USO DE SISTEMAS DE CLASIFICACIÓN SUPERVISADA PARA LA EVALUACIÓN DE LA PROSODIA: APLICACIONES AL HABLA DE PERSONAS CON SÍNDROME DE DOWN

David Escudero[1], Mario Corrales-Astorgano[1], Yolanda Martín de San Pablo[4], Alfonso Rodriguez de Rojas[5] , Valle Flores[2], César González-Ferreras[1], Valentín Cardeñoso[1], Lourdes Aguilar[3]

Departamento de Informática[1]
Departamento de Psicología[2]
Universidad de Valladolid
Valladolid, España
descuder@infor.uva.es

Spanish Philology Department[3]
Universitat Autònoma de Barcelona
Fundación Personas, Valladolid[4]
Asociación Down Valladolid[5]
España

ABSTRACT

Training prosody is a need for people with Down syndrome. Our team has developed a video game aimed to improve the prosody of persons with DS but it was not designed for use as a self-learning resource. To offer the possibility of autonomous play, this study presents the advances achieved to build a component for the automatic prediction of prosodic quality. We present the system architecture with emphasis on the selection of the analyzed acoustic variables and the prosodic aspects of interest considering the target audience of the game. We contrast the performance of the system for predicting different prosodic dimensions.

El entrenamiento de las competencias de producción prosódica es una necesidad para personas con síndrome de Down (SD). Se dispone de un videojuego educativo para el entrenamiento de la producción oral adaptado a personas con SD. Para ofrecer la posibilidad de juego autónomo, se ha desarrollado un módulo automático de predicción. En este trabajo se presenta la arquitectura de dicho sistema de evaluación automática poniendo el acento en las variables acústicas que se analizan y en el contraste del rendimiento del sistema automático empleado a la hora de predecir diversas dimensiones prosódicas.

Keywords: Evaluation of prosodic quality. Automatic assessment of prosodic quality. Supervised classification. Down syndrome voice. Computer assisted pronunciation training.

1. INTRODUCCIÓN

La evaluación de la prosodia es un ámbito de interés que, abordando la complejidad de la prosodia desde perspectivas enriquecedoras, puede ofrecer soluciones a necesidades reales en el ámbito de la enseñanza de la pronunciación asistida por ordenador, tanto para el aprendizaje de idiomas como para el tratamiento del habla patológica. El trabajo que aquí se presenta se enmarca dentro del proyecto Protoaucal (Ministerio de Ciencia, Innovación y Universidades and the European Regional Development Fund FEDER TIN2017-88858-C2-1-R) para el desarrollo de un módulo de evaluación de la calidad prosódica que permita el uso autónomo de un videojuego para la mejora de la comunicación oral por parte de hablantes con síndrome de Down, un colectivo con dificultades en producción y percepción prosódicas (Stojanovik, 2011; O'Leary, 2020).

El enfoque consiste en entrenar clasificadores supervisados (entrenados con muestras que han sido previamente clasificadas por expertos, Bishop, 2006) que se especializan en aspectos relacionados con la calidad de la prosodia previamente definidos en una rúbrica de evaluación y que comparten los etiquetadores del corpus.

La sección 2 presenta el videojuego en el que se integra el módulo de evaluación automática y la arquitectura del nuevo sistema; la sección 3, la información sobre las variables de calidad y los parámetros acústicos. En la sección 4 se comparan

los distintos clasificadores y se muestran las tasas de clasificación, para acabar en la sección 5 con una discusión sobre las limitaciones de la aproximación.

2. EVOLUCIÓN DEL VIDEOJUEGO PARA EL ENTRENAMIENTO AUTÓNOMO

2.1. Descripción del juego

El potencial de los videojuegos para el entrenamiento de personas con discapacidad es indiscutible, según demuestran estudios como los de Prena (2018), Cano (2019). Como resultado de la labor investigadora del equipo, se dispone de un videojuego educativo para la práctica de la comunicación oral (pragmática y prosodia principalmente) de personas con SD desarrollado en el marco de dos proyectos de investigación financiados por Recercaixa-ACUP y BBVA. Su novedad es que el jugador debe superar una serie de desafíos y misiones en una aventura desarrollada en un entorno de juego estimulante mediante el uso adecuado de los rasgos prosódicos. Estos hitos se ubican en un entorno social de situaciones cotidianas que favorecen la mejora de la competencia comunicativa (González-Ferreras, 2017; Aguilar, 2019).

