y para entrenar aspectos concretos de la prosodia. Esta adaptación podría realizarse automáticamente utilizando la información sobre los resultados en las actividades previas realizadas por ese mismo jugador.

La generación de informes sobre los resultados obtenidos por los jugadores en las diferentes sesiones de juego permitiría a los terapeutas poder realizar un seguimiento de los avances de los jugadores y detectar cuales son los problemas concretos de un jugador tanto en la percepción como en la producción de la prosodia. La combinación de ambos aspectos puede ser útil para la integración de este tipo de herramientas en las terapias de habla realizadas en los centros educativos.

El trabajo presentado en esta tesis sobre la evaluación automática de las producciones de los jugadores es un primer acercamiento a la resolución de esta tarea, pero es necesario profundizar más en este aspecto. Aunque el tamaño del corpus utilizado es suficiente para obtener resultados estadísticamente significativos, es necesario seguir recopilando muestras de audio de personas con síndrome de Down con el objetivo de

afianzar y mejorar los resultados obtenidos en la clasificación automática. Este aumento del tamaño del corpus puede facilitar la exploración de nuevas estrategias de clasificación automática, como el uso de redes neuronales profundas, que necesitan un gran número de datos para obtener resultados fiables. Además, actualmente la evaluación de las muestras de audio realizada por expertos es global a todos los aspectos de la prosodia analizados. Sin embargo, los expertos podrían asignar una puntuación diferente a diferentes dimensiones de la prosodia (entonación, acento o ritmo), con el objetivo de utilizar esos datos en clasificadores automáticos más específicos de cada dimensión. Este enfoque es el elegido para desarrollar un proyecto actualmente en curso (*Incorporación de un módulo de predicción automática de la calidad de la comunicación oral de personas con síndrome de Down en un videojuego educativo*, TIN2017-88858-C2-1-R). Por último, la obtención de un corpus de voz más amplio de personas con desarrollo típico con el que poder comparar las grabaciones de las personas con síndrome de Down puede ayudar a mejorar los resultados de evaluación automática.

4.3. Logros y reconocimientos

4.3.1. Publicaciones

4.3.1.1. Publicaciones en revistas

1. La piedra mágica. Un videojuego educativo orientado a la mejora de las habilidades comunicativas orales como ventana a la inclusión social. Lourdes Aguilar, Yurena Gutiérrez-González, Ferran Adell, David Escudero-Mancebo, César González-Ferreras, Valentín Cardeñoso-Payo, Mario Corrales, Patricia Sinobas, Valle Flores. Revista Síndrome de Down. Fundación Síndrome de Down de Cantabria (España) Número 127, volumen 32, páginas 148-157. Diciembre 2015 [2].

2. Acoustic analysis of anomalous use of prosodic features in a corpus of people with intellectual disability. Mario Corrales-Astorgano, David Escudero-Mancebo, César González-Ferreras. International Conference on Advances in Speech and Language Technologies for Iberian Languages (pp. 151-161). Springer, Cham.

2016 [9].

3. Engaging Adolescents with Down Syndrome in an Educational Video Game. César González-Ferreras, David Escudero-Mancebo, Mario Corrales-Astorgano, Lourdes Aguilar-Cuevas and Valle Flores-Lucas. International Journal of Human–Computer Interaction. 2017 [24].

4. Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome. Mario Corrales-Astorgano, David Escudero-Mancebo, César González-Ferreras. Speech Communication. Volume 99, 2018, Pages 90-100 [10].

5. Automatic Assessment of Prosodic Quality in Down Syndrome: Analysis of the Impact of Speaker Heterogeneity. Mario Corrales-Astorgano, Pastora Martínez-Castilla, David Escudero-Mancebo, Lourdes Aguilar, César González-Ferreras, Valentín Cardeñoso-Payo. Applied Sciences. 2019, 9, 1440 [13].

4.3.1.2. Comunicaciones en congresos

1. Arquitectura para la interacción en un videojuego para el entrenamiento de la voz de personas con discapacidad intelectual. Mario Corrales, David Escudero, Valle Flores, César González, Yurena González. XVI Congreso Internacional de Interacción Persona-Ordenador. 2015 [8].

