Universidad de Zaragoza
Escuela de Ingeniería y Arquitectura

Proyecto Fin de Carrera

Diseño de una plataforma web de docencia distribuida aplicada a la logopedia y comunicación en educación especial

Ingeniería de Telecomunicación
Especialidad: Comunicaciones

AUTOR: José Manuel Escartín Villellas

DIRECTOR: Dr. Eduardo Lleida Solano

Departamento de Ingeniería Electrónica y Comunicaciones
Área de Teoría de la Señal y Comunicaciones

Zaragoza, Agosto de 2011

Agradecimientos

Me gustaría mostrar mi agradecimiento a
todas las personas que han colaborado
y me han ayudado a lo largo de la
realización de este proyecto.

Al colegio Alborada y en conreto a
José Manuel por el interés mostrado
y sus opiniones sobre la aplicación.

A Antonio Miguel Artiaga por su
dedicación y ayuda en los
comienzos del proyecto.

A José Vicente Ruiz por su
tiempo y soluciones aportadas.

A José Enrique García, por su
disponibilidad a la hora de
ayudar de forma desinteresada
en todo momento.

A Eduardo LLeida por su disposición
personal y dedicación a la dirección
de este proyecto.

A todos, muchas gracias.

Diseño de una plataforma web de docencia distribuida aplicada a la logopedia y comunicación en educación especial

RESUMEN

Este proyecto se enmarca dentro del Grupo de Tecnologías de las Comunicaciones (GTC) en colaboración con Colegio Público de Educación Especial Alborada (CPEE Alborada) y sigue la línea de proyectos anteriores que ya desarrollaron este tipo de herramientas.

El principal objetivo ha sido el desarrollo de la infraestructura de una plataforma web de docencia distribuida aplicada a la logopedia que finalmente presenta una estructura donde los centros son independientes en funcionamiento. Esta plataforma está dotada de una gestión de perfiles de usuario para profesores, alumnos, administradores y padres, garantizando la confidencialidad, la seguridad de los datos y la eficacia de los métodos de procesado de señal que estarán distribuidos entre cliente y servidor.

Una parte importante de la arquitectura del sistema es que permitirá a dos clientes realizar concurrentemente una misma actividad, de forma que se potencie la participación en grupo de los alumnos en algunas actividades que serán cooperativas. Además incluye la adaptación de las actividades que ya estaban presentes en "Vocaliza".

Se ha implementado un servidor para un centro de educación de pruebas basado en una máquina virtual. Que podrá ser instalado en todos aquellos centros que lo requieran.

También se ha habilitado una aplicación de pruebas en la red, para que los interesados puedan probar la plataforma antes de su instalación.

Otro objetivo ha sido el de mejorar el reconocimiento automático del habla en el caso de los niños, ya que no se dispone de suficientes datos (grabaciones) para generar modelos adaptados a ellos.

Para cumplir este objetivo se ha implementado en el reconocedor del grupo método de Normalización de la Longitud del Tracto Vocal del locutor. Esto mitigará las diferencias fisiológicas entre hablantes adultos y niños reduciendo en parte la tasa de error en el reconocimiento automático del habla en el caso de los niños.

Índice general

Índice de figuras

Capítulo 1

Introducción

1.1 Objetivos

El presente proyecto se enmarca en la colaboración que el Grupo de Tecnologías de las Comunicaciones (GTC) y el Instituto de Investigación en Ingeniería de Aragón (I3A) están llevando a cabo con el Colegio Público de Educación Especial Alborada (CPEE Alborada), y tiene el objetivo fundamental de poner las tecnologías del habla al servicio de aquellas personas que sufren patologías del lenguaje, a través de sistemas de ayuda controlados por voz y de sistemas de asistencia a la logopedia.

Los sistemas de ayuda controlados por voz para personas con patologías del lenguaje conllevan un proceso de adaptación al locutor, esto es, la creación de un modelo acústico adaptado al sujeto que presenta una patología del lenguaje, que permita al reconocedor discernir con mayor precisión las emisiones fonéticas del sujeto. Por tanto, era necesario estudiar el funcionamiento de las técnicas de adaptación habitualmente utilizadas con personas sin problemas de dicción, comprobando si son útiles para personas con patologías en el habla. Esto ya ha sido estudiado en otros proyectos como "Vocaliza"[1] [2], ahora nos centraremos en la mejora para el caso de los niños, en los cuales se presenta una dificultad añadida debido a la diferencia en las características fisiológicas del locutor.

Por ello, el presente proyecto consiste en el desarrollo de una aplicación web cuyo objetivo fundamental es el de servir de apoyo a la logopedia, utilizando las tecnologías del habla para ayudar a las personas con patologías del lenguaje a mejorar su capacidad del habla, permitiendo de forma adicional la adaptación al tracto vocal del locutor, mejorando el reconocimiento en niños.

Este proyecto sigue la línea del proyecto "Vocaliza"[1] [2], una aplicación que permite trabajar con los niveles fonológico, semántico y sintáctico del lenguaje.

Otro fruto de esta colaboración es "Prelingua"[3], herramienta que posibilita el trabajo del prelenguaje, incluyendo presencia/ausencia de voz, control del tono o de la intensidad, y vocalización. El cual se incluirá en la aplicación en un futuro.

De la misma colaboración surgió el proyecto "Cuéntame"[4], en este caso se requería trabajar ciertos aspectos complejos de niveles como el sintáctico y el semántico, que ya cubría Vocaliza, así como comenzar con el nivel superior, el pragmático.

Además siguiendo en esta línea de trabajo, ViVoLab desarrolla un sistema de accesibilidad a páginas web para ciegos [5], que se basa en la síntesis por voz de elementos claves

de la página que son navegados mediante teclado y en comandos orales para avanzar más rápidamente por la página. Un ejemplo de implantación lo podemos ver en Aragón Radio 2 [6].

En el caso que nos ocupa, la aplicación, de nombre "Vocaliza 2.0", pretende proveer tanto a los logopedas como a las personas con patologías en el habla de una herramienta para la mejora del habla, de carácter general, es decir, útil para el máximo número de patologías del lenguaje posible, sirviendo como una potente base para atacar cada caso particular.

Dado el presente marco, este PFC se encuadra en dos líneas diferentes de trabajo, cada una con diversos objetivos:

1. La difusión del proyecto:

   (a) Creación de una aplicación web funcional y accesible.

   (b) Creación de una versión distribuida, para su instalación en las redes locales de los centros que la requieran.

2. La adaptación del reconocedor automático del habla a la longitud del tracto vocal del hablante.

   (a) Calibración de la herramienta para cada usuario, mediante la normalización de la longitud del tracto vocal del locutor

   (b) Realizar las modificaciones requeridas en el reconocedor automático del habla para los nuevos requisitos.

1.2 Estado del arte

En los últimos años, los sistemas basados en tecnologías del habla han conseguido numerosos avances en situaciones de entorno controlado y locutor cooperativo. De esta forma, se han ido introduciendo en la vida cotidiana, sistemas de domótica, sistemas de manos libres para el automóvil, navegación por la red, manejo de los dispositivos móviles, etc. Pero donde está siendo más utilizado es en aplicaciones telefónicas: agencias de viajes, atención al cliente, información etc. La mejoría de estos sistemas de reconocimiento del habla han ido aumentando y su eficacia cada vez es mayor. Actualmente, se trabaja en la expansión de las capacidades de estas tecnologías a otros colectivos y necesidades, como la ayuda a la discapacidad, la logopedia o la enseñanza de idiomas.

En el presente proyecto se aborda la problemática en el diseño de una herramienta web dirigida a la logopedia a nivel de educación especial, a diferencia de otros sistemas, nos interesa que este sea adaptable a los niños, ya que es el público al que está dirigida la herramienta. En consecuencia, habrá que tener en cuenta la estructura morfológica del locutor para poder mejorar los resultados conseguidos con los reconocedores actuales en los niños.

1.2.1 Tecnologías del habla como apoyo a la logopedia

El Reconocimiento Automático del Habla y otras tecnologías del habla han alcanzado buenos resultados en situaciones controladas (ambiente sin ruido, pequeño vocabulario, usuario cooperativo...), pero las prestaciones caen rápidamente cuando se sale de esta situación. Ese es el caso cuando se trabaja con personas que presentan patologías del lenguaje.

El desarrollo de sistemas de logopedia asistida por ordenador puede ayudar a mejorar los problemas en el habla y el desarrollo del lenguaje de pacientes de trastornos y alteraciones del habla. Este tipo de sistemas ha sido siempre muy demandado por parte de profesionales y educadores en el ámbito de la Educación Especial. No obstante, el desarrollo de sistemas comerciales no ha cumplido las expectativas.

Además, existía un problema añadido, y era la inexistencia de herramientas en lengua castellana. En este aspecto, el proyecto "Comunica" fue pionero en el año 2006 y con el lanzamiento de la herramienta "Vocaliza", una aplicación en castellano que permite trabajar con los niveles fonológico, semántico y sintáctico del lenguaje, y que además está ideada para poder trabajar con personas con distintas alteraciones en el habla.

El proyecto "Vocaliza" ha tenido mucho éxito en la comunidad educativa de habla hispana, y por eso se sigue trabajando en su mejora gracias, en gran medida, a la colaboración de los usuarios.

Sin embargo, los resultados del reconocimiento de habla en niños están por debajo de los ofrecidos para los locutores adultos, por lo tanto en "Vocaliza 2.0" se quiere hacer especial hincapié en la adaptación a las características fisiológicas del hablante, con el objetivo de ofrecer mejoras en este terreno.

1.3 Metodología de trabajo

La realización del presente proyecto ha conllevado diversas fases, que incluyen la especificación de los requisitos de la aplicación web y el reconocedor, el desarrollo de las mismas, y el análisis de la mejora de los resultados obtenidos con la normalización de la longitud del tracto vocal del locutor.

1.3.1 Definición de requisitos de Vocaliza 2.0

La primera fase del proyecto consistió en una serie de reuniones llevadas a cabo entre el GTC y el CPEE Alborada, en las que se definieron los requisitos de la aplicación a desarrollar. En dichas reuniones se acordaron los siguientes puntos:

- La aplicación debe ser de libre distribución.
- La aplicación debe tener fundamentalmente fines logopédicos. Esto es, debe ayudar al sujeto que padece patologías en el habla a mejorar su capacidad de comunicación.
- La aplicación debe ser versátil y flexible. Debe ser útil para tratar el máximo número de casos y patologías posibles, y debe proveer a los logopedas y educadores que la usen de una herramienta con amplias y a la vez sencillas opciones de configuración que permitan adecuarla a cada caso.

- Diseño visualmente agradable, pero a su vez sencillo y accesible.
- Debido al ámbito de posible aplicación, era necesario incluir un sistema de gestión de usuarios y de adaptación al locutor.
- Aprovechando la creación de la plataforma web, creación de algún juego colaborativo para la interacción entre usuarios.

1.3.2 Diseño y desarrollo de la aplicación "Vocaliza 2.0"

Definidos los requisitos, se procedió al diseño y desarrollo de la aplicación. Esta fue, con diferencia, la fase de mayor duración del proyecto. La aplicación fue realizada con la herramienta SpringSource Tool Suite, con Groovy and Grails, haciendo uso de tecnología libre y avanzada del mercado para el desarrollo de aplicaciones web.

Durante el proceso de diseño y desarrollo de la aplicación no se dejaron de celebrar reuniones periódicas en el CPEE Alborada, en las que se mostraba la evolución en el diseño de la aplicación para obtener una realimentación con la que ir adaptando el entorno gráfico y las características de la aplicación a las exigencias de futuros usuarios: los educadores, logopedas y sujetos con discapacidades en el habla. Asimismo, esta realimentación fue de gran ayuda a la hora de la corrección de errores.

Las características de la herramienta, así como algunos aspectos importantes de su diseño y desarrollo se exponen en mayor profundidad en el capítulo 2.

1.3.3 Análisis de prestaciones de las tecnologías del habla utilizadas

Para finalizar este proyecto, se decidió realizar una serie de cambios en el reconocedor del grupo para mejorar los resultados obtenidos en niños hasta ahora. Actualmente no se disponen de suficientes datos del habla de niños para poder crear modelos adaptados para el reconocimiento automático habla para estos casos, por lo tanto, se propone la utilización de otro método para mejorar el reconocimiento, el método de la Normalización de la Longitud del Tracto Vocal del hablante (VTLN).

Los detalles sobre la implantación este método se pueden encontrar en el capítulo 3.

Capítulo 2

La aplicación Vocaliza 2.0

2.1 Vocaliza: Objetivos y cuestiones a resolver

Como ya se comentó en el capítulo 1, la aplicación "Vocaliza 2.0" surge a partir de la aplicación "Vocaliza"[1] [2], una aplicación que permite trabajar con los niveles fonológico, semántico y sintáctico del lenguaje.

Dado el éxito de este proyecto anterior, que se basaba en pequeñas actividades o juegos, se optó por continuar con esa filosofía, incluyendo las actividades en la nueva aplicación web, mejorándolas con la realimentación obtenida durante su uso.

Puesto que la mayoría de los pacientes que van a tratarse con la aplicación son niños, ésta ha sido diseñada de forma similar a un juego, en el que el usuario debe intentar pronunciar adecuadamente la palabra o frase apropiada en cada momento, en función de los estímulos audiovisuales que la aplicación provee. De ésta forma la aplicación capta la atención de los sujetos que vayan a ser tratados, resultando divertido para ellos y evita al logopeda parte del esfuerzo que supone estimular a los sujetos para que realicen los ejercicios adecuados.