El videojuego ha demostrado su utilidad al ser capaz de motivar a los usuarios en la realización de ejercicios en compañía de un terapeuta, profesor o familiar. También ha revelado su potencial para recoger corpus de voz de un colectivo de usuarios particularmente difícil. Durante las sesiones de juego se almacena automáticamente tanto información sobre la interacción del usuario con el juego y los resultados en las diferentes actividades como todos los enunciados producidos durante las actividades en que se le pide al jugador que construya, repita o lea una frase. Las grabaciones realizadas durante las sesiones de juego permiten construir un corpus de voz para el estudio de la prosodia en personas con síndrome de Down que ha sido utilizado en varios trabajos previos, como en Corrales-Astorgano (2018), para caracterizar el habla de personas con síndrome de Down.

2.2. Arquitectura para el juego autónomo

En la versión existente del videojuego, un adulto (típicamente el profesor, terapeuta o familiar) se sienta junto al jugador y decide, mediante un teclado accesorio, la corrección o incorrección de los enunciados, de manera que pueda continuar o deba repetir las actividades. Estos juicios se han empleado como medidas de calidad de las producciones orales en trabajos previos como en Corrales-Astorgano (2019), donde se concluye que los juicios de los expertos dependen en gran medida de cada uno de los individuos con SD, tanto de sus capacidades cognitivas como de su estado anímico. A modo de ejemplo, el terapeuta puede permitir continuar en el juego a alguno de los jugadores simplemente para no provocarle situaciones de estrés. En cambio, en otras ocasiones, el terapeuta pide que el jugador repita una actividad si cree que el jugador tiene suficiente habilidad para hacerlo mejor. En una evaluación automática, los factores externos implicados en el desarrollo del juego (nivel de frustración, entre otros) debe dejarse de lado en beneficio del examen de las variables prosódicas.

En una nueva versión del juego, se pretende incluir un módulo de evaluación automática de la calidad del habla para permitir el juego autónomo, de modo que los jugadores puedan practicar la prosodia de forma autónoma empleando sus dispositivos móviles, a la vez que su trabajo queda monitorizado y puede ser supervisado fuera de línea por el terapeuta.

Para disponer de datos enriquecidos sobre los juicios de los expertos, hemos realizado una nueva campaña de grabación en la que los evaluadores no sólo toman la decisión de dejar continuar o no al usuario en la partida, sino que también evalúan con una rúbrica consensuada previamente la calidad de los turnos orales. El sistema automático aprovecha esta información para, además de emitir juicios sobre si el usuario debe o no repetir la actividad, informar sobre aquellos aspectos que el usuario está realizando peor.

3. DESARROLLO DEL MÓDULO DE EVALUACIÓN PROSÓDICA

La figura 1 muestra el esquema de clasificación supervisada seguido. En una primera fase, se debe configurar un corpus de entrenamiento que incluya, junto con las muestras de audio los juicios de calidad emitidos por los expertos. En una segunda fase de entrenamiento, el clasificador analiza las muestras de entrenamiento para especializarse en imitar los juicios de los expertos. En la etapa de test, el clasificador puede ser empleado para hacer el trabajo de los expertos.

En trabajos previos, (Corrales-Astorgano, 2019), hemos utilizado tres tipos diferentes de clasificadores: las redes neuronales, los árboles de decisión y las máquinas de soporte vectorial. Se trata

de tres tecnologías que emplean enfoques diferentes para ajustar los parámetros de clasificación dando resultados que pueden ser complementarios facilitando enfoques posteriores de fusión de expertos. Seguimos aquí un enfoque similar, pero en este caso empleamos los clasificadores en el análisis de los juicios de calidad obtenidos en las rúbricas de evaluación.

En este trabajo, los clasificadores se especializan en la decisión de dejar continuar o no al usuario en el juego y también en el establecimiento de un juicio crítico sobre la calidad de la locución. El uso práctico de este segundo juicio es, por un lado, enriquecer la realimentación que el sistema puede dar al usuario, y de otro lado generar informes para su análisis a posteriori por parte de los expertos.

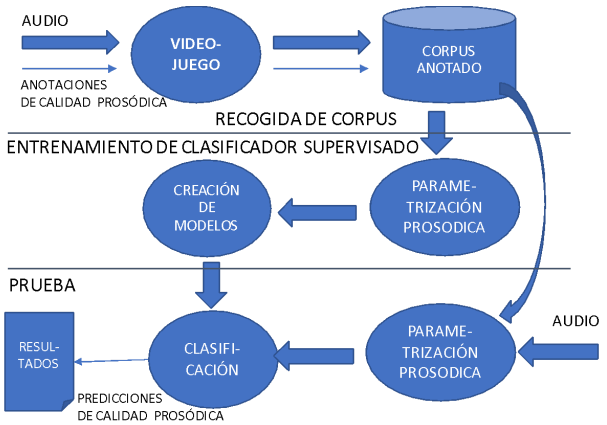

Figura 1: Esquema de entrenamiento de sistemas de clasificación supervisados.