2. On the Use of a Serious Game for Recording a Speech Corpus of People with Intellectual Disabilities. Mario Corrales-Astorgano, David Escudero-Mancebo, Yurena Gutiérrez-González, Valle Flores-Lucas, César González-Ferreras and Valentín Cardeñoso-Payo Language Resources and Evaluation Conference (LREC) 2016 [12].

3. The magic stone: A video game to improve communication skills of people with intellectual disabilities. Mario Corrales-Astorgano, David Escudero-Mancebo, César González-Ferreras, Yurena Gutérrez-González, Valle Flores-Lucas, Valentín Car-

deñoso-Payo, and Lourdes Aguilar-Cuevas. Interspeech 2016. Pages 1565-1566 [11].

4. 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. Iberspeech 2016 [9].

5. Towards an Automatic Evaluation of the Prosody of Individuals with Down Syndrome. Mario Corrales-Astorgano, Pastora Martínez-Castilla, David Escudero-Mancebo, Lourdes Aguilar, César González-Ferreras, Valentín Cardeñoso-Payo. Iberspeech 2018 [14].

4.3.2. Corpus

La tabla 4.1 muestra el corpus de personas con síndrome de Down almacenado al final de este trabajo de tesis. El corpus cuenta con grabaciones de 23 usuarios con síndrome de Down, con un total de 966 grabaciones que, en términos de duración, hacen un total de 3470 segundos. La disparidad entre los datos se debe al número de sesiones que realizaron con el videojuego cada usuario y la versión del mismo. Además, se ha grabado otro corpus con grabaciones de personas con desarrollo típico, que contiene 250 grabaciones de 22 usuarios, que hacen un total de 650 segundos.

El corpus incluye grabaciones de frases de diferentes modalidades (declarativas, interrogativas y exclamativas). Normalmente, los patrones de entonación varían dependiendo de la modalidad de la frase. Las frases declarativas normalmente acaban con una bajada hacia tonos bajos, mientras que las interrogativas acaban con una subida hacia tonos altos. Por otro lado, las frases parcialmente interrogativas, que están caracterizadas por un elemento interrogativo al principio de la frase, empiezan con tonos altos asociados a la parte interrogativa y acaban con una bajada hacia tonos bajos. Las frases exclamativas son, normalmente, variaciones de las frases declarativas, por lo que la variación radica en aspectos como la energía, el volumen o la modulación del tono usado por el hablante. Además, la combinación de diferentes frases con diferentes

Tabla 4.1: Descripción del corpus generado al final del trabajo de esta tesis.#Grabaciones representa el número de grabaciones por usuario. #Actividades diferentes representa el número de actividades diferentes grabadas por cada usuario. La duración representa la duración total en segundos de todas las grabaciones de un mismo usuario.

Usuario	**#Grabaciones**	**#Actividades diferentes**	**Duración (segundos)**
S01	120	32	364
S02	106	32	391
S03	97	35	244
S04	131	34	385
S05	151	40	544
S06	30	26	147
S07	34	25	196
S08	28	24	86
S09	43	30	251
S10	33	29	104
S11	57	7	171
S12	12	7	34
S13	7	6	34
S14	11	7	93
S15	33	7	130
S16	10	7	34
S17	8	6	24
S18	11	7	44
S19	10	7	40
S20	10	7	38
S21	9	7	46
S22	7	6	37
S23	8	6	33
Total	966		3470 (50 minutos)

modalidades permite la inclusión de inflexiones que indican una segmentación en la producción oral. Dependiendo del contexto y de la velocidad de producción de la frase, estas inflexiones pueden corresponder a una pausa (silencio) y a un fin de frase, o a una semi-pausa, que implica un cambio de entonación dentro de la misma frase.

