Metodología para la construcción de voces artificiales de niños para la inclusión educativa

Methodology for the generation of artificial voices of children for inclusive education

Maribel Morales Rodríguez
Marvin Coto Jiménez

Resumen: La integración de voces artificiales en dispositivos tecnológicos es una opción para favorecer la comunicación en personas con discapacidad, ya que permite mayores herramientas para la inclusión de esta población, haciendo valer sus derechos establecidos en la Ley 8661. En la Universidad de Costa Rica se cuenta con el proyecto ED-3416, inscrito desde Acción Social y relacionado con tecnologías del habla para mejorar la calidad de vida de la población con discapacidad Este proyecto apuesta por contribuir en los procesos de inclusión formulando sistemas de comunicación aumentativa con uso de voz artificial para niños, a través de la generación de voces artificiales acordes con el género y la edad a su propia identidad. Los avances recientes en tecnologías del habla, apoyados en sistemas que incorporan inteligencia artificial, hacen factible la generación de voces con sonido más flexibles, abriendo la posibilidad de crear voces personalizadas de acuerdo con acentos y condiciones específicas. Uno de los mayores retos que se tiene es la obtención de datos de calidad, a partir de voces naturales de niños para poder emularlos con ayuda de la computadora. En este artículo se evidencia el diseño de datos y las estrategias de interacción con niños para grabar sus voces de manera que sean aprovechables para crear voces artificiales nuevas que se puedan aplicar en sistemas de comunicación aumentativa que promuevan el acceso de sus usuarios a una participación activa en sus propios procesos de aprendizaje.

Palabras clave: acción social, discapacidad, inclusión educativa, tecnologías del habla.

Abstract: The integration of artificial voices in technological devices is an option to favor communication in people with disabilities, since it allows greater tools for the inclusion of this population, according to their rights established in Law 8661. In the University of Costa Rica the project ED-3416 related to speech technologies to improve the quality of life of the population with disabilities, registered by Acción Social, which is committed to contributing to inclusion processes by formulating augmentative communication systems with the use of artificial voice for children, generating voices that adjust in gender and age to their own identity. Recent advances in speech technologies, supported by systems that incorporate artificial intelligence, make it possible to generate more flexible voices with sound, opening the possibility of creating personalized voices according to specific accents and conditions. One of the biggest challenges is obtaining quality data, from the natural voices of children, to be able to emulate them with the help of the computer. In this paper we show the design of data and the strategies of interaction with children, to record their voices so that they are used to create new artificial voices that can be applied in augmentative communication systems that promote the access of their users in their own learning processes.

Key words: inclusive education, disabilities, social action, speech technologies

1. INTRODUCCIÓN

El uso de computadoras y dispositivos tecnológicos se ha normalizado en la cotidianeidad de gran cantidad de personas, especialmente en los últimos años. Con este incremento en el uso, las posibilidades de aplicación y aprovechamiento de estas tecnologías en gran cantidad de áreas se ha multiplicado, convirtiéndose en un medio de gran interés para la comunicación y el acceso a la información. Al ser el habla la principal forma de comunicación humana, existe un área de investigación dedicada a trasladar la capacidad de comprender y emitir mensajes hablados a la tecnología. Esta área es conocida como tecnologías del habla.

Las dos principales ramas de las tecnologías del habla son el reconocimiento de voz (análogo a la capacidad humana de escuchar y comprender), y la síntesis de voz (análogo a la capacidad humana de emitir mensajes hablados). Cada una de estas áreas tiene subáreas de especialización para modelar e imitar todos los complejos procesos que realizan los sentidos y el cerebro humano (Coto y Morales, 2020).

Existen aún muchos desafíos para construir sistemas sólidos y más naturales que sean útiles en sistemas de comunicación para todas las personas, incluidos los adultos mayores, los niños y las personas con discapacidad. Para el caso de los niños, el empleo de estas tecnologías podría ser beneficioso en varias áreas de aplicación, incluida la seguridad y la educación general e inclusiva (Safavi et al., 2016, p. 1836).

Por ejemplo, en las plataformas de redes sociales que utilizan el reconocimiento de voz, un sistema puede identificar a un niño según su voz y confirmar la identidad de la persona con la que se está comunicando. En educación, un tutor virtual interactivo podría

identificar a cada niño en una clase y podría adaptar su contenido a condiciones o situaciones específicas (Llanos, 2010). Además, el reconocimiento y la síntesis del habla pueden ayudar al desarrollo del habla y el lenguaje en niños pequeños. Por lo tanto, podría decirse que puede ayudar a los niños a mejorar su capacidad de comunicación (Li 2002, p. 1).