3.1. Variables de calidad

Con el fin de disponer de un procedimiento de evaluación de la calidad oral de los enunciados que permita desarrollar el módulo automático de predicción, se ha acordado con los agentes educativos un mapa de criterios de evaluación. Una vez definidos, puede procederse a la recogida de datos para el corpus de entrenamiento. Mediante un dispositivo hardware, el terapeuta valora, en tiempo de juego, las siguientes variables relacionadas con la calidad de la producción oral en general y prosódica en particular:

Inteligibilidad: valora la capacidad del evaluador para transcribir la locución en un mensaje.

Adecuación: indica si el mensaje emitido por el jugador es adecuado al contexto comunicativo del juego.

Omisiones: número de palabras que el locutor ha omitido. Es necesario porque las incorrecciones gramaticales son frecuentes.

Fluidez: mide si la locución contiene puntos de interrupción inapropiados, repeticiones y disfluencias.

Velocidad: penaliza las producciones orales demasiado lentas.

Curva melódica: penaliza construcciones poco expresivas o con una entonación inapropiada.

Continuar: indica si la calidad de la producción es lo suficientemente buena para continuar con el juego o, por el contrario, tiene que repetir la actividad.

Inteligibilidad, adecuación y continuar son variables binarias (SI/NO). Para el resto de variables, se emplea una escala Likert de 1 a 4 y se asignan valores sólo cuando la producción oral se considera inteligible y correcta.

3.2. Parámetros acústicos

Los clasificadores utilizan un conjunto reducido de parámetros prosódicos extraídos empleando el extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), descrito en Eyben et al. (2015) que obtiene con la la herramienta openSmile (Eyben et al., 2013). En particular, estos parámetros están relacionados con F0: frecuencia fundamental y jitter; con la energía: intensidad y shimmer; y con la dimensión temporal: tasa de picos de intensidad por segundo, longitud media y desviación estándar de los segmentos sonoros y no sonoros y la tasa de segmentos sonoros. Se eliminan en este estudio las variables espectrales empleadas en Corrales-Astorgano (2018). Para cada coeficiente se calcula la media y el coeficiente de variación. Para la frecuencia fundamental y la intensidad se mide el percentil 20, 50 y 80, la media y desviación estándar de la pendiente ascendente o descendente, la media global y la diferencia entre el percentil 80 y el 20. A mayores, se incluye el porcentaje de silencio dentro de la locución, la tasa de pausas por segundo y la duración media de las pausas. En total, se utilizan 34 características.

4. RESULTADOS

Empleamos los datos recogidos en la última campaña de evaluación realizada durante el curso 2018/19 en el colegio Tórtola y en la Asociación Síndrome Down Valladolid. Son 601 frases evaluadas de 17 locutores diferentes. Dos profesores

anotaron las variables prosódicas. La tasa de consistencia, entre los etiquetadores va desde 0.33 a 0.9, usando Kendall (1990), dependiendo de la dimensión analizada. La tabla 1 muestra los resultados de clasificación obtenidos para cada una de las variables de clasificación, excepto la inteligibilidad y la adecuación, que no se pueden medir utilizando parámetros acústicos:

Tabla 1. Tasa de clasificación (TC) y porcentaje de falsos negativos (FN) por cada dimensión y clasificador. Las muestras en las que no se ha realizado la evaluación se han eliminado de este estudio. DT significa Decision Trees, SVM, Support Vector Machines y MLP, Multilayer Perceptron.

	DT		**SVM**		**MLP**	
Variable	**TC**	**FN**	**TC**	**FN**	**TC**	**FN**
Fluidez	0.63	0.46	**0.69**	0.41	0.64	0.42
Velocidad	0.68	0.25	**0.7**	0.27	0.65	0.35
Curva	*0.61*	0.15	*0.61*	0.17	*0.62*	0.3
Continuar	0.69	0.14	**0.73**	0.02	0.69	0.18

Los mejores resultados de clasificación se obtienen utilizando el clasificador SVM, salvo en el caso de la curva melódica, en la que los resultados son similares en los tres clasificadores. Con respecto a la tasa de falsos negativos utilizando el clasificador SVM, las tasas más bajas se obtienen para las dimensiones de omisiones (0), continuar (0.02) y curva (0.17). Las tasas de falsos negativos son más altas en las dimensiones de velocidad (0.27) y fluidez (0.41).

