Desarrollo de una base de datos para asistencia a la oralización de niños

Melisa Gisele Kuzman[1], Pablo Daniel Agüero[1], Juan Carlos Tulli[1], Esteban Gonzalez[1], Alejandro Uriz[2] y María Paula Cervellini[1]

[1] Laboratorio de Comunicaciones, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Juan B. Justo 4302, Mar del Plata, Argentina
[1] CONICET - Laboratorio de Comunicaciones, Facultad de Ingeniería, Universidad de Mar del Plata, Juan B. Justo 4302, Mar del Plata, Argentina

E-mail: {melisakuzman,pdaguero}@fi.mdp.edu.ar

Resumen.

En este artículo se presentan las ventajas que conllevaría obtener una base de datos de voz propia para el desarrollo de software de oralización de sordos e hipoacúsicos. La misma sería útil para diversas aplicaciones desarrolladas en la Facultad de Ingeniería, ya que actualmente solo se cuenta con una base de datos de voz con el dialecto ibérico. Para la creación de la base de datos de voz se propone grabar palabras y oraciones, y con la ayuda de un software, realizar una agrupación de los datos en difonemas. Esto dotaría de flexibilidad al sistema de reconocimiento de voz SPHINX, ya que se podrían reconocer palabras que no han sido grabadas. Se realizaron experimentos que arrojan como resultado que la base de datos española es insuficiente para cubrir las 10000 palabras más usadas en el idioma. Para lograr una mejor cobertura será necesario grabar 320 palabras. Como se puede observar, la base de datos de voz realizada será más completa. Al ser más flexible, podrá utilizarse para otros proyectos futuros.

1. Introducción

En la Argentina la deficiencia auditiva corresponde al 18 % de todas las discapacidades. Este dato ha sido publicado por la CAS (Confederación Argentina de Sordomudos)[1], quién mantiene contacto con miles de personas con déficit auditivo.

El período de los primeros años de vida es particularmente crítico para la adquisición del lenguaje y el desarrollo general del niño. Por ello, es importante conocer algunos de los medios para mejorar las posibilidades de comunicación de niños con hipoacusia en edad preescolar. Esto ayuda a que puedan desarrollar su lenguaje, ya que por lo general, estos déficits traen aparejado un retraso en este aspecto [2].

Los niños con hipoacusia tienen dificultades en su adaptación social. Por ello, resulta esencial que se los asista de forma prematura para revertir esta situación. Con este propósito se pueden crear aplicaciones para computadoras que, con un diseño adecuado, pueden ser un importante instrumento para lograrlo.

En la actualidad existen diversos programas que se utilizan para el entrenamiento y rehabilitación de la oralización en personas con déficit auditivo. Entre ellos, se puede destacar "Dr. Speech" [3], que es un software para adiestramiento y entrenamiento de la voz. También se destaca el proyecto desarrollado por la Universidad de Zaragoza, llamado "Alborada 13-A"

[4], un programa de reconocimiento y evaluación de la voz, que permite detectar errores en la pronunciación. Esta aplicación fue realizada con el objeto de ayudar al entrenamiento del habla en personas hipoacúsicas.

Estos programas han sido creados con el objeto de potenciar las posibilidades de comunicación de este sector de la población, permitiendo contribuir con la inserción social de personas que sufren deficiencias auditivas.

Los programas de reconocimiento y evaluación de la voz contienen tres componentes principales: una **base de datos** de voz para deducir modelos estadísticos de representación, un motor de **reconocimiento de voz**, y finalmente el software de **evaluación** y visualización de los resultados a través de una interfaz adecuada.

Un aspecto crítico de las bases de datos de voz que son utilizadas para los trabajos de investigación en la Facultad de Ingeniería de la Universidad Nacional de Mar del Plata, es su origen español. Esto genera problemas en el análisis acústico, ya que las formas lingüísticas y sus correspondientes pronunciaciones suelen diferir considerablemente con respecto a nuestro dialecto.

El primero de los objetivos que se plantea, es la realización de una base de datos en el dialecto español rioplatense. La misma estará dividida en dos grupos: el primero contiene información de niños con alguna deficiencia auditiva, y el segundo voces de niños que no poseen deficiencias.

Una vez obtenida la base de datos de voz, se pretende utilizarla para entrenar el motor de reconocimiento SPHINX [5] (una herramienta de reconocimiento de voz de Java). Esto permite evaluar posibles omisiones, inserciones, distorsiones o sustituciones de fonemas, fenómeno conocido clínicamente como disartria [2]. Finalmente, los datos y el motor de reconocimiento serán utilizados en el desarrollo de programas dedicados a personas hipoacúsicas, como es el caso del Sistema de Práctica de la Oralización [6].