La grabación del sonido de voz en un soporte informatizado constituye un almacenamiento de datos personales según la Agencia Española de Protección de Datos, por lo que los representantes legales de las personas participantes en la investigación tuvieron que firmar un consentimiento para la utilización de estos datos únicamente para fines de investigación y específicamente para el grupo de investigación ECA-SIMM de la Universidad de Valladolid. La distribución de los datos almacenados durante el desarrollo

de la tesis es interesante de cara a que otros grupos de investigación puedan replicar los experimentos realizados y realizar otro tipo de experimentación, por lo que se ha creado un repositorio donde se almacenan algunos archivos que permiten repetir experimentos sin poner en riesgo la privacidad de los usuarios. Entre estos ficheros se encuentran los que almacenan las características extraídas de las grabaciones o algunos scripts utilizados. Los ficheros generados se pueden encontrar en Github (`https://github.com/macoas/tesis`) y Gitlab (`https://gitlab.inf.uva.es/marcorr/tesis`)

4.3.3. Software

Como producto del trabajo realizado en esta tesis, se dispone del videojuego PRADIA, que ha sido utilizado para realizar las grabaciones del corpus utilizado en esta investigación. El videojuego es una aventura gráfica donde el jugador toma el rol de protagonista, y tiene que avanzar en la historia resolviendo diferentes actividades. El videojuego incluye conversaciones con diferentes personajes, el uso de diferentes objetos y la navegación por diversos escenarios. Para interactuar con el videojuego, el jugador debe hacer uso del ratón del ordenador donde se está ejecutando el juego. Las actividades de entrenamiento están incluidas como parte de la narrativa del videojuego, lo que ayuda a la inmersión de los jugadores en el mismo. Existen tres tipos de actividades diferentes: actividades de comprensión, actividades de producción y actividades visuales. Las actividades de comprensión están enfocadas en la comprensión léxico-semántica y en la mejora de la percepción prosódica en diferentes contextos comunicativos. En estas actividades, el jugador tiene que seleccionar entre diferentes opciones cuál cree que es la mejor para continuar la conversación en curso. Las actividades de producción (figura 4.1) se centran en el entrenamiento de la producción oral, teniendo en cuenta los aspectos prosódicos de la comunicación como la entonación, la expresión de emociones o el acento. Al comienzo de estas actividades, el videojuego introduce el contexto de la actividad mediante algún personaje del juego, y el jugador debe responder usando su propia voz. Estas actividades pueden ser de tres tipos: lectura, donde la frase a producir se le muestra mediante texto al usuario; imitación, donde la frase a producir se le repro-

Figura 4.1: Ejemplo de actividad de producción, donde se le muestra al jugador la frase a reproducir, tanto visual como textualmente.

duce oralmente al jugador; y producción espontánea, donde el jugador no tiene ninguna ayuda ni visual ni textual para resolver la actividad, por lo que tiene que interpretar el contexto para producir la respuesta adecuada. Actualmente, estas actividades deben ser evaluadas por un acompañante externo, que suele ser el terapeuta que trabaja con los jugadores. Para ello, utiliza el teclado del ordenador donde se está ejecutando el videojuego. Por último, las actividades visuales están incluidas para añadir variedad al juego y para ejercitar otras habilidades no relacionadas con el lenguaje, como puede ser la visión espacial, la coordinación óculo-manual o la memoria.

El videojuego incluye 3 episodios de más o menos 15 minutos de duración cada uno. El tiempo requerido para completar cada episodio puede variar según las capacidades y los resultados obtenidos por parte de cada jugador. Dentro de estos 3 episodios se incluyen las diferentes actividades, entre las que se incluyen 40 dinámicas de producción, 28 dinámicas de comprensión y 29 actividades visuales. El videojuego permite el acceso a determinadas escenas de forma directa, con el objetivo de que los terapeutas puedan

repetir sólo las partes que más les interesen. En la página del proyecto PRADIA [1] se puede acceder a la descarga del videojuego así como a toda la información sobre el proyecto.