Asimismo, existían los requisitos de ser una aplicación libre y altamente configurable para poder cubrir el mayor número de situaciones posibles a la hora de trabajar con ella.

El desarrollo de la aplicación se ha llevado a cabo en Grails que es un framework para aplicaciones web libre desarrollado sobre el lenguaje de programación Groovy (el cual a su vez se basa en la Java platform). Grails pretende ser un framework altamente productivo basado en la "convención sobre configuración" y la política "DRY" (don't repeat yourself).

El uso de las tecnologías HTML y CSS también ha sido extensivo. Se ha intentado solucionar en la medida de lo posible el temido "cross-browser" o problemas que dan los diferentes navegadores del mercado, sobre todo los no actualizados.

Por último también se ha hecho uso de las tecnologías javascript, con librerías como jQuery principalmente, para la gestión de eventos y del potente AJAX para aumentar la interactividad, usabilidad y velocidad de la aplicación.

Como servidor se utilizará Apache Tomcat, desarrollado por Apache Software Fundation, también de libre distribución.

Por otra parte, al ser una aplicación en red, existía miedo ante la confidencialidad de los datos que ésta trataría, lo cual, sumado al hecho del coste de los recursos necesarios para

su mantenimiento, hizo que finalmente se hiciera necesario el desarrollo de una versión que sería desplegada en la red local de cada centro en lugar de ser una aplicación en la nube como habría gustado en un principio.

Esto hizo replantear la situación y comenzar la búsqueda de nuevas soluciones para nuestro caso. Tras barajar otras soluciones que fueron descartadas para no depender de los sistemas operativos locales o para no tener problemas con otras instalaciones se llegó a la siguiente solución, la máquina virtual.

2.2 La Máquina Virtual

2.2.1 OracleVM VirtualBox

Ante los nuevos hechos y tras algunas reuniones, se ideó la creación de una máquina virtual, en la cual se encontraría todo lo necesario para el uso de la aplicación y quedaría instalado como una máquina más dentro de la red local del centro. Para ello se utilizaría como herramienta principal Oracle VM VirtualBox

Oracle VM VirtualBox es un software de virtualización para arquitecturas x86, creado originalmente por la empresa alemana innotek GmbH. Actualmente es desarrollado por Oracle Corporation como parte de su familia de productos de virtualización. Por medio de esta aplicación es posible instalar sistemas operativos adicionales, conocidos como "sistemas invitados", dentro de otro sistema operativo "anfitrión", cada uno con su propio ambiente virtual.

En nuestro caso dado que uno de los requisitos es que la aplicación fuera de libre distribución se optó como sistema invitado por una distribución Linux, más concretamente Ubuntu, debido a su interfaz gráfica más intuitiva.

Actualmente existe la versión propietaria Oracle VM VirtualBox, que es gratuita únicamente bajo uso personal o de evaluación, y está sujeta a la licencia de "Uso Personal y de Evaluación VirtualBox" (VirtualBox Personal Use and Evaluation License o PUEL) y la versión Open Source, VirtualBox OSE, que es software libre, sujeta a la licencia GPL.

En cuanto a la emulación de hardware, los discos duros de los sistemas invitados son almacenados en los sistemas anfitriones como archivos individuales en un contenedor llamado Virtual Disk Image, incompatible con los demás software de virtualización.

Gracias a esta herramienta conseguimos instalar nuestra máquina virtual como si se tratase de un equipo más de una red local. Cada centro tendrá su servidor, con su base de datos separada del resto de centros.

Las instrucciones para la instalación de la máquina virtual para la creación del servidor de la aplicación están expuestas en el anexo B.

2.3 Estructura del servidor

En la figura 2.1 podemos ver cómo queda la estructura del servidor.

Como vemos, la máquina virtual, que actúa como servidor, estaría dividido en tres partes principales, la aplicación propiamente dicha, una base de datos MySQL donde se

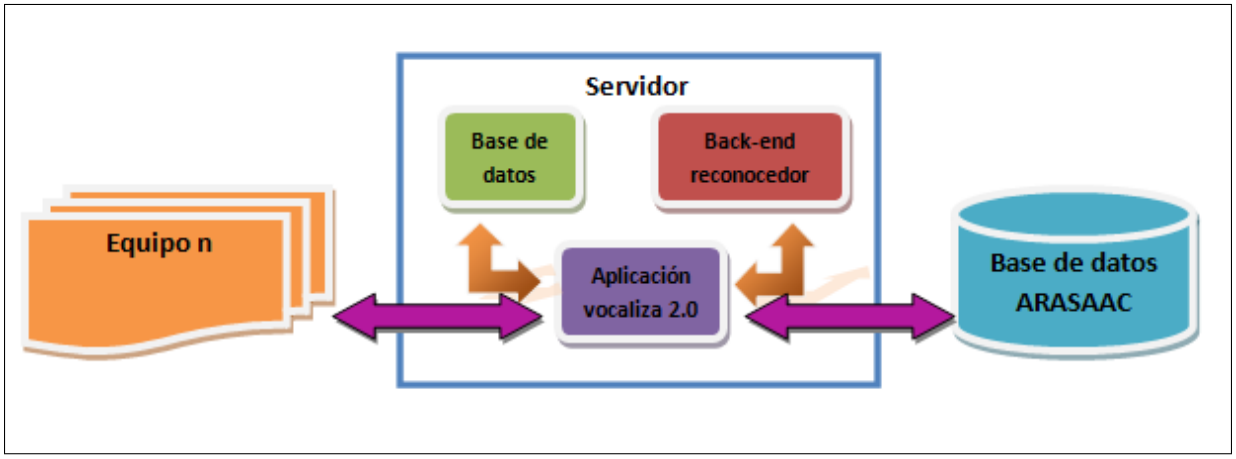

Figura 2.1: Esquema del servidor

almacenan todos los datos de la aplicación, y el back-end del reconocedor de voz, donde se ejecutan las operaciones necesarias para el reconocimiento que no se ejecutan en cliente.

La aplicación también está conectada a la base de datos de pictogramas de ARASAAC de manera que podamos encontrar nuevos pictogramas conforme ésta se actualiza.

2.4 Estructura de la aplicación

2.4.1 Diagrama de bloques

Las distintas funcionalidades de la aplicación se pueden agrupar en diversos bloques, que se relacionan entre sí como se muestra en la figura 2.2.

Cada bloque de la Figura 2.1 agrupa una o varias funcionalidades de la aplicación e intercambia información con otros bloques mediante interfaces representados por flechas de colores. El color de cada flecha indica que tipo de información se intercambia, tal y como se indica en la leyenda, bajo la figura, mientras que la orientación indica en qué sentido va dicha información. A continuación se explican brevemente los distintos tipos de información que pueden intercambiar entre sí los diversos bloques:

- Información de Control y Gestión (flecha roja): es toda la información que se envía desde un bloque para controlar alguna opción o funcionalidad de otro.
- Flujo de Audio (flecha azul): es información de audio, capturada por el micrófono o dirigida a un dispositivo de reproducción (por ejemplo, unos altavoces).
- Información del Habla del Usuario (flecha amarilla): son los parámetros extraídos directamente de la voz del usuario, que definen su forma de hablar, y que definen el modelo acústico adaptado al usuario.

Los apartados siguientes explican los distintos bloques de la Figura 2.2

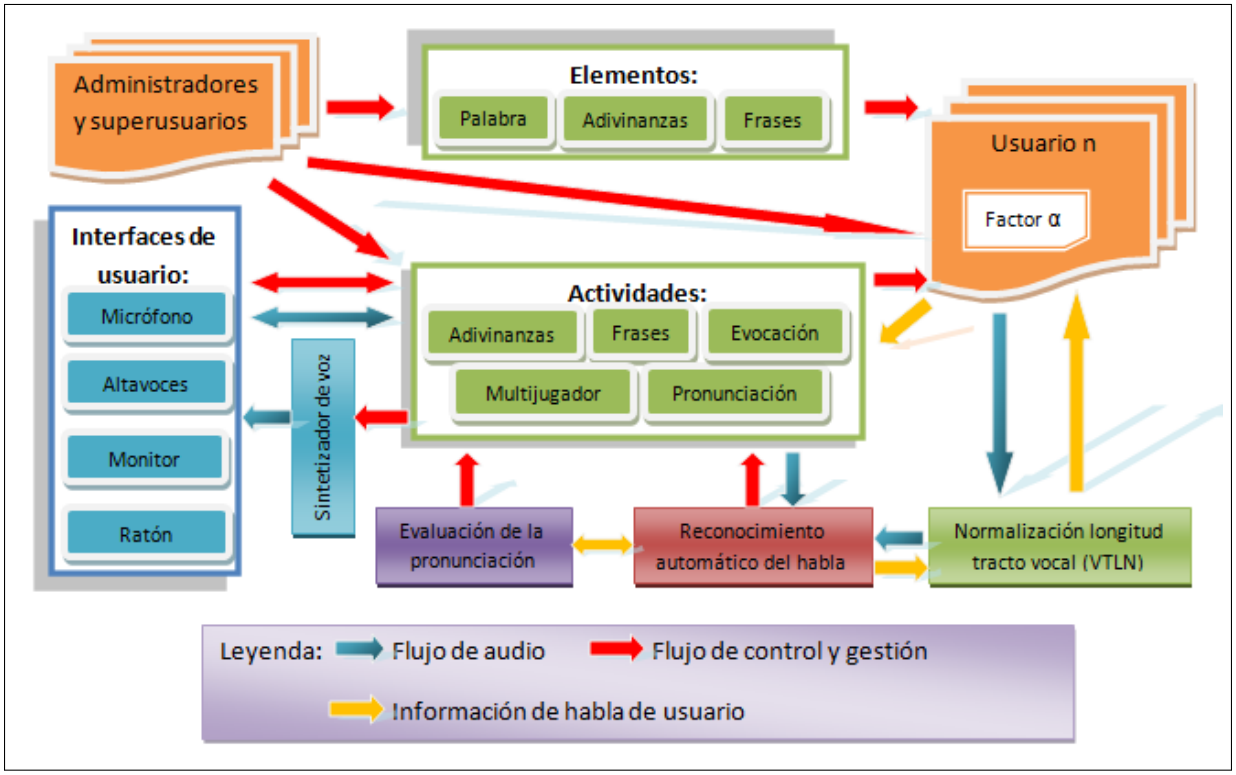

Figura 2.2: Diagrama de bloques

2.4.2 Interfaz de usuario

La aplicación "Vocaliza 2.0" trabaja fundamentalmente con la voz del usuario, debido a ello además de los periféricos habituales para el control de cualquier aplicación se hacen indispensables un micrófono para capturar la voz del usuario, y unos altavoces, ya que en algunas actividades la aplicación va a reproducir las palabras que el usuario debe repetir.

Desde este bloque el usuario puede acceder a las diversas opciones de configuración, tanto para introducir o modificar elementos de las distintas actividades, como para gestionar la información de los usuarios. Igualmente puede acceder a los distintas actividades, o iniciar la calibración, que permitirá adaptar el modelo acústico a la longitud del tracto vocal de un sujeto.

Como requisito de diseño, la aplicación "Vocaliza 2.0" debía ser muy sencilla de manejar para el usuario, y a su vez debía de poseer muchas opciones de configuración, lo que la convertiría en una potente y versátil herramienta de trabajo para los educadores.

Por ello dependiendo del rol que juega en la aplicación el usuario, al intentar acceder a la misma mediante su nombre de usuario y contraseña 2.3, se le mostrará una ventana principal u otra, por ejemplo los administradores o súper usuarios accederán directamente a la ventana principal de gestión de la aplicación como muestra la figura 2.4 , mientras que los usuarios normales accederán directamente a su ventana de actividades como se muestre en la figura 2.5

En el desarrollo de la interfaz gráfica se ha prestado especial atención en que el resultado sea una interfaz accesible, sencilla aunque completa a su vez, y muy intuitiva. De modo que no se requiera de instrucciones especiales para su uso.

Desde aquí, como vemos, el interesado tiene acceso directo a todo aquello que le con-

Figura 2.3: Ventana acceso usuario

Figura 2.4: Ventana principal administrador

Figura 2.5: Ventana principal usuario

cierne, por ejemplo el administrador puede acceder con un solo clic tanto a la información o configuración de usuario (figura 2.6), a la creación de uno nuevo como a la gestión de elementos necesarios para actividades como las frases (figura 2.7).

Cada súper usuario controla un aula, en ella podrá crear tantos usuarios como quiera, y podrá asignar diferentes actividades a cada usuario, así como configurar cada actividad para un usuario concreto a partir de los elementos para las actividades que posee el súper usuario. Por ejemplo el súper usuario crea palabras o adivinanzas (Figuras 2.8, 2.9) y luego las asigna libremente a cada usuario como en el caso de las palabras (Figura 2.10).

2.4.3 Actividades

"Vocaliza 2.0" pretende tratar las patologías del habla mediante diversos juegos, que trabajan distintos niveles del lenguaje. Todos los juegos funcionan de forma similar: la aplicación muestra una o varias imágenes que pueden llevar texto asociado y reproduce un sonido que el sujeto debe relacionar con una o varias palabras y pronunciarlas adecuadamente. En caso de hacerlo correctamente, el sujeto habrá superado el juego con éxito. Por tanto, los juegos necesitarán un reconocedor automático del habla, que interprete la respuesta dada por el sujeto.