En el área de inclusión educativa, el desarrollo de sistemas de comunicación aumentativa y alternativa (SAAC, por sus siglas en inglés) representa un área de oportunidad valiosa, por medio de la cual se puede dotar de capacidades de comunicación verbal a estudiantes con discapacidad que son no verbales. De hecho, las tecnologías más recientes permiten vislumbrar la posibilidad de desarrollar voces personalizadas, de manera que cada usuario cuenta con una voz única, como es única la voz de cada individuo hablante.

A pesar de esta importancia y de las valiosas posibilidades que permite, el desarrollo de tecnologías del habla con niños es un campo con bajo desarrollo en ciertos contextos, como en el caso de Costa Rica. En especial, en cuanto al desarrollo de voces artificiales que puedan formar parte de los SAAC en esta población. Es bien sabido que las características del lenguaje son variables entre los idiomas y los acentos. Por esto, es de importancia establecer con claridad la metodología de desarrollo de voces artificiales en aquellos rangos de edad que representan mayores retos, como en el caso de niños. De esta manera, se pueden evaluar propuestas y establecer los requerimientos de personas e insumos tecnológicos para llevar a la realidad la producción de voces artificiales de niños en Costa Rica.

2. LA EDUCACIÓN INCLUSIVA EN COSTA RICA

La inclusión actualmente es una meta por lograr en los centros educativos, tanto públicos como privados, de Costa Rica. Esto ha sido impulsado por el modelo social con enfoque de derechos humanos al que se circunscribe el Estado Costarricense desde el planteamiento de la Ley 7600 de Igualdad de Oportunidades para las personas con discapacidad en el año 1996. A partir de esta y de la ratificación de la Ley 8661 Convención de Derechos Humanos de las personas con discapacidad en el año 2008, se establece el norte a seguir en cuanto al respeto de garantías para la población en condición de discapacidad.

Asimismo, como parte de la agenda 2030 de la ONU para el desarrollo sostenible, el Estado Costarricense "pretende visibilizar la situación de cinco grupos históricamente excluidos en Costa Rica y los desafíos que enfrentan, a saber: la población indígena, la población afrodescendiente, la población LGTBI, la población con discapacidad y mujeres, niñas y adolescentes" (Programa de Naciones Unidas para el Desarrollo [PNUD] - Costa Rica, 2017, p. 16). Específicamente, en relación con los desafíos que enfrenta la población con discapacidad, se refleja la urgencia de una educación de calidad, de carácter inclusiva e igualitaria en el marco de sus derechos humanos y de acuerdo con las leyes y marcos mencionados anteriormente.

Es sabido que dentro de los derechos fundamentales para que un ser humano pueda vivir en comunidad se encuentra el acceso a la información (Mathiesen, 2008), el cual debe darse en cualquier ámbito en el que se desenvuelva la persona, incluyendo el educativo. Sin embargo, los niños, niñas y adolescentes en condición de discapacidad que requieren de sistemas de comunicación alternativos

se enfrentan constantemente con barreras de acceso a la información y a la comunicación desde el mismo núcleo familiar y ámbito social, lo que deviene en un proceso totalmente excluyente. Como bien lo indica Vega y del Rocío (2018) “existen muchas personas que por su condición de discapacidad cuentan con barreras fisiológicas y contextuales que las exponen a la exclusión social” (p. 29).

Pese al hecho de que la comunicación por medios tecnológicos u otros tipos de formatos que van más allá de la oralidad se contempla en la legislación anteriormente citada, esto no ha sido garantía de que el recurso comunicativo se encuentre al servicio de los usuarios que lo requieren. Algunas de las razones son la falta de capacitación, el desconocimiento acerca de dispositivos de alta tecnología, el costo de estos, entre otros.

Es importante considerar que las actitudes, las estrategias, los recursos y las metodologías actuales son la base para los cambios del mañana. Por lo tanto, es imperativo un cambio actitudinal inmediato ante el uso de la tecnología por parte de los profesionales que trabajan en el área de la discapacidad. Es preciso lograr la construcción de espacios escolares más justos y equitativos para lograr un cambio en la visión de los apoyos que se ofrecen actualmente en las instituciones educativas a las personas con discapacidad.