4.3.4. Proyectos de investigación

La investigación realizada en la presente tesis doctoral ha sido desarrollada en el contexto de los siguientes proyectos de investigación:

- Título: Juguem a comunicar millor! La millora de la competència prosòdica com a via d'integració educativa i d'inclusió social de l'alumnat amb necessitats educatives especials derivades de la discapacitat
  - Investigador Principal: Lourdes Aguilar Cuevas – Universidad Autónoma de Barcelona
  - Fecha de Inicio: 01/01/2014
  - Fecha de Finalización: 01/04/2016
  - Entidad Financiadora: RecerCaixa, ACUP, Obra Social "la Caixa"
  - Cantidad: 77.642,60€
  - Referencia: PZ611683-2013ACUP00202
  - Resumen: El proyecto Recercaixa 2013 «¡Juguemos a comunicar mejor!» pretende impulsar el desarrollo de la competencia comunicativa en el alumnado con Necesidades Específicas de Apoyo Educativo derivadas de la discapacidad. El objetivo es desarrollar una herramienta de autoaprendizaje basada en el uso de un videojuego cuyos niveles se irán superando mediante pruebas de voz. De esta forma, creemos que los alumnos podrán solucionar sus dificultades de fluidez, ritmo, etc. con un método dinámico y sin necesidad de supervisión profesional continuada.

- Título: PRADIA: La aventura gráfica de la pragmática y la prosodia
  - Investigador Principal: Lourdes Aguilar Cuevas – Universidad Autónoma de Barcelona
  - Fecha de Inicio: 20/10/2016
  - Fecha de Finalización: 31/12/2017
  - Entidad Financiadora: Fundación BBVA
  - Cantidad: 59.957,92€
  - Referencia: CF613399
  - Resumen: El propósito principal del proyecto es desarrollar una herramienta de intervención educativa, específicamente un videojuego, en el ámbito de las relaciones entre prosodia (entonación, fluidez, entre otros aspectos) y pragmática (comprensión de las situaciones reales de comunicación) que permita impulsar las habilidades comunicativas del colectivo de Síndrome de Down. La investigación, de carácter aplicado, se orienta hacia el desarrollo de una aventura gráfica (subgénero de los videojuegos de aventura), que permita al alumnado con diversidad funcional adquirir de manera inadvertida unos objetivos de aprendizaje relacionados con sus habilidades comunicativas. Se plantea así un uso innovador de las tecnologías de la información en el tratamiento de una cuestión propia de las humanidades, como es el caso de la vehiculación de contenidos de lingüística en la educación especial.

- Título: Incorporación de un módulo de predicción automática de la calidad de la comunicación oral de personas con síndrome de Down en un videojuego educativo
  - Investigadores Principales: David Escudero Mancebo y Valentín Cardeñoso Payo, Universidad de Valladolid

- Fecha de Inicio: 01/01/2018
- Fecha de Finalización: 31/12/2020
- Entidad Financiadora: Ministerio de Economía, Industria y Competitividad.
- Cantidad: 79,140€
- Referencia: TIN2017-88858-C2-1-R
- Resumen: Se dispone de un videojuego serio para la práctica de la comunicación oral (pragmática y prosodia principalmente) de personas con discapacidad intelectual desarrollado a lo largo de los cuatro últimos años en el marco de dos proyectos de investigación financiados por Resercaixa y BBVA. El videojuego ha demostrado su utilidad al ser capaz de motivar a los usuarios en la realización de ejercicios de tipo práctico en compañía de un entrenador (típicamente un profesor, logopeda o un familiar). También ha demostrado su potencial para recoger corpus de voz de un colectivo de usuarios particularmente difícil. El objetivo de este proyecto es la incorporación al videojuego de un componente inteligente que permita a los usuarios realizar los ejercicios de forma autónoma, sin la necesidad de constante de la presencia del entrenador.