Cada una de las actividades se comenta brevemente a continuación.

A- Pronunciación

Este juego trabaja el lenguaje en su nivel fonológico, es decir, pretende que el usuario practique la correcta pronunciación de los fonemas o sonidos que componen una palabra.

Figura 2.6: Ventana de información de usuario

Figura 2.7: Ventana de creación de frases

Figura 2.8: Ventana de creación de palabras

Figura 2.9: Ventana de creación de adivinanzas

Figura 2.10: Ventana asignación de palabras

La actividad de pronunciación es el más básico de los cinco juegos que se plantean. En él la aplicación muestra un pictograma por pantalla y reproduce un sonido asociado a la palabra que representa dicho pictograma, normalmente la misma palabra. El sujeto simplemente debe pronunciar dicha palabra con la mayor precisión posible. Si la palabra se ha pronunciado correctamente el sujeto superará el juego, y recibirá una calificación estimada por el bloque de Evaluación de Pronunciación, en función de lo bien que lo haya hecho.

Podemos ver la actividad de pronunciación en la figura 2.11.

B- Adivinanzas

Este juego trabaja el lenguaje en su nivel semántico, esto es, obliga al usuario a razonar y asociar imágenes y sonidos con ideas y significados concretos. El juego consiste en plantear una adivinanza al sujeto, que en términos de la aplicación es simplemente una pregunta con tres posibles respuestas (figura 2.12). La aplicación reproduce la pregunta que también se muestra por pantalla, al tiempo que muestra tres imágenes asociadas a las posibles respuestas. El usuario debe pronunciar adecuadamente la respuesta correcta para superar el juego con éxito.

C- Frases

El juego de las frases trabaja el lenguaje en su nivel sintáctico, ayudando al usuario a comprender como se forman oraciones y el orden que tienen las palabras en las mismas. El juego consiste en mostrar una sucesión de imágenes que forman una oración (figura 2.13). El usuario debe decir la oración correspondiente, en esta ocasión no es tan importante la pronunciación de cada palabra como el orden en las palabras y la continuidad a la hora de pronunciarlas. El juego se supera con éxito si todas las palabras que forman la oración son pronunciadas en el orden mostrado y con cierta fluidez.

D- Evocación

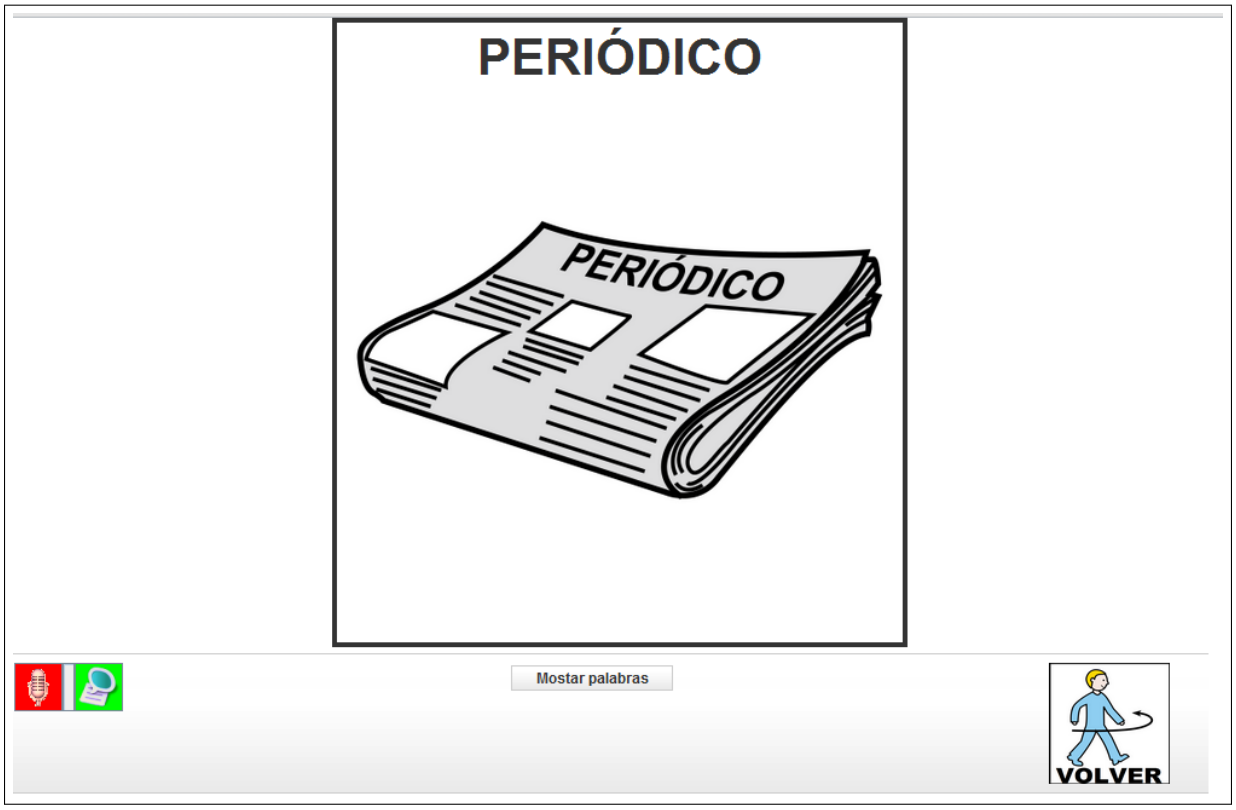

Figura 2.11: Actividad Pronunciación

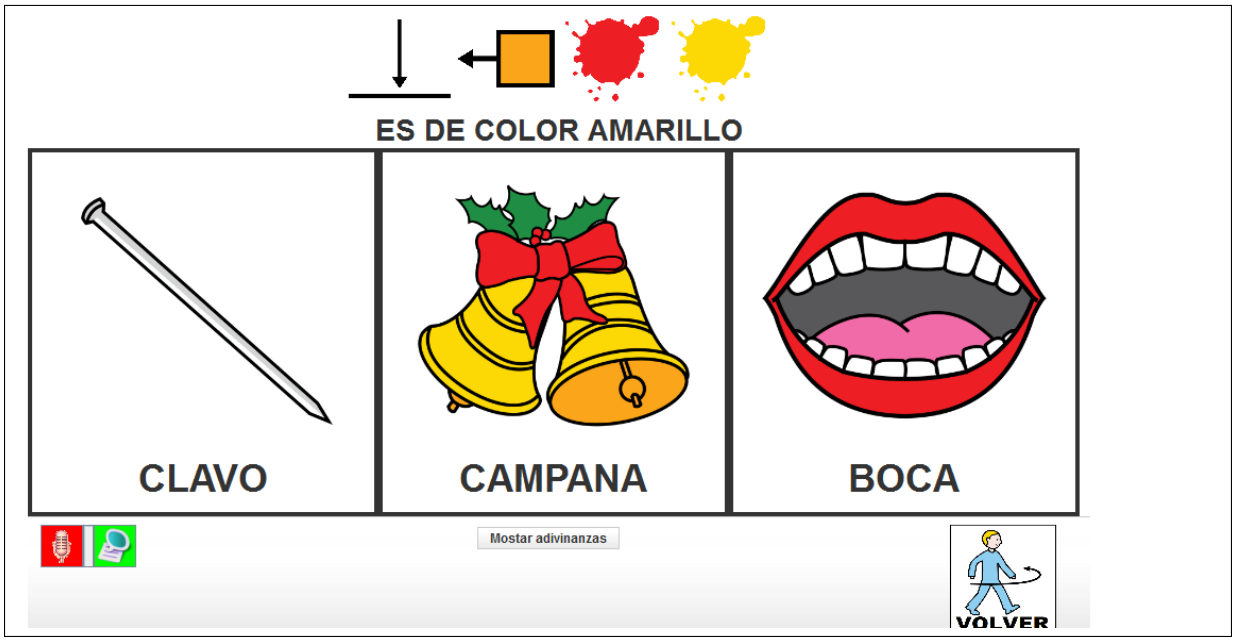

Figura 2.12: Actividad Adivinanzas

Figura 2.13: Actividad Frases

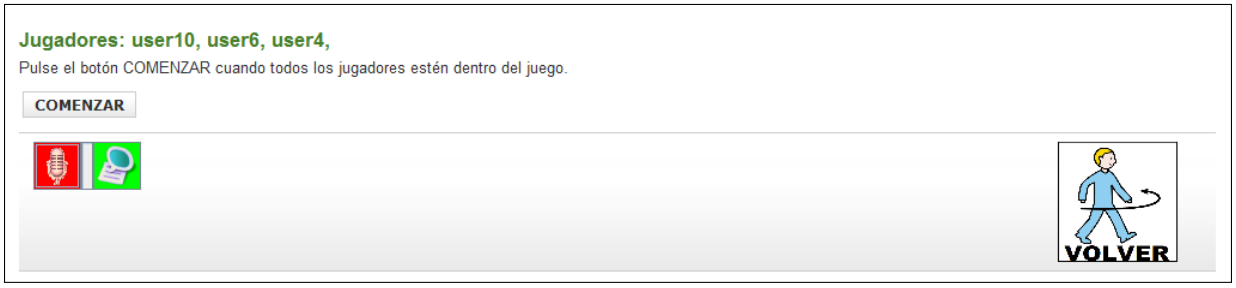

Figura 2.14: Sala de espera actividad multijugador

El juego de evocación no está diseñado para trabajar ningún nivel del lenguaje en concreto sino que permite al sujeto que lo utiliza practicar libremente las palabras que desee sin que sea la aplicación la que obligue a pronunciar esas palabras, como sucedía en el juego de pronunciación.

Este juego no muestra ninguna imagen ni reproduce ningún sonido a priori. Es el usuario el que debe pronunciar la palabra que el desee practicar, de forma que si lo hace adecuadamente, la aplicación mostrará la imagen asociada a dicha palabra, confirmando la buena pronunciación del sujeto.

En principio esta actividad era la menos valorada por los logopedas, pero se ha mantenido ya que se ha mejorado levemente su funcionamiento.

E- Tablero multijugador

Es la principal novedad en cuanto a actividades, es la primera multijugador. Está basado en peticiones ajax al servidor tanto síncronas como asíncronas. Pensada como actividad colaborativa, en esta actividad varios usuarios entran en una sala de espera, y cuando alguno lo cree conveniente, pulsa el botón de comenzar actividad (figura 2.14). Entonces se carga el tablero con pictogramas y los usuarios deben pronunciar las diferentes palabras que corresponden a los pictogramas que se muestran, conforme la aplicación reconoce las palabras que el usuario pronuncia, tacha la palabra del tablero(figura 2.15), cuando todas las palabras son tachadas, el juego concluye y muestra el tiempo total que se ha empleado en resolver el tablero.

Figura 2.15: Actividad Tablero Multijugador

El proceso que sigue la aplicación cuando un usuario decide utilizar uno de los juegos es el siguiente:

En primer lugar el usuario elige el juego deseado, y la aplicación carga las palabras, adivinanzas o frases del usuario desde la base de datos del propio usuario, confeccionada por el administrador.

A continuación el bloque de Actividades muestra las imágenes, el texto y reproduce los sonidos asociados para que el usuario pronuncie la palabra o palabras correspondientes. El bloque de Actividades recibe la información de audio capturada por el micrófono y la envía al bloque de Reconocimiento Automático del Habla, que devuelve al bloque de Actividades la palabra o palabras reconocidas. En su caso, el bloque de Actividades recibirá información sobre la calidad de la pronunciación desde el bloque de Evaluación de Pronunciación.

Si el usuario tiene éxito en la actividad, un dibujo animado aparecerá en la pantalla, que dependerá de la calidad de la pronunciación en el caso del juego de Pronunciación y del número de intentos. En caso de fallo, el juego continuará, sin ningún estímulo para el sujeto, para evitar que éste encuentre divertido el hecho de cometer errores.

Finalmente los resultados se almacenan en las estadísticas de usuario (figura 2.16) para que luego el súper usuario correspondiente pueda consultarlas

2.4.4 Normalizacion de la longitud del tracto vocal del locutor (VTLN)

En diversos estudios [7] se ha comprobado que los resultados en el reconocimiento automático del habla sufren un gran deterioro en el caso de que los locutores sean niños, entre otras cosas debido a sus características físicas.