El artículo se encuentra organizado de la siguiente manera. La Sección 2 describirá las distintas posibilidades de adquisición de datos y el análisis de los mismos. En la Sección 3 se mostrará la cobertura de difonemas que logra la base de datos de voz española, y cómo se podría mejorar. Finalmente, en la Sección 4 se detallarán las conclusiones del trabajo, realizando un análisis de las cuestiones a mejorar con respecto a la base de datos española.

2. Métodos

Para poder llevar a cabo el desarrollo de proyectos de software para personas hipoacúsicas, se va a utilizar un sistema de reconocimiento de voz [7], también conocido como **ASR**∗. Esta herramienta computacional es capaz de procesar una señal de voz y reconocer la información contenida convirtiéndola en texto. Debido a que estos sistemas utilizan modelos estadísticos, resulta necesario grabar cantidades adecuadas de datos para lograr así precisión en el modelado acústico.

2.1. Elección de los datos a grabar

Dependiendo de la aplicación, existen diferentes posibilidades de agrupar los fonemas [8], las cuales se mencionan a continuación:

- Fonemas: Es una de las unidades mínimas usadas en los sistemas de voz∗∗. Dentro del español existen 18 consonantes y 5 vocales [9]. Es este trabajo se utilizará un inventario de 24 fonemas y 7 alófonos, definidos por Alarcos (1950) [10]. Debido a que estas unidades están sometidas a variaciones contextuales, solamente podrán ser grabadas en el contexto de palabras u oraciones.

∗ Siglas que provienen de Automatic Speech Recognition

∗∗El fonema es la unidad mínima desprovista de sentido delimitable en la cadena hablada.

- Difonemas: Son las unidades que se consideran coarticuladas, ya que dependen del contexto que se encuentre a sus lados (derecho o izquierdo). En otras palabras, es la unión de dos fonemas, y constituyen como máximo un total de 961 difonemas, provenientes del cálculo 31x31 (deberán ser eliminados de la lista aquellos que nunca aparecen en un idioma).
- Trifonemas: Estos tienen en cuenta las coarticulaciones generadas a partir de los contextos derecho e izquierdo. Las combinaciones posibles en los trifonemas resulta significativamente mayor a los casos mencionados anteriormente, alcanzando un máximo de 29791 posibilidades (31x31x31), de las cuales deberán ser eliminadas aquellas que no existen en un idioma.
- Palabras y oraciones: se realiza la grabación directa, con lo cual se obtiene mayor naturalidad en la voz.

Para el desarrollo de la base de datos de voz se decide dividir a las palabras y oraciones grabadas en difonemas. Estas unidades dotarán al sistema de una flexibilidad que resultaría imposible con las otras, necesitando un número singificativamente más reducido de grabaciones. Si se realizan bases de datos específicas para distintas aplicaciones, con las palabras o frases para cada caso particular, se elimina la componente de flexibilidad de las mismas, y no haría posible la reutilización de las bases de datos entre aplicaciones existentes o futuras. Esto generaría un aumento significativo del trabajo, al ser necesario repetir todo el proceso de creación de una nueva base de datos. Cabe destacar que las palabras y oraciones serán elegidas cuidadosamente, para obtener la mejor cobertura de los difonemas que existen en el idioma.

2.2. Condiciones de grabación

Las grabaciones se van a realizar con WaveSurfer [11], una herramienta de código abierto para la manipulación y visualización del sonido. La misma dispone de una interfaz lógica y sencilla, que provee de funcionalidad de una forma intuitiva y que puede ser adaptada a diferentes tareas. Puede ser utilizada para la investigación del habla y la educación. Las aplicaciones del programa que resultan de interés para el trabajo son la grabación, el análisis del sonido y del habla, y la transcripción fonética. Además es una herramienta que permite trabajar con diferentes tipos de formatos como WAV y NIST/Sphere.

Para la grabación se utilizará un micrófono dinámico (SKP Pro-92 XRL, unidireccional, con una respuesta en frecuencia entre los 50-16000Hz, impedancia de salida de 600 Ohms y una sensibilidad de -52dBV $\pm$ 3dB).

Además, la grabación se efectuará con la participación de un especialista que instruya a los niños. Esto se debe a que se busca generar un ambiente propicio para el desenvolvimiento de los mismos, lo cual se reflejará en una mayor naturalidad en su habla [2].