- Título: Herramientas software ludificadas para la evaluación y entrenamiento de la pronunciación
  - Investigadores Principales: Valentín Cardeñoso Payo
  - Fecha de Inicio: 2018
  - Fecha de Finalización: 2020
  - Entidad Financiadora: Junta de Castilla y León

- Cantidad: 12,000€
- Referencia: VA050G18
- Resumen: Recientemente, las tecnologías del habla, un ámbito tradicionalmente dedicado a la síntesis y el reconocimiento de voz de locutor está siendo aplicada a la práctica de la pronunciación asistida por ordenador. El objetivo de este proyecto es profundizar en la aplicación de la experiencia del GIR sobre tecnologías del habla en el ámbito de la pronunciación asistida por ordenador en dos dominios diferentes: la práctica de la pronunciación por parte de personas con discapacidad intelectual, y la práctica de la pronunciación de estudiantes de un segundo idioma. Puesto que el grupo de investigación ha desarrollado ya proyectos en estas líneas, la financiación solicitada en este proyecto servirá para consolidar las investigaciones en estas líneas. Los resultados principales del proyecto a realizar en los próximos tres años serán: el diseño de un módulo de evaluación automática de prosodia de personas con discapacidad intelectual que sirva para mejorar el componente Dashboard del videojuego ya desarrollado; el diseño de un módulo de evaluación y diagnosis de la calidad de la producción de fonemas y prosodia en las herramientas de entrenamiento de pronunciación basadas en pares mínimos; nos centraremos principalmente en el español y el inglés como idiomas L2. Durante el proyecto está previsto culminar las dos tesis doctorales en desarrollo y servir de apoyo para solicitudes de proyectos de investigación competitvos de ámbito nacional y europeo.

Bibliografía

[1] Aguilar, L. Pradia, misterio en la ciudad. `http://www.pradia.net`, 2019. Last accessed: 2018-07-18.

[2] Aguilar, L., Gutiérrez-González, Y., Adell, F., Escudero-Mancebo, D., González-Ferreras, C., Cardeñoso-Payo, V., Corrales-Astorgano, M., and Sinobas, P. La piedra mágica: un videojuego educativo orientado a la mejora de las habilidades comunicativas orales como ventana a la inclusión social. *Revista Síndrome de Down. Fundación Síndrome de Down de Cantabria (España)* (2016), 148–157.

[3] Black, B. Educational software for children with Down syndrome-an update. *Down Syndrome News and Update 6*, 2 (2006), 66–68.

[4] Boone, D. R., McFarlane, S. C., Von Berg, S. L., and Zraick, R. I. *The voice and voice therapy*. Pearson/Allyn & Bacon Boston, 2005.

[5] Cagatay, M., Ege, P., Tokdemir, G., and Cagiltay, N. E. A serious game for speech disorder children therapy. In *2012 7th International Symposium on Health Informatics and Bioinformatics* (2012), IEEE, pp. 18–23.

[6] Chapman, R., and Hesketh, L. Language, cognition, and short-term memory in individuals with Down syndrome. *Down Syndrome Research and Practice 7*, 1 (2001), 1–7.

[7] Corral, S., Arribas, D., Santamaría, P., Sueiro, M., and Pereña, J. Escala de inteligencia de wechsler para niños-iv. *Madrid: TEA Ediciones* (2005).

[8] Corrales, M., Escudero, D., Flores, V., González, C., and 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* (2015), pp. 445–448.

[9] Corrales-Astorgano, M., Escudero-Mancebo, D., and González-Ferreras, C. Acoustic analysis of anomalous use of prosodic features in a corpus of people with intellectual disability. In *International Conference on Advances in Speech and Language Technologies for Iberian Languages* (2016), Springer, pp. 151–161.

[10] Corrales-Astorgano, M., Escudero-Mancebo, D., and González-Ferreras, C. Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome. *Speech Communication 99* (2018), 90–100.

[11] Corrales-Astorgano, M., Escudero-Mancebo, D., González-Ferreras, C., Gutiérrez-González, Y., Flores-Lucas, V., Cardeñoso-Payo, V., and Aguilar-Cuevas, L. The magic stone: a video game to improve communication skills of people with intellectual disabilities. In *Proceedings of Interspeech 2016* (2016), pp. 1565–1566.

[12] Corrales-Astorgano, M., Escudero-Mancebo, D., Gutiérrez-González, Y., Flores-Lucas, V., González-Ferreras, C., and Cardeñoso-Payo, V. On the use of a serious game for recording a speech corpus of people with intellectual disabilities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (2016), pp. 2094–2099.