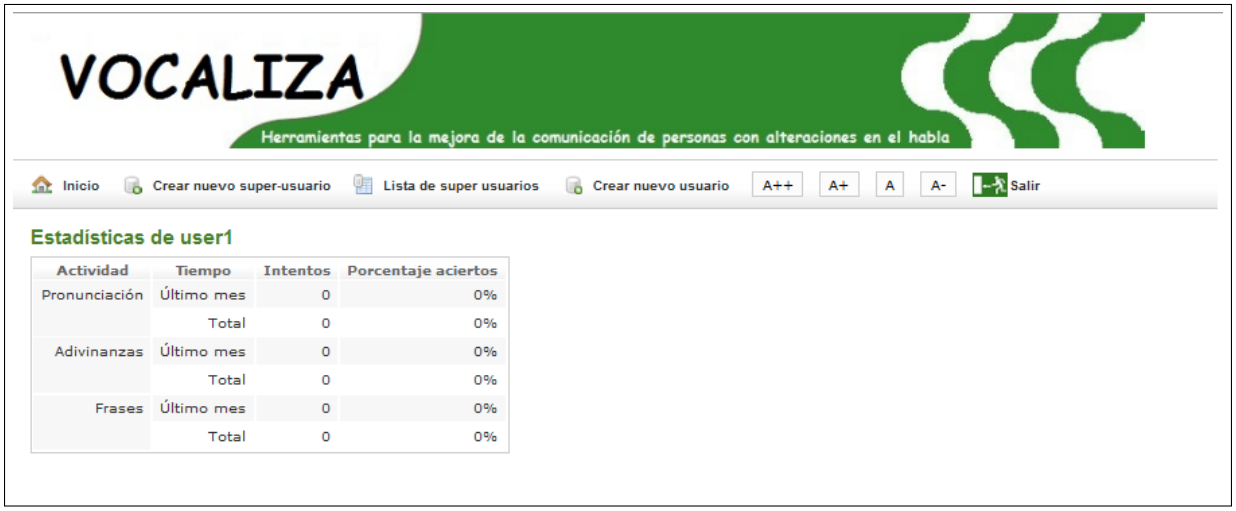

Figura 2.16: Estadísticas de usuario

En el caso de nuestra aplicación, el público principal al que está dirigido es fundamentalmente el de los niños. Por lo cual, para intentar mejorar estos resultados en parte, ya que no disponemos de datos suficientes para crear modelos adaptados a los niños, pretendemos adaptar el reconocimiento a las características fisiológicas del locutor.

Para ello, como novedad, utilizamos la técnica de la normalización de la longitud de tracto local del locutor, a partir de ahora VTLN el cual explicaremos en profundidad en el capitulo 3.

2.4.5 Gestión de usuarios

Este bloque, denominado Usuarios en la figura 2.2, es gestionado principalmente por el administrador, que tiene la capacidad de crear y eliminar otros súper usuarios además de la posibilidad de gestionar una clase de usuarios. Un nivel por debajo se encuentran los súper usuarios, los cuales pueden gestionar una clase de usuarios, y finalmente los usuarios. Ver figura 2.17

Aquí se almacena la información relativa a cada usuario de la aplicación, que el resto de los bloques consultará para el correcto funcionamiento de la aplicación. Dicha información incluye:

- Código del usuario.
- Contraseña del usuario, encriptada con un algoritmo de tipo Hash, que son los más utiles para almacenar este tipo de información sensible; ya que son algoritmos de solo ida, es decir, no es posible desencriptar lo que se ha encriptado. En este caso se utiliza el algoritmo SHA-256 que proporciona seguridad de 128 bits. Para hacer más complicados los ataques de fuerza bruta también se le añade un texto generado aleatoriamente a la contraseña antes de generar el Hash, más conocido como “Salt”. Esto ayuda a proteger las contraseñas más debiles. Para más información consultar Documentación de Spring Security Core [8].

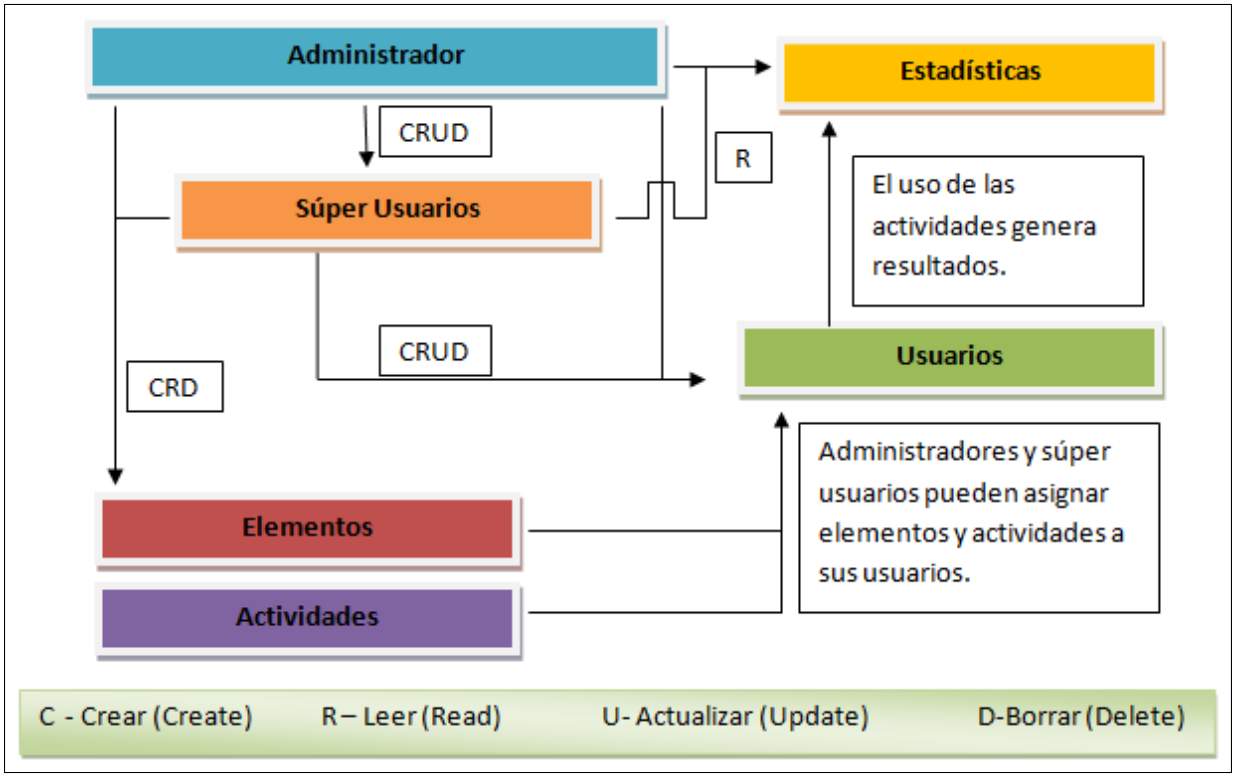

Figura 2.17: Gestión de Usuarios

- Elementos asociados al usuario: las palabras, adivinanzas y frases que tenga asociadas ese usuario, que serán aquellas que aparecerán en los diversos juegos. Controladas por el súper usuario correspondiente.
- Actividades asociadas al usuario: cada usuario puede tener acceso a diferentes actividades en función de lo que el súper usuario crea conveniente.
- Configuración de los diversos bloques de tecnologías del habla: incluyendo los parámetros de adaptación al locutor.
- Estadísticas del uso de las diferentes actividades, para que el súper usuario pueda realizar un seguimiento completo.

Finalmente, el bloque de Gestión de Usuarios permite añadir y eliminar usuarios a los súper usuarios, así como modificar la configuración de cualquiera de ellos. Las funcionalidades que aporta son:

- Añadir y eliminar palabras, adivinanzas y frases del usuario. El súper usuario selecciona los elementos deseados de las listas de palabras, adivinanzas y frases de su base de datos para asignarlas a miembros de su clase.
- Añadir y elminar actividades de usuario.
- Calibración, para la adaptación del modelo acústico a la longitud del tracto vocal del locutor.
- Cambio de contraseñas de usuario.
- Mostrar estadísticas de usuario.

2.4.6 Gestión de Elementos

Como se ha explicado en el bloque de actividades, la aplicación utiliza palabras, adivinanzas y frases como elementos imprescindibles en las distintas actividades que ofrece. El bloque de Gestión de Elementos permite gestionar estos componentes, es decir, permite añadir, eliminar y modificar palabras, adivinanzas y frases de la aplicación.

En esta ocasión se ha separado la gestión de elementos en dos partes. Por un lado el súper usuario crea su base de datos de palabras, frases y adivinanzas, por otro podrá asignar estos elementos a los usuarios de su clase desde otro menú según considere oportuno.

Los distintos componentes o elementos que se pueden gestionar se explican con mayor detalle a continuación:

- Palabras: las palabras constituyen el elemento fundamental de la aplicación "Vocaliza 2.0" Una palabra debe tener uno o más pictogramas asociados. Las palabras son utilizadas directamente por los juegos de pronunciación, de evocación y por el tablero multijugador e indirectamente por el resto de las actividades. Para añadir nuevas palabras a la base de datos, el súper usuario introduce una palabra en la aplicación y ésta busca la palabra indicada dentro de la base de datos de ARASAAC(Portal Aragonés de la Comunicación Aumentativa y Alternativa [9] que ofrece recursos gráficos y materiales para facilitar la comunicación de aquellas personas con algún tipo de dificultad en este área.) mediante un servicio web, permitiendo seleccionar entre los pictogramas que esta palabra tiene asociados que se adecúen mejor a nuestras necesidades. En el menú de asignación al usuario final existe un buscador para hacer más facil la labor de búsqueda y asignación de elementos al usuario final.

- Adivinanzas: las adivinanzas están formadas por una pregunta o pista y tres palabras que son las posibles respuestas. Sólo una es correcta, y todas ellas deben ser palabras contenidas en la aplicación. Las adivinanzas son utilizadas únicamente por el juego de adivinanzas. Al insertar el texto de la adivinanza se nos permite la selección del pictograma más adecuado para cada palabra insertada, así como la posibilidad de reordenarlos si es necesario.

- Frases: las frases son una sucesión de palabras. Al igual que las adivinanzas deben ser palabras contenidas en la aplicación. Al igual que en el caso anterior cuando insertamos una frase se nos permite la selección del pictograma más adecuado para cada palabra y su reordenación en caso de ser necesario.

Además de las palabras contenidas en la aplicación se pueden introducir ciertas palabras derivadas de estas, las conjugaciones de los verbos y los cambios de singular a plural, y el cambio de género son reconocidos por la aplicación.

También se incluye un buscador dentro del menú de asignación de elementos a los usuarios finales, lo que facilita la asignación de elementos por bloques de contenido o simplemente la facilita en grandes bases de datos de súper usuario.

Con este modelo de gestión se puede trabajar con cada usuario con los elementos más adecuados para su caso concreto.

2.4.7 Bloque de Reconocimiento Automático del Habla

El reconocedor integrado en la aplicación es el que el GTC está desarrollando y utilizando para la investigación en el ámbito de las tecnologías del habla. Dicho reconocedor está basado en la teoría de Modelos Ocultos de Markov, explicada en el Anexo A. Tal como está implementado, permite utilizar modelos acústicos muy variados, cuyas unidades acústicas pueden ser cualquiera de las descritas en el Anexo A. En concreto la aplicación "Vocaliza 2.0" utiliza modelos acústicos contextuales, ya que cada unidad acústica permite modelar los distintos contextos en los que se puede encontrar un fonema con gran precisión y versatilidad. Las unidades acústicas contextuales utilizadas se obtienen de dividir un fonema en 3 unidades acústicas tal y como se explica en el Anexo A. Manejando unidades acústicas tan pequeñas, bastará con que cada una tenga asociado un único estado en el modelo de Markov. La función de densidad de probabilidad asociada de cada estado de la cadena oculta de Markov se modela con una mezcla de 16 gaussianas.

Como parámetros de entrada, el reconocedor obtiene de la voz del usuario un total de 13 medidas acústicas en una ventana temporal de 25 ms, siendo el desplazamiento de dicha ventana de 10 ms, es decir, se extraen 13 medidas acústicas cada 10 ms. Dichas medidas acústicas son la energía localizada de la señal y 12 parámetros Mel-cepstrum localizados a los que se les ha sustraído el valor medio de cada uno de ellos a lo largo del tiempo, mediante el proceso conocido como CMS (sustracción de la media del cepstrum) que minimiza la influencia del canal (resonancia de la sala, micrófono, tarjeta de sonido) en la señal adquirida, cancelando dicha influencia por completo si el canal es lineal e invariable en el tiempo. Además de dichas medidas, el reconocedor utiliza la primera y segunda derivada de los valores del cepstrum de la señal, y la primera derivada del logaritmo de la energía. El concepto de cepstrum de la señal de voz se explica en el Anexo A.

Este bloque puede considerarse el núcleo de las tres actividades que componen la aplicación. Para cada uno de ellos, necesita cierta información para inicializarse, lo que incluye el modelo adaptado al usuario, el vocabulario permitido, con la correspondiente transcripción a subfonemas de cada palabra, y las gramáticas que reglan la construcción de respuestas válidas. Tanto el Bloque de Usuarios como el Bloque de Actividades son los encargados de proveer de esta información antes de iniciar cada uno de las actividades.

Para la aplicación "Vocaliza 2.0" ha sido necesaria una dedicación especial a la generación dinámica de las gramáticas que utiliza el reconocedor, ya que éstos dependen de la actividad y usuario activos.

Una vez iniciada la actividad, el bloque se coloca en estado de espera, hasta que reciba la señal de audio procedente del Bloque de Actividades, devolviendo la palabra o palabras reconocidas, en el caso de haberlas. Asimismo, realiza la extracción de parámetros y su envío al bloque siguiente.