Ante esta realidad, surge el proyecto llamado “Tecnologías del habla para mejorar la calidad de vida de la población con discapacidad”, inscrito en la Universidad de Costa Rica por medio de la Vicerrectoría de Acción Social. Este apuesta a contribuir en los procesos de inclusión formulando sistemas de comunicación aumentativa con uso de voz artificial para niños. Si bien es cierto que existen dispositivos en el mercado que cuentan con voz artificial, la gran mayoría hacen uso de voces de adultos, y los que tienen voces de niños son de un

costo económico alto. Por otra parte, las características de las voces no siempre presentan coincidencias con género, edad o identidad del país de los usuarios. Para el uso de los sistemas comerciales, es común el requerimiento de un proceso de entrenamiento, el cual pocos profesionales han recibido en el país.

El proyecto ED-3416, "Tecnologías del habla para mejorar la calidad de vida de la población con discapacidad", inscrito desde la Escuela de Ingeniería Eléctrica, busca generar voces artificiales que se ajusten en género y edad a la propia identidad de los niños y niñas que así lo requieran. Para ello, se ha conformado un equipo transdisciplinario de profesionales en áreas tan distintas como complementarias para los fines de la investigación. Se cuenta, por ejemplo, con un ingeniero electricista especialista en tecnologías del habla, una terapeuta física especialista en primera infancia, una orientadora especialista en familia y dos docentes de enseñanza especial especialistas en derechos humanos, comunicación y neurociencia. El equipo se presenta con el objetivo de ampliar la visión a un modelo integral que concibe al individuo como un todo y respeta no solo sus particularidades sino también la globalidad del acceso al entorno social en igualdad de oportunidades.

En el entorno escolar, se pretende que toda la comunidad educativa pueda tener el mismo derecho de acceso a la información y a la comunicación, por lo que el uso de un SAAC no solo ofrece al usuario el ejercicio de su derecho a comunicarse, sino que también provee a sus compañeros de aula y docentes ese mismo derecho en relación con el estudiante con discapacidad.

Pero la educación inclusiva no se basa únicamente en la atención de la población con discapacidad, sino que busca que este enfoque, dentro de las escuelas, esté fundamentado en el acceso a una

educación de calidad para todos y todas. Tal y como lo menciona Arnaiz (1996, p. 3), "la atención a la diversidad no es solamente atención a los alumnos con necesidades educativas especiales, sino a todos los alumnos escolarizados en un centro educativo". Por ello, dicho enfoque ha obligado a las escuelas a modificar sus prácticas educativas con el fin de brindar igualdad de oportunidades y equidad a todo el estudiantado, independientemente de si presenta alguna condición de discapacidad, respetando los diferentes estilos y ritmos de aprendizaje y ofertando una enseñanza individualizada.

3. PROCEDIMIENTO GENERAL DE CREACIÓN DE VOCES ARTIFICIALES

El objetivo técnico de crear una voz artificial se puede establecer como el convertir un texto cualquiera en una señal de sonido que sea indistinguible del habla humana (Coto-Jiménez y Goddard-Close 2016a, p. 2). El procedimiento requerido para llevarla a cabo puede dividirse en dos etapas: la primera se trata de convertir un texto dentro de un dispositivo (como un mensaje, una instrucción o respuesta) en una especificación lingüística, y el segundo convertir esta especificación en la señal de audio (Zen et al., 2009, p. 3). La especificación lingüística realiza la transcripción fonética del texto y la especificación de la pronunciación de los fonemas. La conversión en señal de audio debe considerar varios parámetros, como el tono, la amplitud y la duración (Coto-Jiménez y Goddard-Close 2016b).

Estos procedimientos son necesarios cuando se desea que un SAAC pueda emitir mensajes arbitrarios a solicitud del usuario. Cuando en una aplicación se tiene solamente una cantidad limitada de palabras o posibilidades, las implementaciones pueden realizarse con frases pregrabadas. En un entorno educativo, dado que se

trata de un proceso de aprendizaje y en un contexto de interacción general, no es posible pensar en pregrabar todos los mensajes que un usuario vaya a utilizar. En estos casos es indispensable realizar síntesis de voz para que el estudiante pueda expresar de forma verbal lo que desee comunicar.