Para la grabación de la base de datos de voz se tienen en cuenta un conjunto de especificaciones. Algunas de estas características son:

- Relación señal a ruido.
- Frecuencia de muestreo utilizada para la grabación de la voz.
- Cantidad de bits utilizados para muestrear y almacenar la información de voz.
- Tiempo de reverberancia.
- Aprovechamiento de los bits de información.
- Presencia de saturaciones (principalmente en plosivas).

A continuación son explicados cada uno de los los parámetros, y cuáles son las medidas que se tomarán para mejorarlos.

Relación señal a ruido. El recinto donde se harán las grabaciones deberá ser pequeño, y poseer irregularidades en las paredes para evitar inconvenientes debidos a la reflexión de las ondas sonoras. Para mejorar la **relación señal a ruido**, la grabación se realizará con un ordenador portátil, que se encontrará trabajando con batería (sin conexión a la línea eléctrica), para evitar interferencias de señales de baja frecuencia. Por otra parte, se tomará la precaución de usar un cable apropiado para no generar atenuaciones elevadas u otro tipo de interferencias. Se alejará al hablante de otras fuentes de ruido externo, tales como los ventiladores en las computadoras, aires acondicionados o teléfonos.

La relación señal a ruido es un parámetro importante en cualquier sistema, ya que indica una relación entre la señal bajo estudio, y ruidos provenientes de otras fuentes.

Frecuencia de muestreo y bits útiles de información. Al utilizar WaveSurfer se puede seleccionar la frecuencia de muestreo de la grabación y el número de bits para almacenar la información. Lo ideal es que la frecuencia de muestreo y la cantidad de bits con los que se guarda la información sea lo más elevada posible, ya que mejoraría la calidad de los datos que se obtienen. Sin embargo, estos parámetros se encuentran limitados por la placa de sonido instalada en la computadora con la que se realice la grabación. Los valores que se considerarán para realizar las primeras grabaciones serán los tomados del desarrollo del programa Vocaliza [4]. La frecuencia de muestreo resulta entonces de 16Khz y 16 bits para discretizar la información.

Tiempo de reverberancia. El tiempo de reverberancia es fundamental para generar una grabación de buena calidad. Este se define como el tiempo en el que la energía de la señal tarda en decaer 60dB, luego de finalizada la oralización. Es decir, se considera que las reflexiones finalizan cuando la intensidad con la que se perciben es una millonésima de su valor original. Minimizar este tiempo ayudará a que los datos obtenidos sean considerados fieles.

Presencia de saturaciones y aprovechamiento de los bits de información. En las grabaciones suelen producirse sonidos no deseables, como los sonidos plosivos. Estos se generan cuando el tracto vocal se cierra en algún punto, lo que causa que el aire se acumule para después salir expulsado repentinamente. Esta expulsión de aire se caracteriza por estar precedida de un silencio. Si el micrófono es ubicado directamente enfrente de la boca del hablante, entonces es muy susceptible a que las ráfagas de aire ocasionadas por los sonidos plosivos produzcan saturaciones. En estos casos se repetirán las grabaciones, intentando lograr un buen compromiso entre el aprovechamiento de los bits de información y la eliminación de saturaciones.

2.3. El motor de reconocimiento de voz

El objetivo de un sistema de reconocimiento automático del habla consiste en obtener una secuencia de palabras (o etiquetas) que son la representación textual de una señal acústica. Dicha tarea no es trivial, ya que en muchas circunstancias la señal acústica contiene no solamente palabras y pausas, sino también disfluencias, sonidos ambientales, ruidos articulatorios (labios, respiración), etc.

Para realizar esta tarea, los sistemas de reconocimiento automático del habla utilizan una serie de herramientas de modelado estadístico y decodificación. En ellas se aplican una serie de simplificaciones en el enfoque para obtener una solución implementable.

Una de las primeras suposiciones es que el vocabulario a reconocer estará limitado. Su tamaño puede ir de unas pocas palabras (por ejemplo: sistemas de reconocimiento de órdenes verbales) a decenas de miles de palabras (sistemas de reconocimiento de gran vocabulario). No es posible para un ASR reconocer palabras desconocidas, debido a que no le sería posible encontrar las fronteras de las mismas. Por ejemplo, la oración *"la casa de la pradera está habitada por un*

ermitaño" se podría pronunciar como *"lacasadelapradera estáhabitada porunermitaño"*. Como se puede observar, existen pausas después de las palabras *pradera* y *habitada*. Sin embargo, el resto de las palabras son pronunciadas sin dar ningún indicio del fin de una y del comienzo de la otra. Por lo tanto, es necesario conocer las palabras para poder encontrar las fronteras de las mismas.