2.4.8 Bloque de Verificación de la Pronunciación

El bloque de Evaluación de Pronunciación permite evaluar la calidad de la articulación del usuario. Para ello se utiliza un algoritmo basado en la teoría de Verificación de Pronunciación (Utterance Verification, [10]) que plantea la posibilidad de dotar de mayor robustez a los sistemas de reconocimiento comprobando el ratio de verosimilitud entre la probabilidad de que la articulación de entrada al reconocedor esté asociada a una determinada palabra o serie de palabras según el modelo utilizado por el reconocedor, conocido

como hipótesis nula, y la probabilidad de que la articulación esté asociada a la misma o mismas palabras, en base a otro modelo alternativo, conocido como hipótesis alternativa. Dicho ratio de verosimilitud ofrece una medida de confianza que indica si realmente se ha pronunciado lo que el reconocedor ha reconocido, por tanto puede establecerse un umbral de decisión a partir del cual puede considerarse que la salida del reconocedor es correcta.

El algoritmo utilizado para evaluar la pronunciación fue desarrollado en el proyecto "Vocaliza" [1].

2.4.9 Bloque de Síntesis de Voz

El bloque de Síntesis de Voz realiza las conversiones de texto a voz que demanda la aplicación durante el proceso de ejecución. Sólo el bloque de Actividades y VTLN hacen uso de esta funcionalidad, que no es imprescindible para la aplicación, pero sí facilita la labor de los usuarios del programa, tanto a la hora de configurar la aplicación como a la hora de utilizar las diversas actividades.

El sintetizador de voz entra en funcionamiento en las diversas actividades sintetizando una palabra o frase. Su función es simplemente la de ofrecer la opción de acompañar con audio las imágenes y textos que se muestran en las actividades.

Para nuestra plataforma podemos usar el sintetizador de voz utilizado que el GTC está desarrollando en sus investigaciones en el ámbito de síntesis de voz en el caso de la versión distribuida por los centros. En el caso de la plataforma en la nube podría estudiarse la implantación del sintetizador de "Loquendo"[11] mediante la aquisición de la correspondiente licencia.

En cualquier caso esto le es indicado al cliente del reconocedor mediante un parámetro.

2.5 Modelos del lenguaje: Generación Dinámica

En la aplicación "Vocaliza 2.0" se decidió que lo más provechoso era la generación de la gramática dinámicamente para cada intento de reconocimiento del habla dentro de cada actividad, en función de ésta y del usuario que la realiza se genera una gramática para las respuestas, de este modo obtenemos un mayor control sobre las respuestas del usuario, acotando en la medida más conveniente los posibles resultados.

Capítulo 3

VTLN: Normalización de la Longitud del Tracto Vocal del locutor

3.1 Descripción del trabajo

Actualmente no se dispone de suficientes datos (grabaciones) del habla de los niños como para generar modelos adaptados al habla de éstos y así poder usar los métodos de adaptación que se vienen utilizando para los adultos.

En este caso introducimos un nuevo método para intentar resolver en parte el problema de las diferencias fisiológicas de diferentes locutores sin necesitad de tener muchos datos sobre el habla de un locutor.

VTLN o Normalización de la Longitud del Tracto Vocal [12] es un método para reducir los efectos de la diferencia de longitud en el tracto vocal de diferentes hablantes.

3.2 El Tracto Vocal

El tracto vocal incluye todo lo que se encuentra entre las cuerdas vocales y los labios. Las partes principales son la faringe, la cavidad nasal y las diferentes partes de la boca (Figura 3.1). Cuando cambiamos la forma del tracto vocal se obtienen diferentes sonidos debido a que diferentes formas del tracto vocal se convierten en diferentes frecuencias de resonancia. En las vocales y en las consonantes sonoras, las frecuencias de resonancia son llamadas formantes.

Para calcular las frecuencias de resonancia necesitamos un modelo simplificado del tracto vocal.

Un modelo simple es el del tubo sin perdidas. Para alcanzarlo, enderezamos el curvado tracto vocal y hacemos de él un cilindro. Dado que un cilindro es un modelo demasiado simple para nuestro propósito, concatenamos varios cilindros de diferentes radios. Si asumimos que no hay perdidas por fricción entre el aire y el tubo y que el tubo no vibra, entonces este es el modelo del tubo sin perdidas o el modelo de los tubos concatenados sin perdidas [13].

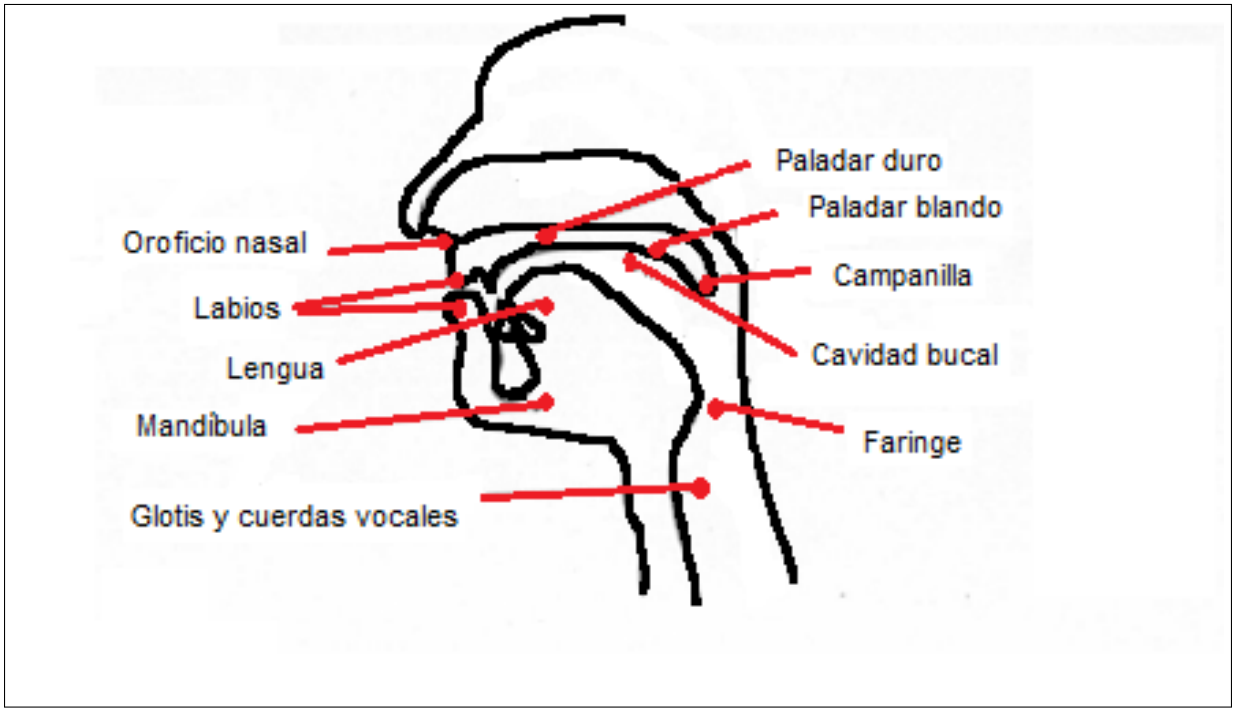

Figura 3.1: Órganos de producción del habla

De acuerdo con Holmes [12], las formantes, o frecuencias de resonancia, están equiespaciadas a:

$$f_{Hz} = \frac{(2n+1)V_s}{4L} \tag{3.1}$$

(con n=0, 1, 2, 3, ...). Donde V_s es la velocidad del sonido (en metros por segundo)y L es la longitud del tubo (en metros) que simula la longitud del tracto vocal. Esto explica porque las formantes en los niños están desplazadas a frecuencias más altas. Una mayor complejidad presenta el hecho de que las longitudes de las diferentes partes del tracto vocal no varían con la misma proporción con la edad de los niños. Por ejemplo, normalmente la garganta crece más que la boca. Para compensar completamente este efecto se requeriría una transformación acústica muy compleja pero solo el escalado lineal ha sido modelado en los sistemas de reconocimiento del habla.

3.3 Método VTLN

Actualmente las técnicas RAH funcionan bastante bien con voces de adultos para permitir aplicaciones prácticas. Pero cuando se reconoce a niños, los resultados decaen considerablemente. ¿Por qué ocurre esto?

La principal causa de este hecho es que no se disponen de suficientes datos para construir modelos adaptados al habla de los niños(grabaciones del habla) y poder usar los métodos de adaptación que se vienen utilizando para los adultos correctamente. Debido a este hecho debemos recurrir a otros métodos que ayuden a la adaptación sin necesidad de tener tantos datos sobre el habla para su correcto ajuste.

Los niños difieren tanto físicamente como en pronunciación con los adultos. Las diferencias en condiciones físicas vienen por el hecho de que los niños tienen un tracto vocal

más corto y cuerdas vocales más cortas. Esto hace que su frecuencia fundamental y su frecuencia formante estén desplazadas hacia frecuencias superiores. Para niños de cinco a siete años las primeras tres formantes están desplazadas alrededor de un 65 % comparadas con los adultos [12].

Cuando los niños pierden los dientes de leche, lo que ocurre cuando tienen alrededor de 6 años, su pronunciación pierde calidad por algún tiempo hasta que obtienen sus dientes permanentes. En los niños más pequeños también puede ocurrir que no hayan aprendido a pronunciar correctamente todavía.

Algunas de estas diferencias pueden ser compensadas, pero aun así los resultados empeoran. Esto es debido a que los niños tienen menos vocabulario y hablan más espontáneamente y menos gramaticalmente que los adultos.

También se encuentra que los resultados del reconocimiento del habla de niños varía de muy pobre hasta tan bueno como adultos [14] Esto hace a uno creer que las diferencias no pueden ser compensadas tan solo adaptando los modelos al habla de los niños o escalando el VTLN para hacer más parecido el habla al de los adultos. La diferencia entre ambos seguirá presente.

Las voces más brillantes de los niños son debidas en gran parte a la longitud más corta de su tracto vocal. Para compensar este hecho, el método VTLN puede ser de ayuda. La idea básica es estrechar el eje frecuencial para concentrar la potencia más o menos en el mismo lugar que se encuentra en el habla en adultos.

Como ya hemos comentado, si vocales idénticas producidas por dos tractos vocales de diferentes longitudes son comparadas, se podrá ver que las formantes del tracto vocal más corto están desplazadas a frecuencias más altas comparadas con las formantes del tracto vocal más largo. Esto puede ser compensado estrechando o comprimiendo el eje frecuencial del espectro, antes del reconocimiento automático del habla. Si se aplica el método VTLN, el estrechamiento/compresión se produce durante la parametrización (MFCC), mediante la reestimación de los bancos de filtros de Mel, después del efectuar el enventanado y la transformada de Fourier.

Para la implementación del metodo necesitaremos efectuar un barrido del factor de transformación α y recalcular los bancos de filtros de Mel para cada caso, el factor de transformación afecta en el cálculo de los centros del banco de filtros del siguiente modo:

$$f_{Hz}^{\alpha}(k\triangle f_{mel}) = 700(10^{k\triangle f_{mel}/2595} - 1)/\alpha \tag{3.2}$$

3.4 Implementación del método

En nuestro caso nos interesa reducir la diferencia entre adultos y niños el máximo posible debido a que vamos a trabajar principalmente con niños, y nos interesa que el reconocedor ofrezca resultados aceptables para ellos, por lo cual vamos a implementar el método VTLN dentro del reconocedor del grupo.

Para ello han sido necesarios dos tipos de cambios en el reconocedor

Por un lado los que tienen que ver con el cálculo del factor de transformación óptimo (Modo calibración), y los que permiten al reconocedor usar el factor de transformación

óptimo en el reconocedor durante el funcionamiento normal de las actividades (Modo normal), describiremos estos cambios a continuación.

3.4.1 El reconocedor

El reconocedor sigue una arquitectura cliente-servidor, en la que el cliente es un applet java (módulo o programa realizado en java, y que se puede incrustar en cualquier página web). Dicho cliente se ejecutará en cualquier navegador web, siempre que éste tenga habilitado java.

El cliente de reconocimiento de voz realiza las funciones de adquisición de audio, parametrización (MFCC), y compresión de los parámetros acústicos (DVQ) y envio en tiempo real al servidor, para reducir el ancho de banda de envío hasta 2.1 Kbps.

Por otra parte, el servidor de reconocimiento es un programa capaz de gestionar varias conexiones simultáneas de distintos applets , de forma que recibe los parámetros acústicos codificados, y detecta el fin del audio mediante un VAD(Voice Activity Detection) extrae los vectores de características MFCC a partir de éstos, y aplica el algoritmo de reconocimiento enviando la respuesta al applet. Además, en cada proceso de reconocimiento, cada conexión asociada a un applet debería haber recibido la gramática enviada por éste previamente.

Más información en ViVoLab [15]

3.4.2 Modo calibración

En adultos se parte del factor de transformación $\alpha = 1$, normalmente para niños este factor estará en torno al 0.8, en nuestro caso efectuaremos un barrido desde 0.7 a 1.1 para este factor.

En este caso fue necesario cambiar el modo de comunicación habitual entre el cliente y el servidor del reconocedor del grupo.

Normalmente el cliente enviaba trama a trama, casi en tiempo real, el audio hacia el servidor mientras se producía el habla, y éste era el encargado de la detección del habla para después calcular la respuesta y la enviarla al cliente, cortando la comunicación.

- Detección de principio y fin del habla en el cliente. Implementación de un detector de actividad del habla VAD.
- Parametrización con recalculo de los centros del banco de filtros de Mel para cada factor de transformación como se ha explicado anteriormente y envío al servidor.
- Captura de las confianzas devueltas por el servidor para la estimación del factor de transformación óptimo que se devolverá a la aplicación para ser guardado en la base de datos de usuario y su posterior uso, sin cerrar la comunicación hasta que el bucle se haya completado.