Desde la década de 1990, el método dominante para realizar síntesis de voz ha sido la selección de unidades, en el cual se toman segmentos de audio de longitud variable de una base de datos adecuadamente etiquetada y se unen para conformar nuevas frases. Este método tiene requerimientos considerables en cuanto a cantidad de datos almacenados y procesamientos necesarios (Rabiner 2007, p. 20). Para esto se siguen tres etapas (Hunt y Black 1996, p. 373):

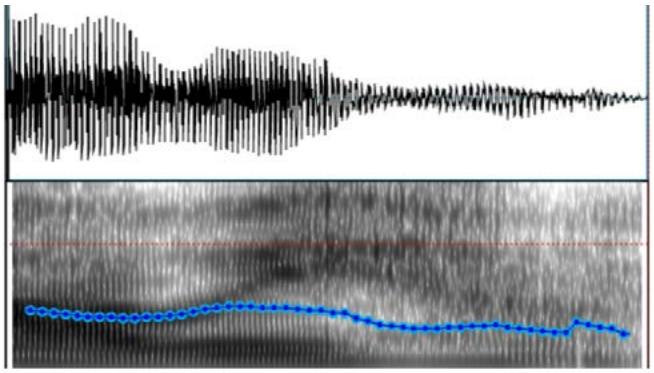
1. Grabar en lenguaje natural habla suficiente, de manera que se abarquen todos los sonidos y las combinaciones (a nivel de difonos).
2. Segmentar o etiquetar las unidades obtenidas en las grabaciones.
3. Seleccionar las unidades más apropiadas al generar nuevas frases.

Los tres problemas principales que se encuentran al elaborar un sintetizador concatenativo son

1. La distorsión producida por las discontinuidades en los puntos de concatenación.
2. Altos requerimientos de memoria.
3. La recolección y etiquetado de datos requiere gran cantidad de tiempo
   (Gonzalvo et al., 2007, p. 7; Zen et al., 2007, p. 294).

Por ejemplo, para producir una voz de calidad como la generada en los dispositivos más actuales, se requieren decenas de horas de grabación. Esta debe realizarse en condiciones controladas de ruido, como en estudios de grabación, y con locutores o actores profesionales que logren homogeneidad en la emisión de frases a lo largo del tiempo. De esta manera, cuando se tomen y se unan fragmentos de esas grabaciones, el resultado seguirá siendo natural y con transiciones suaves entre los sonidos. Cada frase grabada debe ser analizada con cuidado para establecer su conveniencia dentro de la expresión que se desee dar a la voz, por ejemplo a través de la observación de distintas representaciones de la misma dentro de la computadora, como los oscilogramas y espectrogramas que se muestran en la Figura 1. El oscilograma sirve para comparar la intensidad del sonido a lo largo del tiempo. Aquí se pueden observar patrones como tartamudeo, silencios entre palabras y duración general de la emisión. En el espectrograma se observa la intensidad de diferentes frecuencias (como tonos agudos o graves) y se pueden comparar con otros para identificar y establecer similitudes o diferencias (Mascorro y Torres, 2013, pág. 15).

Figura 1. Dos representaciones de la señal de habla:
Oscilograma (parte superior) y espectrograma (parte inferior).



Fuente: Elaboración propia a partir del programa Praat.

Como puede observarse, con ambas se puede mostrar la presencia de un sonido (voz) que baja en intensidad hacia el final de la pronunciación. En la primera se observa una reducción de la amplitud, mientras que en la segunda se observa con una disminución de los tonos de gris.

Si bien este método de producir voz artificial concatenando unidades logra resultados de calidad, el insumo requerido y la cantidad de almacenamiento hizo que diversos grupos de investigadores a lo largo del mundo buscaran nuevas formas de producirla. Así, a partir de la década del año 2000, surge el método de síntesis estadística paramétrica de voz, el cual está basado en modelos matemáticos del habla en lugar de grabaciones preexistentes.

Este nuevo método contempla los siguientes pasos, según Zen (2009, p. 1142):

1. Grabar en lenguaje natural habla suficiente, de manera que se abarquen todos los sonidos del lenguaje o acento particular.
2. Extraer representaciones matemáticas de los sonidos.
3. Segmentar o etiquetar las unidades obtenidas en las grabaciones y hacer una correspondencia entre las representaciones matemáticas y las lingüísticas.
4. Al pronunciar una nueva frase, buscar entre los modelos matemáticos aquellos que mejor representen los sonidos que se deseen emitir.

Como se puede observar, el proceso es un poco más complejo que el indicado para concatenación de unidades. Sin embargo, las nuevas ventajas que se pueden obtener son:

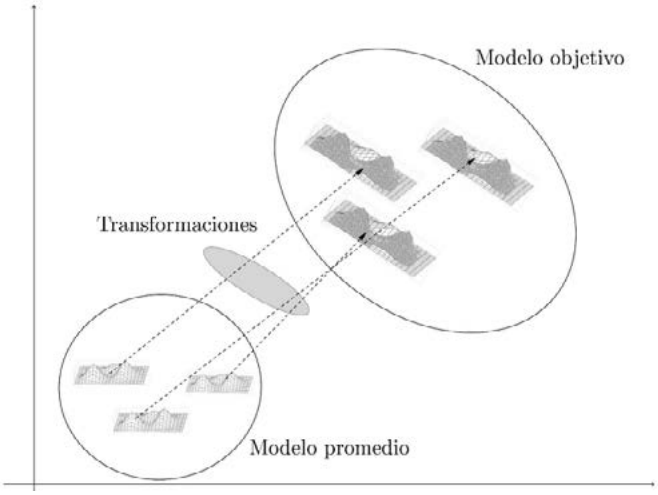
1. Un menor espacio de almacenamiento requerido.
2. Menor cantidad de grabaciones necesarias para producir una voz.
3. Mayor flexibilidad al operar sobre modelos matemáticos en lugar de grabaciones reales.

A pesar de los menores requerimientos, la cantidad de grabaciones necesarias para producir una voz de calidad son del orden de horas. En la Tabla 1 se muestran algunas referencias representativas para distintos idiomas. Estas serán de importancia para mostrar los retos que representa generar datos útiles para producir voces artificiales de niños.

En la segunda mitad de la década del 2000, surgen nuevas posibilidades para mejorar la calidad de las voces generadas con esta nueva técnica. Estas parten del hecho de que los modelos matemáticos que representan una voz en particular pueden requerir solo pequeñas modificaciones para que suene como la de una

persona distinta. Esta nueva posibilidad fue llamada síntesis de voz por adaptación (Yamagishi et al., 2009). En la Figura 2 se ilustra esta posibilidad. El modelo promedio se refiere al conjunto generado por medio de grabaciones de calidad, mientras que el modelo objetivo es aquel del cual se cuenta con pocos datos.

Figura 2: Ilustración del procedimiento de síntesis de voz por adaptación



Fuente: Elaboración propia

Por ejemplo, esta nueva posibilidad permite generar voces de niños, de los cuales es muy difícil generar una cantidad de grabaciones con las características deseadas, tales como las longitudes mostradas en la Tabla 1.

Tabla 1: Tiempo de grabación utilizado para generar voces artificiales en distintos idiomas

Año	Idioma	Cantidad de datos
2002	Inglés	1 hora (Tokuda et al., 2002)
2006	Croata	1 hora 25 minutos (Ipsic y Martincic-Ipsic, 2006)
2007	Alemán	3 horas (Krstulović et al., 2007)
2008	Griego	1 hora y media (Karabetsos, 2008)
2008	Catalán	1 hora (Bonafonte et al., 2008)
2010	Checo	5 horas (Hanzlíček, 2010)
2013	Tamil	5 horas (Boothalingam et al., 2013)

Fuente: elaboración propia

Ahora, con un conjunto mucho menor de datos (aún no especificado en las referencias, por lo que requiere intensa experimentación), emerge la posibilidad de generar un modelo objetivo a partir de pocas grabaciones de niños, así como de generar uno de calidad a partir del modelo promedio (Coto-Jiménez y Goddard-Close 2016c, p. 418).

Por lo anterior, es también necesario partir de grabaciones de voces reales de niños. Y dado que es deseable dotar a cada usuario potencial de voces artificiales en SAAC de una voz personalizada, también es deseable en este contexto generar una gran cantidad de grabaciones de voces de niños y niñas. De manera que aunque la posibilidad tecnológica ha emergido, la generación de datos de voces de menores de edad como primer paso ya presenta retos considerables.

4. ESTABLECIMIENTO DE UN CORPUS LINGÜÍSTICO PARA NIÑOS COSTARRICENSES

Uno de los mayores retos en tecnologías del habla, y específicamente en la creación de voces artificiales de niños, es la obtención de audios de calidad con los cuales se puedan aplicar los sistemas y algoritmos disponibles para generarlas. Si bien la técnica de adaptación requiere poca cantidad de grabaciones, las que se obtengan deben tener ciertas condiciones mínimas, como claridad en los sonidos, dicción adecuada y libre de ruidos.

Con el fin de proporcionar estos insumos para la base de datos de voz infantil, se decidió realizar una primera sesión de grabación con niñas y niños, en un estudio profesional. Por otra parte, la selección de palabras o frases que los y las menores participantes podrían generar, así como la estrategia de interacción para obtener las grabaciones, fueron consideraciones primordiales y de un previo diseño cuidadoso. La sesión se realizó en el mes de enero de 2019 y contó con la participación de 5 menores de edad entre los 6 y los 12 años. El género, edad y lugar de procedencia de la población infantil participante se describen en la Tabla 2.