Los sistemas de reconocimiento automático del habla que se desarrollan hoy en día asumen que es posible hacer su tarea basándose en un modelado estadístico de la señal acústica (modelado de la generación acústica de las palabras del idioma) y el lenguaje (modelado de la construcción del discurso del idioma utilizando palabras) [12]. Para ello se utilizan datos de entrenamiento (señal acústica, y transcripción ortográfica y fonética de la misma) para obtener los parámetros de los modelos estadísticos. La cantidad de datos de entrenamiento esta acotada por el volumen del corpus disponible, y por ello es necesario usar modelos cuyos parámetros puedan ser obtenidos en forma confiable teniendo en cuenta esta limitación.

El uso de la tecnología de reconocimiento automático del habla para la detección de errores articulatorios, impone un condicionamiento a su funcionamiento. En este caso el sistema de reconocimiento no deberá deducir que es lo que se pronunció, sino que deberá analizar aquello que le fue encomendado al usuario pronunciar. En estas condiciones es posible mejorar la capacidad de análisis de estos sistemas, incorporando las posibles pronunciaciones alternativas que puede producir el usuario en el momento del análisis.

Con el objeto de ser consistentes en el uso de JAVA en el desarrollo de los sistemas de apoyo a la oralización de personas sordas e hipoacúsicas, se decidió utilizar el sistema de reconocimiento de voz SPHINX para la segmentación del audio en fonemas y el análisis de los mismos.

Sphinx-4 es un sistema de reconocimiento de voz muy actualizado, que se encuentra escrito enteramente en el lenguaje de programación JAVA. El sistema fue creado mediante una colaboración conjunta entre el grupo Sphinx de la Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs (MERL), y Hewlett Packard (HP), con contribuciones de la University of California de Santa Cruz (UCSC) y el Massachusetts Institute of Technology (MIT).

3. Experimentos

Con el objeto de analizar la capacidad de cobertura de fonemas de la base de datos Alborada, se realizó la transcripción fonética automática de las 57 palabras diferentes presentes en este corpus. Para ello se utilizó un software que convierte la transcripción ortográfica de una palabra en su transcripción fonética. Luego, dicha transcripción fonética se procesó para obtener la transcripción de las palabras en difonemas. El número total de difonemas presentes en la base de datos Alborada es de 147. En el Cuadro 1, puede observarse la transcripción fonética realizada por el software, a un conjunto de palabras.

El análisis de la capacidad de cobertura de dichos difonemas con respecto a las palabras del español se hizo utilizando el corpus CREA escrito. El Corpus de referencia del español actual (CREA) es un conjunto de textos de diversa procedencia, almacenados en soporte informático, del que es posible extraer información para estudiar las palabras, sus significados y sus contextos [13]. La Real Academia Española explica que el CREA es un corpus representativo del estado actual de la lengua, de manera que los materiales que lo integran han sido seleccionados de acuerdo con los parámetros habituales.

Las 10000 palabras más frecuentes del corpus CREA fueron transcritas a su representación en difonemas. El número total de difonemas presentes en este subconjunto del corpus es de 440. En consecuencia, los 147 difonemas presentes en la base de datos Alborada no son suficientes para analizar la pronunciación de este subconjunto.

Con el objeto de obtener un listado de palabras que permita cubrir este número de difonemas se utilizó un algoritmo *greedy* de selección de palabras para minimizar el número de palabras

Cuadro 1. Representación de palabras mediante difonemas

Palabra	Fonemas
boca	_+b b+o o+k k+a a+_
bruja	_+b b+r r+u u+x x+a a+_
cabra	_+k k+a a+B B+r r+a a+_
campana	_+k k+a a+m m+p p+a a+n n+a a+_
casa	_+k k+a a+s s+a a+_
caramelo	_+k k+a a+r r+a a+m m+e e+l l+o o+_
casa	_+k k+a a+s s+a a+_
clavo	_+k k+l l+a a+B B+o o+_
cuchara	_+k k+u u+tS tS+a a+r r+a a+_
dedo	_+d d+e e+D D+o o+_
ducha	_+d d+u u+tS tS+a a+_
escoba	_+e e+s s+k k+o o+B B+a a+_
flan	_+f f+l l+a a+n n+_
fresa	_+f f+r r+e e+s s+a a+_
fuma	_+f f+u u+m m+a a+_
gafas	_+g g+a a+f f+a a+s s+_
globo	_+g g+l l+o o+B B+o o+_
gorro	_+g g+o o+rr rr+o o+_
grifo	_+g g+r r+i i+f f+o o+_
indio	_+i i+n n+d d+j j+o o+_
jarra	_+x x+a a+rr rr+a a+_
jaula	_+x x+a a+w w+l l+a a+_
lápiz	_+l l+a a+p p+i i+T T+_

necesarias para tener una cobertura total de los 440 difonemas. El número de palabras resultantes es de 320.