Una vez haya finalizado el barrido, nos quedaremos con el factor de transformación que produjo la máxima confianza en el reconocedor. Devolviéndolo a la aplicación y guardándolo en la base de datos.

Este método es más lento como es lógico debido a que efectuará el reconocimiento de múltiples secuencias (barrido del factor de transformación entre 0.7 y 1.1) pero solamente es necesario efectuarlo una vez por locutor.

Para que el reconocedor funcione en modo calibración deberemos pasarle el parámetro calibración(calibracion="true")

3.4.3 Modo normal

En este caso se sigue el método de comunicación habitual. La diferencia es que se incluye un nuevo parámetro externo, "alpha". Éste es el parámetro que se calculó en la calibración de usuario y viene marcado por el usuario de la aplicación. Este factor "alpha" provocará una reestimación de los bancos de filtros de Mel, para que estén adaptados al locutor que esté usando la aplicación.

Para el uso del reconocedor en este modo debemos pasarle el parámetro de transformación "alpha" (ej:alpha="0.9")

Si no se le envía ningún factor de transformación "alpha" al reconocedor, éste sigue funcionando como hasta ahora, lo mismo ocurre con el parámetro calibración, asegurando la compatibilidad del reconocedor actualizado con las aplicaciones que lo utilizan hasta ahora, con el factor de transformación "alpha=1" .

3.5 Resultados del método

Este método ha probado ser particularmente efectivo cuando existe poca información de adaptación del locutor (como es nuestro caso) incluso en modo no supervisado, así que lo utilizaremos reduciendo el tiempo de calibración de la herramienta respecto a métodos como la estimación de Máxima Verosimilitud (ML), o incluso a la estimación Máximo a Posteriori (MAP) pese a que no es tan efectivo como éstos. Recordar que para usar estos métodos es necesario disponer de más información sobre el habla del locutor para la elaboración de modelos del habla, información de la cual no disponemos en el caso de los niños.

Al ser necesaria la grabación de una única frase para la calibración de la herramienta no parece que vaya a ser un proceso incomodo para los usuarios como ha ocurrido en otras ocasiones.

De este modo conseguimos reducir el impacto de la diferencia de la longitud del tracto vocal en niños.

Con el método VTLN en solitario se pueden producir reducciones en la tasa de errores de hasta el 10 % [13]

Capítulo 4

Conclusiones

En el capítulo 1 se definía una serie de objetivos en las dos líneas diferenciadas de trabajo que comprende este proyecto. En este capítulo se recopilan las conclusiones alcanzadas en ambas líneas, así como la evaluación de lo conseguido y las posibilidades de mejora en un futuro.

4.1 Difusión del proyecto

El objetivo principal del proyecto era el desarrollo de "Vocaliza 2.0", una plataforma libre y en castellano de apoyo a la logopedia y comunicación en educación especial, que aprovecha las diversas tecnologías del habla estudiadas. La aplicación ha sido creada de un modo funcional, accesible e intuitiva para el usuario final, gracias en parte a la realmientación recibida por los usuarios colaboradores del Colegio Público de Educación Especial Alborada (CPEE Alborada), dentro del marco de colaboración con el Grupo de Tecnologías de las comunicaciones (GTC) de la Universidad de Zaragoza.

Por otro lado, aunque la aplicación en principio fue pensada para ser una aplicación en la nube, este punto tuvo que variarse debido a diversos inconvenientes como vimos en el capítulo 2(Confidencialidad de los datos, mantenimiento del servidor principal....). La solución adoptada finalmente nos llevó a la creación de una máquina virtual para generar una versión distribuida, instalable en las redes locales de cada centro que desee utilizarlo, que puede funcionar incluso sin conexión al exterior.

Esta aplicación será utilizada, tras la finalización del proyecto, en el CPEE Alborada, y en cualquier otro centro público que la considere necesaria, para que todas aquellas personas con patologías del lenguaje puedan practicar y mejorar su habla.

Más allá de las actividades que se han incluido en esta versión, se han sentado las bases para la inclusión dentro de la plataforma creada de otros proyectos como "Prelingua"[3] y "Cuéntame"[4]. Debido a la modularidad de la aplicación es fácilmente extensible a las necesidades que vayan surgiendo, igualmente serán aplicables todas aquellas mejoras que se practiquen en el reconocedor del grupo sin necesidad de reconfiguración de la aplicación.

Por supuesto una aplicación de este tipo siempre es mejorable, tanto desde el punto de vista gráfico como tecnológicamente, a continuación se exponen las principales mejoras que podrían aplicarse en la siguiente versión.

- Dotar de mayor configurabilidad a los administradores y súper usuarios sobre las aplicación. Permitir la asignación de deberes o permitir realimentaciones positivas y negativas configurables en función del usuario.
- Creación de nuevas actividades multiusuario útiles para los logopedas, aprovechando el modo de comunicación creado en la actividad Tablero Multijugador.
- Con la llegada en los próximos meses de Grails 2.0, mejorar la comunicación intercentros, con su mejorada gestión de múltiples bases de datos.
- Mejora en las estadísticas de usuario en función de los datos más útiles para los logopedas.
- Y lo más importante, seguir mejorando su funcionalidad logopédica y educativa.

En cierto momento del proyecto se meditó la posibilidad de crear un editor de actividades para que los logopedas pudieran crear sus propias actividades y pudieran integrarlas en la plataforma, pero la tecnología usada actualmente no permite este tipo de integración al menos de un modo asequible para el desarrollador desde el punto de vista temporal, así que fue descartado. Seguramente en futuras versiones de las tecnologías usadas se pueda desarrollar este tipo de editor, lo que haría la plataforma todavía más útil e interesante.

4.2 La adaptación del reconocedor automático del habla a la longitud del tracto vocal del hablante

Otro objetivo importante del proyecto era la mejora de resultados del reconocimiento automático del habla en niños, para ello se decidió implementar en el Reconocedor Automático del Habla del grupo, el método de Normalización de la Longitud del Tracto Vocal del locutor. Con este método se intenta mermar el efecto nocivo que las características fisiológicas del hablante causan en los resultados de reconocimiento en el caso de los niños, como ya hemos visto en el capítulo 3. Esta normalización mejora los resultados en niños en parte, ya que este método no puede corregir, por ejemplo, la falta de aprendizaje del idioma propia de los niños y la espontaneidad con la que los niños a veces se comunican.

Se buscaba que esta mejora se produjera sin la necesidad de mucho tiempo de grabación por parte del usuario, que en muchas ocasiones es tedioso, y según las realimentación recibida del uso de otras herramientas de este tipo, estos métodos de grabación acaban por no ser usados. En este caso solamente será necesaria una frase de grabación, lo que en principio no será muy incomodo. Con esta mejora aunque todavía se está algo lejos del resultado de reconocimiento por parte de una voz adulta, ha hecho que las distancias se recorten considerablemente como se ha visto en otros estudios [7].

Sería conveniente crear un estudio sobre los resultados reales que la implantación del método VTLN ofrece en nuestro caso, en esta ocasión no ha sido posible debido a la limitación en tiempo que tiene el proyecto así como por la dificultad para realizar el estudio con niños de las edades necesarias.

Por supuesto lo más interesante es poder crear modelos adaptados para los niños, pese a que todavía no se disponen de datos suficientes para crearlos, éste es un requisito

fundamental para ofrecer unos resultados excelentes en el reconocimiento automático del habla en niños

En esta línea de trabajo se puede seguir trabajando, intentando mejorar los resultados obenidos. Combinando otros técnicas con el VTLN podemos optimizar los resultados, por ejemplo, combinando este método con la técnica de Maximum Likelihood Linear Regresion (MLLR) [13].

También sería conveniente la integración dentro de la plataforma de otros servicios web creados por el GTC. Por ejemplo el applet para cliente que permite hacer grabaciones y adaptación, usando un servidor distinto (que también valdría para la verificación).

4.3 Líneas futuras: La aplicación Vocaliza 2.0 en la nube

El objetivo futuro de este proyecto es poder presentar una plataforma libre, accesible y en castellano de apoyo a la logopedia en la nube. Este hecho facilitaría mucho la comunicación intercentros, la realimentación para la mejora de la aplicación, etc. En el momento que haya recursos suficientes y evolucione el tratamiento de los datos sensibles, este objetivo podría ser cumplido.

Mientras tanto debería aprovecharse para integrar todas las tecnologías del habla disponibles y unificar los diferentes proyectos y servicios realizados por el GTC en la plataforma, y así poder tratar todos los aspectos del lenguaje desde la misma herramienta, para que llegado el momento, esta plataforma evolucionada pueda ser la referencia en apoyo a la logopedia.

Apéndice A

Introducción a las tecnologías del habla

En este anexo se pretende explicar brevemente las bases teóricas de las diversas tecnologías del habla que se han utilizado en la aplicación Vocaliza 2.0. Existe literatura [13] que trata los distintos temas aquí explicados con mayor detalle.

A.1 Reconocimiento automático del habla

El reconocimiento automático del habla (RAH) es el proceso por el cual una máquina transcribe en texto una señal acústica de voz. Dicho proceso puede permitir una posterior interpretación del texto de forma que la máquina "comprende" el mensaje que transporta la señal de voz, actuando en consecuencia.

A.1.1 Clasificación

Los sistemas de RAH admiten diversas clasificaciones, atendiendo a distintos criterios: En función del locutor, un sistema de RAH puede ser:

- Independiente del locutor: el sistema de RAH está preparado para trabajar con cualquier persona, no siendo tan fiable como un sistema que estuviera adaptado a un único locutor, pero siendo mucho más flexible.
- Adaptado al locutor: el sistema de RAH extrae automáticamente ciertos parámetros del habla del locutor adaptándose al mismo, aumentando así la fiabilidad del sistema.

En función del tamaño del vocabulario que el sistema de RAH puede reconocer, se puede hacer la siguiente clasificación:

- Sistemas de pequeño vocabulario: pueden reconocer entre 1 y 99 palabras.
- Sistemas de vocabulario medio: pueden reconocer entre 100 y 999 palabras.
- Sistemas con gran vocabulario: pueden reconocer más de 1000 palabras.

En función de lo que se pretende reconocer, un sistema de RAH puede ser:

- Reconocedor de palabras aisladas: el sistema reconoce cadenas de palabras pronunciadas con suficiente pausa entre ellas. Es muy fiable, pero requiere que el locutor utilice un modo de comunicación no natural.
- Reconocedor de habla continua: el sistema reconoce palabras contenidas en frases pronunciadas sin ningún tipo de restricción. De esta forma, el locutor se comunica de forma natural, pero el sistema es menos fiable.

A.2 Fundamento teórico

Un sistema de RAH pretende, dada una secuencia de medidas acústicas de la voz de un locutor, decidir que secuencia de palabras pertenecientes a un vocabulario finito fueron pronunciadas por dicho locutor. Suponiendo que O es la secuencia de medidas acústicas de la voz del locutor, que sirven de entrada al sistema de RAH, la probabilidad de que el locutor haya pronunciado una secuencia de palabras W pertenecientes al vocabulario del que dispone el reconocedor vendrá dada por $P(W|O)$. De esta manera, el sistema decidirá que la secuencia de palabras pronunciada $\underline{W}$ sea aquella que satisfaga la expresión del Máximo a Posteriori(MAP):

$$\underline{W} = argmax_w[P(W|O)] \tag{A.1}$$

Aplicando el teorema de Bayes, $P(W|O)$ puede obtenerse de la siguiente forma:

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \tag{A.2}$$

Así, la expresión final que utilizará el reconocedor para obtener la secuencia de palabras W , teniendo en cuenta que $P(O)$ es independiente de W será:

$$\underline{W} = argmax_w[P(O|W)P(W)] \tag{A.3}$$

Donde:

- $P(O|W)$ es la probabilidad de que al pronunciar la secuencia de palabras W se obtenga la secuencia de medidas acústicas O. Dicho término vendrá dado por un modelo acústico, que contiene información sobre las unidades acústicas (fonemas u otras unidades) que componen cada palabra y la información que caracteriza estadísticamente la secuencia de medidas acústicas asociada a cada unidad.
- $P(W)$ es la probabilidad de que se pronuncie la palabra W. Dicho término vendrá dado por un modelo del lenguaje, que contiene información sobre las posibles transiciones de una unidad a otra según las reglas gramaticales y léxicas del idioma a reconocer.

Para obtener ambos modelos se parte de un modelo de producción del habla. El modelo de producción del habla más utilizado por los sistemas de RAH consiste en considerar el proceso de producción del habla como un proceso estocástico regido por la teoría de los Modelos Ocultos de Markov, como se explica en el apartado I.1.4.

A.2.1 Parámetros acústicos: MEL-cepstrum

Existen numerosos parámetros o medidas acústicas de la voz del locutor que el sistema de RAH puede utilizar para determinar que palabras han sido pronunciadas, pero los más extendidos, con diferencia, son los coeficientes Mel-cepstrum (Mel-Frequency Cepstrum Coefficients, MFCC), de la señal de voz.