5. DISCUSIÓN

El uso de métodos de clasificación supervisada para predecir la calidad prosódica permiten alcanzar altas tasas con un número de variables relativamente pequeño. En próximos desarrollos se pondrá en evidencia si esta tasa es lo suficientemente alta para el propósito que se persigue, que no es otro que el de permitir el juego autónomo de los usuarios. Hemos puesto el acento en presentar bajas tasas de falsos negativos porque pueden provocar en el usuario una sensación de frustración deteriorando en su predisposición hacia el entrenamiento con el videojuego.

Las tasas de predicción de las variables son útiles para ofrecer indicadores a los terapeutas sobre el rendimiento de los alumnos y para ser relacionadas con configuraciones anómalas en la producción oral, por ejemplo, alargamiento excesivo de pausas, o monotonía en los contornos de F0.

Apuntamos los siguientes aspectos que deben considerarse en trabajos futuros que pretendan mejorar las tasas de predicción:

1. Usar corpus con un número elevado de grabaciones y con categorías equilibradas. Se trata de un punto débil importante, porque disponer de un corpus de habla mayor de hablantes con discapacidad intelectual es muy complicado, como se discute en Corrales-Astorgano (2018).
2. Asegurar en la medida de lo posible la consistencia de los juicios de los expertos que evalúan la calidad prosódica. Evaluar prosodia siempre es difícil porque un mismo mensaje puede emitirse con diferentes realizaciones prosódicas, todas diferentes y correctas.
3. Tener en cuenta la heterogeneidad de los locutores. En Corrales-Astorgano (2019) mostramos que el impacto del locutor es importante por la gran diversidad de perfiles que se encuentran en el colectivo con síndrome de Down.
4. Emplear múltiples dominios de evaluación para focalizar el entrenamiento en uno de los ámbitos con el fin de lograr mayor efectividad.

6. CONCLUSIONES

Se ha presentado un trabajo de implementación de un módulo de predicción de calidad prosódica que se va a incluir en un videojuego educativo para el entrenamiento de la producción oral, en general, y prosódica, en particular ,por parte de usuarios con síndrome de Down. La precisión del clasificador valora su futura utilidad en este dominio con la inclusión de sesgos para penalizar la presencia de falsos negativos. Es trabajo futuro estudiar la relación entre las tasas de predicción de las variables de calidad prosódica y las posibles configuraciones de parámetros acústicos anómalas. Se destacan como puntos clave para mejorar las tasas de predicción el uso de corpus grandes equilibrados con anotaciones fiables sobre múltiples dimensiones y la consideración como variable de la elevada diversidad de los informantes.

6. BIBLIOGRAFÍA

Aguilar, L. (2019). Learning Prosody in a Video Game-Based Learning Approach. Multimodal Technologies and Interaction, 3(3), 51.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Cano, A. R., García-Tejedor, Á. J., Alonso-Fernández, C., & Fernández-Manjón, B. (2019). Game Analytics Evidence-Based Evaluation of a Learning Game for Intellectual Disabled Users. *IEEE Access*, *7*, 123820-123829.

Corrales-Astorgano, M., Martínez-Castilla, P., Escudero-Mancebo, D., Aguilar, L., González-Ferreras, C., & Cardeñoso-Payo, V. (2019). Automatic Assessment of Prosodic Quality in Down Syndrome: an Analysis of the Impact of Speaker Heterogeneity. *Applied Sciences*, *9*(7), 1440.

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

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190-202.

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

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. International Journal of Human–Computer Interaction, 33(9), 693-712.

Kendall, M. G., & Gibbons, J. D. (1990). Rank correlation methods. New York, NY : Oxford University Press

O'Leary, D., Lee, A., O'Toole, C., & Gibbon, F. (2020). Perceptual and acoustic evaluation of speech production in Down syndrome: A case series. *Clinical linguistics & phonetics*, *34*(1-2), 72-91.

Prena, K., & Sherry, J. L. (2018). Parental perspectives on video game genre preferences and motivations of children with Down syndrome. *Journal of Enabling Technologies*.

Stojanovik, V. Prosodic deficits in children with Down syndrome. J. Neurolinguist. 2011, 24, 145–155.