Tabla 2: Características de las personas menores participantes en la sesión de grabación

Género	Edad	Procedencia
Masculino	6 años	San José
Femenino	8 años	Alajuela
Femenino	8 años	San José
Masculino	11 años	San José
Femenino	12 años	Alajuela

Fuente: elaboración propia

La sesión fue realizada en el estudio de grabación de la Escuela de Ciencias de la Comunicación Colectiva y se contó con el consentimiento informado de los padres de las niñas y niños participantes. Para el diseño de la base de datos se desarrolló una estrategia de interacción con las niñas y niños participantes en donde se utilizó el juego imaginativo, la conversación sobre temas de interés y el uso de instrumentos formales y no formales propios de la valoración del lenguaje oral. Por ejemplo, dentro de las valoraciones no formales se incluyó la lectura, la repetición de palabras, la terminación de historias, entre otros, y el test de articulación como valoración formal. La función original de estos instrumentos es la de evaluar la articulación de cada fonema, la construcción de oraciones o la agrupación de categorías semánticas según la prueba usada, teniendo en esta ocasión el claro objetivo de recolectar las muestras de voz, en lugar de una identificación de los componentes del lenguaje oral anteriormente citados.

Inicialmente, se utilizó un test denominado Test de Articulación (Morales, 2016, pp. 59-66), el cual contempla la emisión de los sonidos iniciales, medios y finales de los fonemas del alfabeto en español. Posteriormente, se realizó la grabación de palabras por grupos semánticos con el propósito de contar con un banco de información por categorías de alto uso en el lenguaje infantil, tanto en actividades de vida diaria como dentro del currículo escolar propio de niveles iniciales (colores, animales, alimentos) con el propósito de generar en un futuro próximo voces artificiales para niños de dicho nivel educativo, y posteriormente ir evolucionando en currículos propios de niveles escolares y colegiales.

Se grabaron también palabras que contemplaran dichas categorías semánticas en singular y plural y finalmente la construcción de oraciones según temas de interés de las personas participantes,

con el fin de generar bancos relacionados a temas de alto agrado que por ende cuentan con impacto emocional y permite registrar las inflexiones en la voz. Dentro de estos temas libres se recolectó información sobre mitología, en especial hadas y ninfas; Pokémon; animales, en especial perros; y viajes o paseos.

Dentro de las estrategias usadas para la toma de las muestras de voz en la interacción con las personas menores de edad, es importante considerar que se requiere de conocimiento en el área de desarrollo infantil, no solo por la forma en que es propicio relacionarse con las personas participantes, sino también porque se debe tener conocimiento de los aspectos de percepción y atención, ya que por ejemplo el tiempo de atención sostenida de un niño de 6 años dista mucho del que puede presentar un niño de 12 años, entendiendo que "la atención sostenida es a la que tradicionalmente nos referimos cuando decimos que estamos <<concentrados>>, permite que la activación de los circuitos dirigidos hacia uno o varios estímulos, se mantenga durante un período de tiempo que permita su procesamiento adecuado" (Carazo, 2009, p. 60). Esta atención puede ir aumentando en tiempo de permanencia en un estímulo por un mayor lapso, conforme el individuo avanza en su neurodesarrollo y además se encuentra en estrecha relación con el impacto que ese estímulo tenga a nivel perceptual y emocional. De esta manera, se pueden aprovechar al máximo los insumos que se necesite recopilar.

Algunas de las estrategias con las que se trabajaron fueron el uso de material con contraste visual, el juego y competencias para la pronunciación de palabras, alternar la recolección de datos con espacios de interacción no formal o descansos, así como el uso de reforzamiento positivo de carácter verbal. La principal razón de esta forma de interacción con la población participante se fundamenta

en el hecho anteriormente señalado sobre la importancia de un estímulo que logre captar la atención mediante sus características perceptuales que permiten, como bien indica Ellis (2011, p. 217), el almacenaje necesario de la información en la memoria de trabajo. Estas características de los estímulos responden al tamaño, la intensidad, la novedad, la incongruencia, la emoción y el significado personal de la tarea, así como al material al que se expone, en este caso a las niñas y niños participantes. En la Figura 3 se ilustra el proceso de interacción durante las grabaciones, utilizando imágenes de frutas y diversos equipos para registrar el habla.

Figura 3: Interacción con un niño durante el proceso de grabación, utilizando imágenes que son señaladas para obtener la pronunciación de cada palabra

Fuente: elaboración propia

Previo a la integración de las grabaciones a los sistemas de generación de habla artificial se requiere un proceso de edición detallado para separar, de la totalidad de grabaciones, aquellos fragmentos de

habla que resulten útiles. Para esto, se requiere la escucha atenta y selección manual utilizando un programa de edición de audio.