Esto se debe a que dicha representación Mel-cepstrum ofrece varias ventajas. En primer lugar, el cepstrum de la señal de voz permite separar de forma sencilla la influencia del canal de transmisión o el sistema de captación electrónico de la señal de voz, mientras que la escala de frecuencias Mel se adecua mucho mejor al comportamiento perceptual del ser humano.

Dada una señal de voz $x[n]$ su cepstrum complejo se calcula así:

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} ln[X(e^{jw})] e^{jwn}\, dw \tag{A.4}$$

Donde $X(e^{jw})$ es la transformada de Fourier de $x[n]$.

Es decir, el cepstrum complejo se define como la transformada de Fourier inversa del logaritmo neperiano de la transformada de Fourier de la señal de voz. Dicha transformación es homomórfica y se conoce como el análisis homomórfico de la señal de voz. Así, las distintas operaciones de convolución que sufra la señal $x[n]$ se convertirán en sumas en el dominio cepstral.

Puesto que la fase de la señal de voz no contiene información necesaria para comprender las palabras que han sido pronunciadas, los sistemas de RAH suelen utilizar el cepstrum real, que se define como la parte real del cepstrum complejo y se puede hallar directamente de la siguiente forma:

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} ln \left| X(e^{jw}) \right| e^{jwn}\, dw \tag{A.5}$$

Es decir, el cepstrum real se define como la transformada de Fourier inversa del logaritmo neperiano del módulo de la transformada de Fourier de la señal de voz. La representación Mel-cepstrum consiste en obtener el cepstrum de la señal de voz utilizando una escala logarítmica de frecuencias para realizar la transformación al dominio frecuencial, concretamente la escala de frecuencias Mel.

Los sistemas de RAH suelen utilizar como medidas acústicas los MFCC localizados de la señal de de voz. Puesto que la representación Mel-cepstrum comprime notablemente la información de la señal de voz, basta con tomar únicamente los doce primeros parámetros cepstrales. Dicho parámetros acústicos comprenden una representación estática de la señal de voz. La representación dinámica de la señal se suele realizar con la primera y segunda derivada de los parámetros cepstrales.

Una práctica muy habitual en los sistemas de RAH es sustraer el promedio temporal de cada uno de los coeficientes Mel-cepstrum. Dicha operación elimina la parte estacionaria de la señal, que es justamente la influencia del canal de transmisión y los dispositivos de adquisición.

A.2.2 Modelos Ocultos de Markov (HMM)

Como ya se ha indicado, la forma más habitual de modelar la producción del habla es mediante Modelos Ocultos de Markov (Hidden Markov Models, HMM). Un Modelo Oculto de Markov es básicamente una cadena de Markov en la que la salida observada es una variable aleatoria X generada de acuerdo a una función de densidad de probabilidad de salida asociada a cada estado. Esto implica que no hay una correspondencia clara entre la secuencia de salida observada y la secuencia de estados, por lo que no es posible determinar unívocamente la secuencia de estados que se ha producido dada una secuencia de observaciones. La cadena de Markov permanece, por tanto, "oculta". Es por ello por lo que se dice que son modelos ocultos.

De esta forma, cada secuencia de parámetros acústicos observada lleva asociada una cadena de Markov oculta, cuyos estados serán segmentos de la cadena de palabras que produjo la mencionada secuencia de parámetros. Así, el sistema de RAH debe estimar que serie de estados produjeron la secuencia de medidas acústicas observada, descubriendo así la cadena de Markov oculta y, por tanto, las palabras que produjeron la secuencia observada.

Los parámetros que definen un modelo oculto de Markov son:

- El conjunto de posibles estados existente en la cadena oculta de Markov.
- La matriz de probabilidades de transición entre los posibles estados.
- El conjunto de funciones de densidad de probabilidad asociadas a los distintos estados.
- La probabilidad que cada estado tiene de ser el estado inicial.

Los HMM utilizados en los sistemas de RAH presentan algunas restricciones, debidas justamente a que pretenden modelar la producción del habla. En primer lugar, la cadena de Markov que permanece oculta es tal que desde un estado sólo existen dos posibles transiciones, que son la permanencia en ese mismo estado y pasar al estado siguiente, lo que condiciona notablemente la matriz de transiciones. Igualmente, dado un estado, se asume que la probabilidad de que se produzca transición al estado siguiente o permanencia en el mismo estado depende únicamente del estado actual.

Además, se cumple también que los parámetros acústicos emitidos desde un determinado estado sólo dependen de ese estado y no de la secuencia de estados anteriores.

Todo ello permite utilizar el algoritmo de Viterbi como método de decodificación, para obtener la secuencia de estados más probable a partir de una secuencia de medidas acústicas, una vez se conocen los parámetros que rigen el Modelo Oculto de Markov.

Como se ha explicado anteriormente, los sistemas de RAH utilizan dos modelos para calcular la secuencia de palabras asociada a una secuencia de medidas acústicas: un modelo acústico y un modelo del lenguaje. El modelo acústico es justamente el Modelo Oculto de Markov.

A.2.3 Modelado acústico

El modelado acústico pretende estimar los parámetros que rigen el Modelo Oculto de Markov asociado al proceso de producción del habla.

En primer lugar, se deben definir los posibles estados del modelo, de forma que para cada palabra del vocabulario de un sistema de RAH exista una única secuencia de estados que la defina. Para ello hay que determinar, previamente, la unidad o segmento acústico que sea apropiada para llevar a cabo el mejor reconocimiento posible. Dicha unidad debería cumplir las siguientes condiciones:

- Exactitud: la unidad escogida debería tener una representación de forma exacta como secuencia acústica aunque aparezca en diversos contextos.
- Entrenabilidad: se deberán tener suficientes realizaciones de cada unidad para poder estimar de forma correcta sus parámetros asociados.
- Generabilidad: cualquier palabra que se quiera introducir al vocabulario de palabras reconocibles se deberá poder obtener a partir del conjunto de unidades ya determinado.

En base a dichas condiciones, las posibles unidades que se pueden elegir son:

- Palabras: las palabras son la unidad más grande que puede tomarse en los sistemas de RAH. Aunque pueden servir en determinados sistemas, no es habitual tomarlas como unidad acústica mínima de reconocimiento, ya que la entrenabilidad y generabilidad propia de esta unidad son pésimas, requiriéndose muchas realizaciones de cada palabra para extraer los parámetros adecuadamente, y no permitiendo añadir nuevas palabras al vocabulario, sin una estimación previa de los parámetros asociados a cada una de ellas.
- Fonemas: el fonema es la unidad básica del lenguaje, y su caracterización está íntimamente ligada al proceso de generación de voz. Los fonemas son fácilmente entrenables, ya que el número de fonemas es muy reducido, sólo 24 en castellano, por lo que es fácil obtener muchas realizaciones de cada fonema. También el uso de fonemas ofrece gran generabilidad, ya que todas las palabras están formadas por un número finito de fonemas. Sin embargo, los fonemas son muy difíciles de entrenar con exactitud para todos los contextos en los que pueden aparecer, ya que la realización acústica de un fonema se ve altamente alterada por los fonemas que tiene como vecinos.
- Subfonemas: los subfonemas son las unidades más pequeñas que se emplea en sistemas de RAH. Se obtienen al dividir cada fonema en tres segmentos. El segmento central es único para cada fonema, pero los segmentos inicial y final tienen en cuenta el contexto en el que se encuentra el fonema. De esta forma se consigue gran exactitud, corrigiendo el problema de contexto en los fonemas, pero se pierde entrenabilidad, ya que el número de subfonemas existentes es mucho mayor que el de fonemas. Los subfonemas están siendo muy utilizados últimamente en la mayoría de los sistemas de RAH.

Una vez decidida la unidad acústica que se utilizará para reconocer, deben asignarse a la misma una secuencia de estados. La longitud de esa secuencia dependerá de la unidad elegida: por ejemplo, los fonemas suelen modelarse con tres estados, que indican el comienzo, el centro y el fin del fonema, mientras que los subfonemas, como ya realizan dicha división a nivel de unidad, se suelen modelar con un único estado.

Conocidos los estados del HMM, el algoritmo de Baum-Welch [13] permite calcular el resto de parámetros que definen el Modelo Oculto de Markov, si se dispone de un número de medidas acústicas observadas suficientemente grande.

A.2.4 Modelado del lenguaje

El conocimiento lingüístico es una necesidad importante en los sistemas de reconocimiento del habla continua. La obtención de un modelo del lenguaje es necesaria para poder determinar la probabilidad de que se de una agrupación de palabras dada según las reglas sintácticas y gramaticales. De esta forma, se podrán descartar en la fase de reconocimiento frases que el modelo acústico da como muy probables, pero que no tengan sentido lingüístico.

Supongamos una secuencia específica de palabras $W = w_1 w_2..w_q$, la probabilidad de que se de esa secuencia de palabras se podrá calcular teniendo en cuenta la dependencia de cada palabra con las anteriores:

$$P(W) = P(w_1 w_2..w_q) = P(w_1)\Delta P(w_2|w_1)\Delta P(w_3|w_1 w_2)\Delta..\Delta P(w_q|w_1..w_{q-1}) \quad \text{(A.6)}$$

Este modelo es muy complejo, especialmente cuando se considera la probabilidad de aparición de una palabra condicionada a las $Q-1$ precedentes. Para ello, se reduce este cálculo a considerar la probabilidad condicionada a las N palabras más próximas. El caso más sencillo sería considerar sólo pares de palabras válidas o no, definiendo:

$$P(w_j|w_k) = \begin{cases} 1\, w_k w_j\, vlido \\ 0\, resto \end{cases} \quad \text{(A.7)}$$

En el caso general de usar N palabras para el cómputo de la probabilidad, ésta queda:

$$P_N(W) = \prod_{i=1}^{Q} P(w_i|w_{i-1}, w_{i-2}, .., w_{i-N+1}) \quad \text{(A.8)}$$

El cálculo de estas probabilidades se realiza simplemente con las frecuencias de aparición de conjuntos de palabras en un texto suficientemente largo y significativo. Así:

$$\hat{P}(w_i|w_{i-1}, w_{i-2}, .., w_{i-N+1}) = \frac{P(w_i, w_{i-1}, w_{i-2}, .., w_{i-N+1})}{P(w_{i-1}, w_{i-2}, .., w_{i-N+1})} \quad \text{(A.9)}$$

Esta forma de cálculo acarrea problemas debido a las limitaciones impuestas por el tamaño del texto, ya que puede haber secuencias de palabras válidas que, sin embargo, no aparezcan en el texto y asignemos $P(w_i, w_{i-1}, w_{i-2}, .., w_{i-N+1}) = 0$ sin serlo. Para evitar esto, se hacen algoritmos de alisado que no sólo tienen en cuenta la probabilidad de todo

el conjunto de palabras, sino de subconjuntos de palabras dentro del propio conjunto. Para el caso $N = 3$, se define:

$$\hat{P}(w_3|w_1, w_2) = p_1\Delta\frac{F(w_1, w_2, w_3}{w_1, w_2} + p_2\Delta\frac{F(w_1, w_2}{w_1} + p_3\Delta\frac{F(w_1)}{\sum F(w_i)} \tag{A.10}$$

Donde tomamos en cuenta la probabilidad de que aparezcan las tres palabras juntas, y la probabilidad de que se den dos o una de ellas por separado. Los pesos incluidos en la fórmula de alisado deben cumplir la condición de ser no negativos y sumar uno.

A.2.5 Evaluación

Para evaluar las prestaciones de un sistema de RAH se suele utilizar el número de errores que se cometen en las palabras reconocidas, medidos como WER (Word Error Rate). En esta medida influyen los tres tipos de errores que se pueden tener en reconocimiento:

- Sustituciones: Se produce una sustitución cuando una palabra de la frase correcta se reconoce como una palabra diferente. Por ejemplo, reconocer: "La noche del ganador" en vez de "La noche del cazador".
- Borrados: Un borrado se da cuando una palabra que aparece en la frase correcta es omitida en la frase reconocida. Por ejemplo: "Noche del cazador" en vez de "La noche del cazador".
- Inserciones: Tendremos una inserción cuando en la frase reconocida se introduce una palabra que no existía en la frase correcta. Por ejemplo: "La noche del la cazador" en vez de "La noche del cazador"

Contando estos tres tipos de errores se obtiene la WER:

$$WER(\,\%) = \frac{Sustituciones + Borrados + Inserciones}{Palabras}\Delta 100\,\% \tag{A.11}$$

Donde en el denominador se cuentan las palabras de la frase correcta, y no las de la frase reconocida, que pueden ser diferentes según las inserciones y borrados producidos. De esta forma, si por ejemplo la frase a reconocer es "La noche del cazador", y el reconocedor devuelve la frase "Noche del no cazador", la WER será del 50 %, ya que el número de palabras es cuatro, y se han producido dos errores, un borrado ("la"), y una inserción ("no").

La obtención de la tasa de error no es tan directa y sencilla como pueda parecer, ya que si hacemos una correspondencia de palabras directa entre las frases del ejemplo anterior, obtendríamos:

Frase Correcta	Frase Reconocida
La noche del cazador	La noche del no cazador

Donde se podría pensar que la tasa de error es del 75 %, con tres sustituciones ("Noche" en lugar de "La"; "del" en lugar de "noche", y "no" en lugar de "del") sobre cuatro palabras totales; cuando ya hemos visto que la WER real es el 50 %.