[13] Corrales-Astorgano, M., Martínez-Castilla, P., Escudero-Mancebo, D., Aguilar, L., González-Ferreras, C., and Cardeñoso-Payo, V. Au-

tomatic assessment of prosodic quality in Down syndrome: Analysis of the impact of speaker heterogeneity. *Applied Sciences 9*, 7 (2019), 1440.

[14] Corrales-Astorgano, M., Martínez-Castilla, P., Escudero-Mancebo, D., Aguilar, L., González-Ferreras, C., and Cardeñoso-Payo, V. Towards an automatic evaluation of the prosody of people with Down syndrome. In *Proc. IberSPEECH 2018* (2018), pp. 112–116.

[15] Cutler, A., Dahan, D., and Van Donselaar, W. Prosody in the comprehension of spoken language: A literature review. *Language and speech 40*, 2 (1997), 141–201.

[16] Dunn, L. M., Dunn, L., and Arribas, D. Peabody, test de vocabulario en imágenes. *Madrid: TEA ediciones* (2006).

[17] Eadie, P. A., Fey, M., Douglas, J., and Parsons, C. Profiles of grammatical morphology and sentence imitation in children with specific language impairment and Down syndrome. *Journal of Speech, Language, and Hearing Research* (2002).

[18] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing 7*, 2 (2016), 190–202.

[19] Eyben, F., Weninger, F., Gross, F., and Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia* (2013), ACM, pp. 835–838.

[20] Feng, J., Lazar, J., Kumin, L., and Ozok, A. Computer usage by children with Down syndrome: Challenges and future research. *ACM Transactions on Accessible Computing (TACCESS) 2*, 3 (2010), 13.

[21] Fidler, D. J., and Nadel, L. Education and children with Down syndrome: Neuroscience, development, and intervention. *Mental retardation and developmental disabilities research reviews 13*, 3 (2007), 262–271.

[22] Fisch, H., Hyun, G., Golden, R., Hensle, T. W., Olsson, C. A., and Liberson, G. L. The influence of paternal age on Down syndrome. *The Journal of urology 169*, 6 (2003), 2275–2278.

[23] González, C., Noda, A., Bruno, A., Moreno, L., and Muñoz, V. Learning subtraction and addition through digital boards: a Down syndrome case. *Universal access in the information society 14*, 1 (2015), 29–44.

[24] González-Ferreras, C., Escudero-Mancebo, D., Corrales-Astorgano, M., Aguilar-Cuevas, L., and Flores-Lucas, V. Engaging adolescents with Down syndrome in an educational video game. *International Journal of Human–Computer Interaction 33*, 9 (2017), 693–712.

[25] Hirst, D. J. *Speech Prosody. From Acoustics to Interpretation.* Springer Verlag GmbH u. Company, 2017.

[26] Huete García, A. Demografía e inclusión social de las personas con síndrome de Down. *Revista Síndrome de Down* (2016).

[27] Kent, R. D., and Vorperian, H. K. Speech impairment in Down syndrome: A review. *Journal of Speech, Language, and Hearing Research* (2013).

[28] Lanfranchi, S., Cornoldi, C., and Vianello, R. Verbal and visuospatial working memory deficits in children with Down syndrome. *American journal on mental retardation 109*, 6 (2004), 456–466.

[29] Lanzoni, M., Kinsner-Ovaskainen, A., Morris, J., and Martin, S. Eurocat – surveillance of congenital anomalies in Europe: epidemiology of Down syndrome 1990-2014. `http://dx.doi.org/10.2760/331810`, 2019. Last accessed: 2019-06-18.

[30] Laws, G., and Bishop, D. V. Verbal deficits in Down's syndrome and specific language impairment: a comparison. *International Journal of Language & Communication Disorders 39*, 4 (2004), 423–451.

[31] Mahatody, T., Sagar, M., and Kolski, C. State of the art on the cognitive walkthrough method, its variants and evolutions. *Intl. Journal of Human–Computer Interaction 26*, 8 (2010), 741–785.

[32] Martin, G. E., Klusek, J., Estigarribia, B., and Roberts, J. E. Language characteristics of individuals with Down syndrome. *Topics in language disorders 29*, 2 (2009), 112.