Aplicando el mismo procedimiento fue posible hallar la totalidad de los trifonemas en ese subconjunto de palabras del idioma español. El mismo arroja un resultado de 2919 trifonemas, lo que conllevaría a la necesidad de grabar 1778 palabras.

Por lo tanto, el uso de difonemas implica un buen compromiso entre precisión de los modelos estadísticos de los fonemas y sus contextos, y el volumen de datos que resulta necesario grabar para lograr una buena cobertura.

4. Conclusiones

En el presente artículo se han abordado las consecuencias que sufren las personas que poseen disminución auditiva. Además, se han nombrado algunas aplicaciones en el campo de la computación, que han sido creadas con el objeto de brindar nuevas herramientas para el entrenamiento y mejora del lenguaje de personas con hipoacusia. También se destaca la importancia de una base de datos de voz confiable en los sistemas.

La base de datos de voz que se posee en la Facultad de Ingeniería es de origen español. El principal problema que surge es la imposibilidad de implementación de las aplicaciones creadas. Esto se debe a las diferentes formas lingüísticas que poseemos con respecto a los españoles. Obtener una base de datos de voz propia, permitirá mejorar proyectos desarrollados en la Facultad de Ingeniería y poder llevarlos a la práctica, como podría ser el caso del Sistema

de Práctica de la Oralización (SPO) [6, 14, 15, 16].

En el trabajo se plantea como objetivo crear una base de datos de voz en español rioplatense, y mejorar la calidad con respecto a la existente de origen español ibérico. Por otra parte, se pretende utilizar un método diferente de selección del contenido fonético, que dotaría al sistema de reconocimiento de voz de una mayor flexibilidad.

Los experimentos realizados arrojan como resultado que la base de datos española existente es insuficiente para cubrir las 10000 palabras más usadas en el idioma español. Para lograr una mejor cobertura será necesario grabar 320 palabras. Esto posibilitaría crear y desarrollar nuevos proyectos, usando una base de datos de voz confiable y versátil.

Referencias

[1] "http://www.cas.org.ar," Tech. Rep.
[2] R. L. Belloto, *Voz y pronunciación en el niño discapacitado auditivo.* Ediciones ARES, 1974.
[3] "http://www.drspeech.com," Tech. Rep.
[4] O. Saz, E. Lleida, C. Vaquero, and W.-R. Rodríguez, "The alborada-i3a corpus of disordered speech," Tech. Rep., 2010.
[5] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," CMU, Tech. Rep., 2004.
[6] J. M. Garin, P. D. Agüero, J. C. Tulli, E. L. Gonzalez, and A. J. Uriz, "Spo: una ayuda en java para la oralización de hipoacúsicos," Tech. Rep., 2010.
[7] H. van den Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa, "Tc-star: New language resources for asr and slt purposes," in *Proceedings LREC 2006*, 2006, pp. 2570–2573.
[8] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing:A Guide to Theory, Algorithm and System Development.* Prentice Hall, 2001.
[9] J. Llisterri and J. B. Mariño, "Spanish adaptation of sampa and automatic phonetic transcription," Tech. Rep., 1993.
[10] E. Alarcos, *Fonología española.* Gredos, 1950.
[11] "http://www.speech.kth.se/wavesurver/man.html," Tech. Rep.
[12] S.-C. Yin, R. Rose, O. Saz, and E. Lleida, "Verifying pronunciation accuracy from speakers with neuromuscular disorder," in *In INTERSPEECH-2008*, 2008, pp. 2218–2221.
[13] "http://www.rae.es," Tech. Rep.
[14] J. M. Garín, P. D. Agüero, J. C. Tulli, and E. L. Gonzalez, "Sistema de práctica de la oralización en plataforma java," in *SICA*, Noviembre 2009.
[15] J. M. Garin, P. D. Agüero, and J. C. Tulli, "Nuevos aportes al entrenamiento de personas hipoacúsicas sistema de entrenamiento para personas hipoacúsicas en plataforma java," in *TISE*, Diciembre 2008.
[16] J. M. Garin, P. D. Agüero, and J. C. Tulli, "Nuevos horizontes - aportes al entrenamiento de personas hipoacúsicas," in *SICA*, Septiembre 2008.