Una vez finalizado este proceso se cuenta, entonces, con el primer elemento de la cadena de requerimientos para generar el habla. En etapas posteriores debe contarse con información complementaria que utilice la totalidad de fonemas de nuestro idioma, grabadas en condiciones de calidad por personas adultas. De esta manera, se hace posible aplicar los sistemas de síntesis de voz por adaptación para la pronunciación y la prosodia particular de Costa Rica.

Dado que se trata de una innovación tecnológica realizada en el país, es posible que en los procesos subsecuentes se requiera desarrollar herramientas de análisis, de experimentación y de evaluación propias.

5. APLICACIONES POTENCIALES

El uso del recurso obtenido en las grabaciones de voces infantiles costarricenses pretende potencializar sistemas de comunicación aumentativos y alternativos de costo accesible implementados en una variedad de dispositivos de fácil manejo para los usuarios, sus familias y el personal docente, así como para los profesionales de los servicios de apoyo que interactúan con el estudiantado en condición de discapacidad.

Entre los posibles usos de este banco de datos de voces infantiles se encuentran muchas posibilidades de aplicación, a saber:

1. Procesamiento de las señales del habla para la creación de voces que sean únicas y personalizadas y que, con investigaciones posteriores, puedan llegar incluso a evolucionar con el estudiante.

Esto lo consideramos como hipótesis a confirmar, la cual puede llegar a afectar positivamente el desarrollo emocional de un usuario infantil al cual como a sus pares etarios le va cambiando la voz mientras crece.

2. Dispositivos como sistemas de comunicación iniciales. En estos, la persona usuaria puede comenzar a indicar gustos y preferencias, emociones, frases cortas y por supuesto oraciones en una evolución similar a la que tiene el desarrollo del lenguaje oral. Es decir, se brinda la posibilidad de crear dispositivos con voz artificial infantil para las diversas etapas del desarrollo con patrones vocales propios de la etapa y con las características esperadas en cuanto al tipo de aproximación comunicativa oral según la edad.
3. Implementación en aplicaciones de aprendizaje de lenguaje y comunicación, semejantes a las disponibles en la actualidad, pero con voces personalizadas y de acuerdo con la edad del estudiante. Esto puede aplicar tanto a comunicación por medio de pictogramas, como aprendizaje de sílabas y otros elementos del lenguaje.

6. REFLEXIONES FINALES

Se ha presentado una visión general sobre los elementos, tanto teóricos como prácticos, para la creación de una voz artificial infantil y su contextualización con la primera experiencia de recolección de voces costarricenses para este fin. El banco o corpus lingüístico es un primer e importante paso para dotar mediante un sistema de comunicación aumentativa y alternativa al estudiantado en condición de discapacidad de un medio de acceso inclusivo a su propio entorno.

Dado que el proceso requiere una gran cantidad de información, también es necesario contar con suficientes datos de otras voces que sean coincidentes con la de las personas menores de edad en idioma y cercanas en su acento. Se trata aquí de novedades tecnológicas que han abierto las puertas para generar voces infantiles en contextos específicos.

Estas voces pueden ser únicas o personalizadas, y abren la puerta del acceso y la inclusión desde la misma interacción social en la familia, pasando por el proceso de aprendizaje del currículo educativo que les corresponde según su edad y características y ampliando su impacto al acceso e información de la sociedad a la que pertenece. Desde la comunidad universitaria, a través de la investigación, el desarrollo tecnológico y trabajo transdisciplinar se puede incidir en el sistema social donde se construyan espacios justos e inclusivos para todos los ciudadanos.

7. REFERENCIAS BIBLIOGRÁFICAS

Arnaiz, Pilar. (1996). Las escuelas son para todos. Siglo Cero.

Bonafonte, Antonio., Adell, Jordi., Esquerra, Ignacio., Gallego, Silvia., Moreno, Asunción., y Pérez, Javier. (2008). Corpus and Voices for Catalan Speech Synthesis. In LREC.

Boothalingam, Romani., Solomi, Sherlin., Gladston, Anushiya., Christina, Lilly., Vijayalakshmi, P., Thangavelu, Nagarajan., y Murthy, Herma. (2013). Development and evaluation of unit selection and HMM-based speech synthesis systems for Tamil. In 2013 National Conference on Communications (NCC), 1-5. IEEE.

Carazo, Viviana., y López, Luis. (2009). Aprendizaje, coevolución neuroambiental. Coordinación Educativa y Cultural Centroamericana (CECC). San José, Costa Rica.