[33] Martínez, M. H., Duran, X. P., and Navarro, J. N. Attention deficit disorder with or without hyperactivity or impulsivity in children with Down's syndrome. *International Medical Review on Down Syndrome 15*, 2 (2011), 18–22.

[34] Martínez-Castilla, P., Sotillo, M., and Campos, R. Prosodic abilities of spanish-speaking adolescents and adults with williams syndrome. *Language and Cognitive Processes 26*, 8 (2011), 1055–1082.

[35] McFarlane, A., Sparrowhawk, A., Heald, Y., et al. *Report on the educational use of games.* TEEM (Teachers evaluating educational multimedia), Cambridge, 2002.

[36] Nielsen, J. Usability inspection methods. In *Conference companion on Human factors in computing systems* (1994), ACM, pp. 413–414.

[37] Patel, R., Kember, H., and Natale, S. Feasibility of augmenting text with visual prosodic cues to enhance oral reading. *Speech Communication 65* (2014), 109–118.

[38] Peppé, S. J. Why is prosody in speech-language pathology so difficult? *International Journal of Speech-Language Pathology 11*, 4 (2009), 258–271.

[39] Pettinato, M., and Verhoeven, J. Production and perception of word stress in children and adolescents with Down syndrome. *Down Syndrome Research and Practice* (2009).

[40] Polson, P. G., Lewis, C., Rieman, J., and Wharton, C. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of man-machine studies 36*, 5 (1992), 741–773.

[41] Presson, A. P., Partyka, G., Jensen, K. M., Devine, O. J., Rasmussen, S. A., McCabe, L. L., and McCabe, E. R. Current estimate of Down syndrome population prevalence in the united states. *The Journal of pediatrics 163*, 4 (2013), 1163–1168.

[42] Quesenbery, W. The five dimensions of usability. In *Content and complexity*. Routledge, 2014, pp. 93–114.

[43] Raven, J., Raven, J. C., et al. *Test de matrices progresivas: manual/Manual for Raven's progessive matrices and vocabulary scalesTest de matrices progresivas*. No. 159.9. 072. Paidós,, 1993.

[44] Rodríguez, W. R., Saz, O., and Lleida, E. A prelingual tool for the education of altered voices. *Speech Communication 54*, 5 (2012), 583–600.

[45] Saz, O., Yin, S.-C., Lleida, E., Rose, R., Vaquero, C., and Rodríguez, W. R. Tools and technologies for computer-aided speech and language therapy. *Speech Communication 51*, 10 (2009), 948–967.

[46] Stojanovik, V. Prosodic deficits in children with Down syndrome. *Journal of Neurolinguistics 24*, 2 (2011), 145–155.

[47] Tan, T.-S., Ariff, A., Ting, C.-M., Salleh, S.-H., et al. Application of malay speech technology in malay speech therapy assistance tools. In *2007 International Conference on Intelligent and Advanced Systems* (2007), IEEE, pp. 330–334.

[48] Tejedor-García, C., Cardeñoso-Payo, V., Machuca, M. J., Escudero-Mancebo, D., Ríos, A., and Kimura, T. Improving pronunciation of spanish as a foreign language for l1 japanese speakers with japañol capt tool. *Proc. IberSPEECH 2018* (2018), 97–101.

[49] Torrente, J., Del Blanco, Á., Moreno-Ger, P., and Fernández-Manjón, B. Designing serious games for adult students with cognitive disabilities. In *International Conference on Neural Information Processing* (2012), Springer, pp. 603–610.

[50] Wells, B., Peppé, S., and Vance, M. Linguistic assessment of prosody. *Linguistics in clinical practice* (1995), 234–265.

[51] Wuang, Y.-P., Chiang, C.-S., Su, C.-Y., and Wang, C.-C. Effectiveness of virtual reality using wii gaming technology in children with Down syndrome. *Research in developmental disabilities 32*, 1 (2011), 312–321.

[52] Yussof, R. L., and Zaman, H. B. Scaffolding in early reading activities for Down syndrome. In *International Visual Informatics Conference* (2011), Springer, pp. 180–192.