Coto-Jiménez, Marvin., Goddard-Close, John. (2016a). LSTM deep neural networks postfiltering for improving the quality of synthetic voices. In Mexican Conference on Pattern Recognition, 280-289.

Coto-Jiménez, Marvin., Goddard-Close, John. (2016b). Hidden Markov Models for Artificial Voice Production and Accent Modification. In Ibero-American Conference on Artificial Intelligence, 415-426.

Coto Jiménez, Marvin y Morales Rodríguez, Maribel. (2020). Tecnologías del habla para la educación inclusiva. Revista Actualidades Investigativas en Educación, 20(1), 1-24. Doi.10.15517/aie.v20i1.40129.

Coto-Jiménez, Marvin., Goddard-Close, John. (2016c). Speech Synthesis Based on Hidden Markov Models and Deep Learning. Research in Computing Science, 112, 19-28.

Ellis, Jeanne. (2005). Aprendizaje humano. Editorial Person Prentice Hall.

Gonzalvo, Xavier., Iriondo, Ignasi., Socoró, Joan. C., Alıas, Francesc., y Monzo, Carlos. (2007). HMM-based Spanish speech synthesis using CBR as F0 estimator. ITRW on NOLISP, 788-793.

Hanzlíček, Zdenek. (2010). Czech HMM-based speech synthesis. In International Conference on Text, Speech and Dialogue, 291-298. Springer, Berlin, Heidelberg.

Hunt, Andrew., Black, Allan. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In 1996

IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings [Vol. 1, pp. 373-376]. IEEE.

Ipsic, Ivo., Martincic-Ipsic, Sandra. (2006). Croatian HMM-based speech synthesis. Journal of Computing and Information Technology, 14(4), 307-313.

Karabetsos, Sotiris., Tsiakoulis, Pirros., Chalamandaris, Almilios., y Raptis, Spyros. (2008). HMM-based speech synthesis for the Greek language. In International Conference on Text, Speech and Dialogue (pp. 349-356). Springer, Berlin, Heidelberg.

Krstulović, Sacha., Hunecke, Anna., y Schröder, Marc. (2007). An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements. In Eighth Annual Conference of the International Speech Communication Association.

Llanos, Leonardo. (2010). Tecnologías del habla y análisis de la voz. Aplicaciones en la enseñanza de la lengua. Laboratorio de lingüística informática, Universidad Autónoma de Madrid, 1-41.

Mascorro, Guillermo. (2013). Reconocimiento de voz basado en MFCC, SBC y Espectrogramas. Ingenius, 10, 12-20.

Mathiesen, Kay. (2008). Access to Information as a Human Right. https://ssrn.com/abstract=1264666

Morales, Maribel. (2016). De los sonidos a las palabras: Métodos y técnicas para la estimulación del lenguaje oral 1. EUNED.

Li, Qun., Russell, Martin. (2002). An analysis of the causes of increased error rates in children's speech recognition. In Seventh International Conference on Spoken Language Processing.

Programa de Naciones Unidad para el Desarrollo – Costa Rica. (2017). Visión 2030 Poblaciones Excluidas en Cota Rica: No dejar a nadie atrás. PNUD-Costa Rica. San José, Costa Rica.

Rabiner, Lawrence., Schafer, Ronald. (2007). Introduction to digital speech processing. Foundations and Trends® in Signal Processing, 1(1–2), 1-194.

Safavi, Saeid., Najafian, Maryam., Hanani, Abualsoud., Russell, Martin., Jancovic, Peter., Carey, Michael. (2016). Speaker recognition for children's speech. arXiv preprint arXiv:1609.07498.

Tokuda, Keiichi., Zen, Heiga., y Black, Allan. (2002). An HMM-based speech synthesis system applied to English. In IEEE Speech Synthesis Workshop (pp. 227-230).

Vega, Deliyore., y del Rocío, María. (2018). Comunicación alternativa y aumentativa: Acciones y reflexiones para romper el silencio en las aulas.

Yamagishi, Junichi., Kobayashi, Takao., Nakano, Yuji., Ogata, Katsumi., y Isogai, Juri. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Transactions on Audio, Speech, and Language Processing, 17(1), 66-83.

Zen, Hen., Tokuda, Keiichi., Black, Allan. (2009). Statistical parametric speech synthesis. speech communication, 51(11), 1039-1064.

Zen, Heiga., Nose, Takashi., Yamagishi, Junichi., Sako, Shinji., Masuko, Tasuko., Black, Allan, y Tokuda, Keiichi. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In SSW, 294-299.