A.3 Adaptación al locutor

La adaptación de un sistema RAH al locutor es el proceso por el cual se estiman los parámetros del Modelo Oculto de Markov que rige la producción del habla de ese locutor determinado, en base a un conjunto de observaciones o corpus del habla del locutor.

Existen varios métodos para adaptar un sistema de RAH al locutor. A continuación se describen los dos métodos más utilizados.

A.3.1 Estimación de Máxima Verosimilitud (ML)

Dado un conjunto de secuencias de medidas acústicas del habla de un locutor $X_1, X_2, .., X_N$, la Estimación de Máxima Verosimilitud [16] trata de estimar el modelo que maximiza la verosimilitud conjunta vista como una probabilidad:

$$\phi = argmax_{\phi}[\prod_{i=1}^{N} P(X_i|\phi)] \tag{A.12}$$

Esto es justamente lo que hace el algoritmo de Baum-Welch [13] para estimar los parámetros de un Modelo Oculto de Markov a partir de un conjunto de secuencias de medidas acústicas.

El principal inconveniente de este método es que requiere una gran cantidad de observaciones para conseguir una buena estimación. Este método se utiliza normalmente partiendo de medidas acústicas pertenecientes a diversos locutores, con la intención de estimar un modelo para un sistema RAH independiente del locutor, pero no es resulta muy interesante para crear modelos adaptados al locutor, pues requeriría adquirir gran cantidad de observaciones del habla del usuario.

A.3.2 Estimación Máximo a Posteriori (MAP)

Ciertamente, aunque distintas observaciones de medidas acústicas asociadas a un mismo estado pueden presentar diferencias considerables, que se pueden acentuar cuando las observaciones provienen de locutores distintos, todos los locutores pronuncian los mismos fonemas de forma similar, esto es, las medidas acústicas asociadas a un mismo estado tendrán ciertas semejanzas que las harán muy distintas de otras medidas asociadas a otros estados.

Por tanto, se puede aprovechar la información contenida en un modelo acústico estimado con múltiples locutores, independiente del locutor, para ir adaptándola al habla de un locutor concreto, a partir de unas pocas observaciones de su habla. Esto resolvería el problema de requerir gran cantidad de observaciones del habla del locutor, pues bastaría con un número mucho menor que simplemente matizara el modelo independiente del locutor para convertirlo en un modelo adaptado al mismo.

Para ello se utiliza el método de estimación MAP [17], o estimación bayesiana.

Apéndice B

Instalación del servidor

En este anexo se pretende mostrar el proceso de instalación de la parte servidora en la red local del centro donde se quiere instalar la aplicación. Se incluyen varias imagenes para explicar los pasos de manera más visual.

B.1 VM VirtualBox

B.1.1 Instalación

- Abrir un explorador de red y visitar la siguiente URL : http://www.oracle.com/technetwork/server-storage/virtualbox/downloads/index.html
- Una vez ahí, descargar la última versión para el sistema operativo que tengamos instalado en el equipo servidor.
- Ejecutamos el instalador una vez descargado y nos encontramos con la figura B.1 , en el caso de Windows:
- Hacemos click en Next, opciones por defecto y continuamos. La instalación se completará automáticamente.

B.1.2 Configuración de la máquina virtual

- Abrimos VM VirtualBox y nos encontramos con la figura B.2
- Hacemos click sobre nueva, para crear una nueva máquina virtual.
- En el siguiente paso le damos un nombre a la nueva máquina virtual, y elegimos sistema operativo (Ubuntu,Linux). Como se indica en la figura B.3
- El siguiente paso es asignar memoria RAM a nuestra máquina virtual como se ve en la figura B.4
- Después debemos seleccionar el disco duro de la máquina, usaremos un disco duro existente:“Vocaliza.vdi”, que estará incluido en el paquete de instalación. Figura B.5

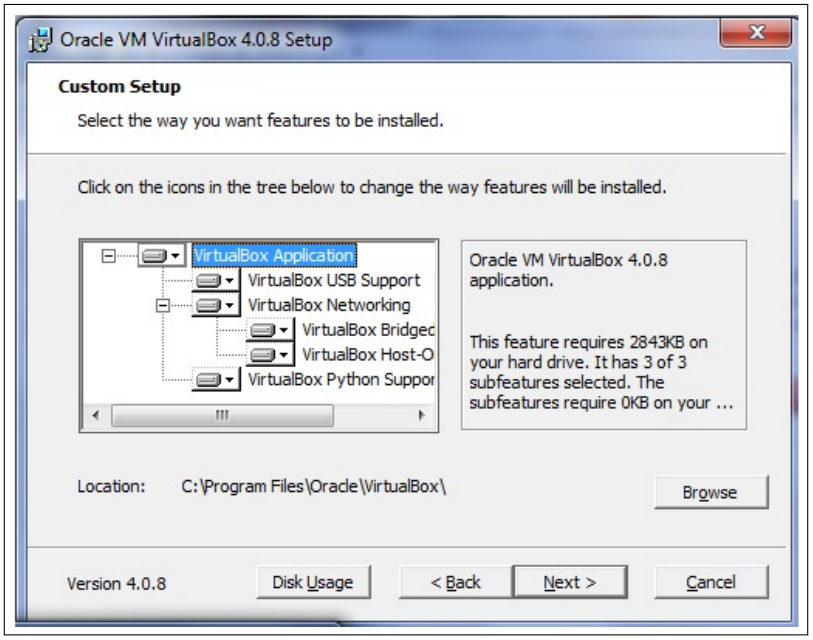

Figura B.1: Instalador Oracle VM

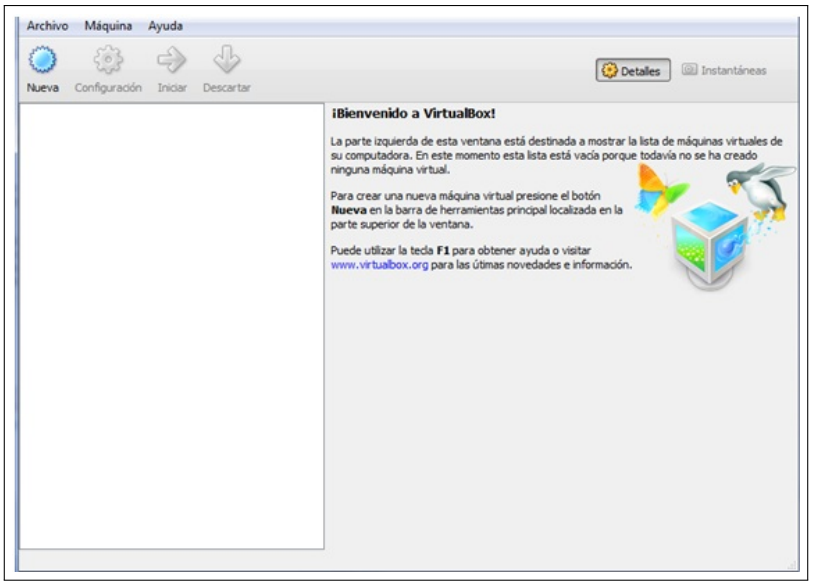

Figura B.2: Ventana de inicio Oracle VM

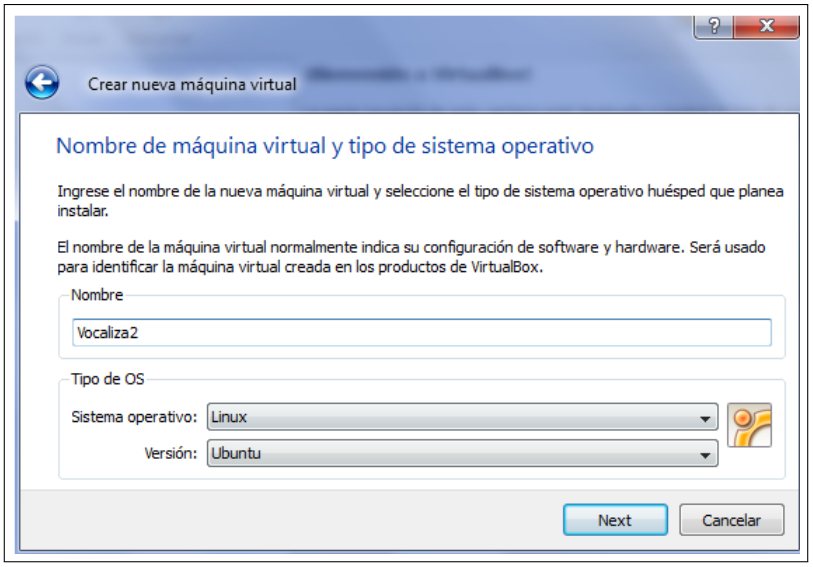

Figura B.3: Nombre nueva máquina virtual y sistema operativo

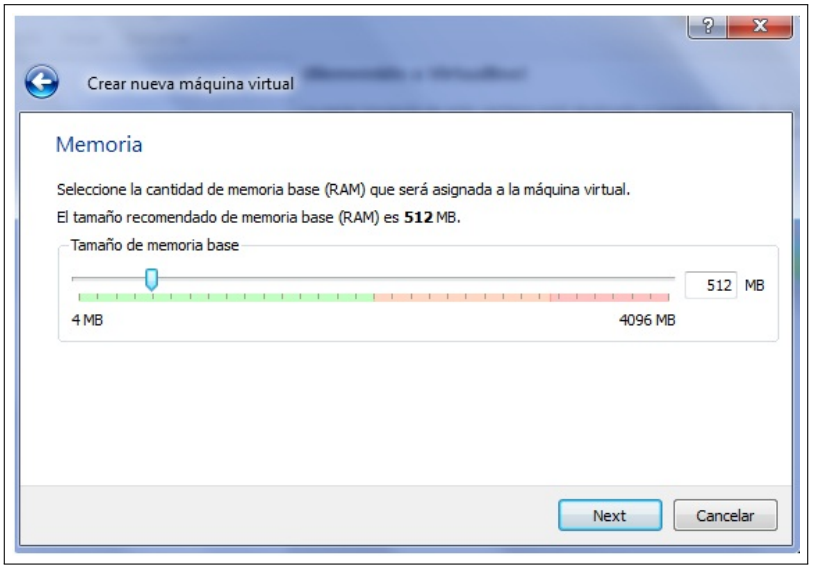

Figura B.4: Asignación de memoria a la máquina virtual

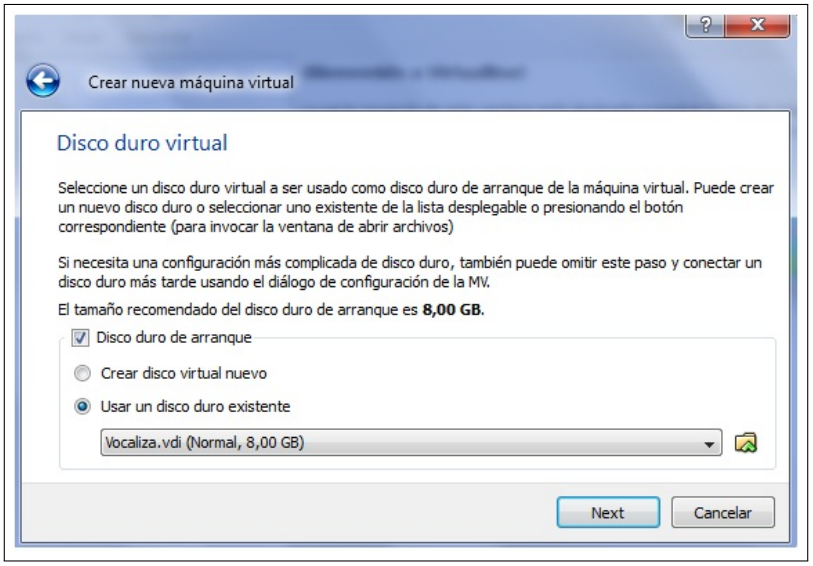

Figura B.5: Disco duro virtual

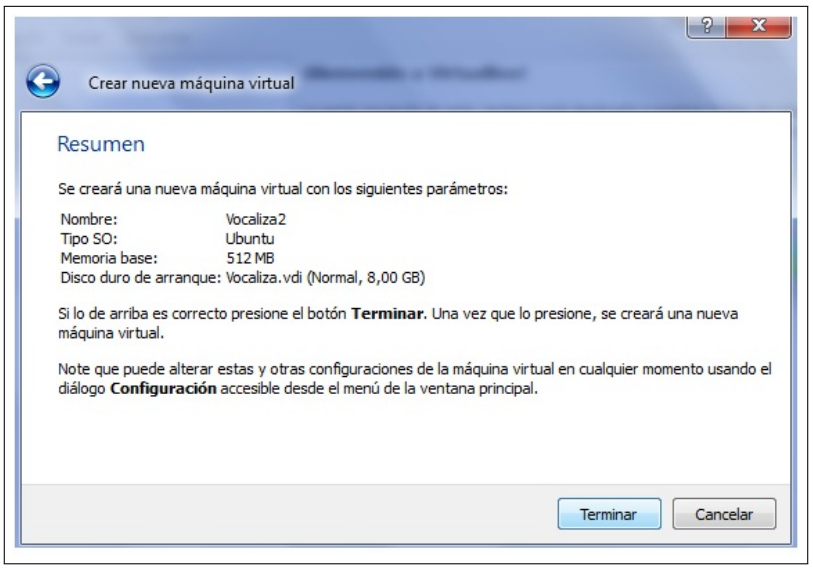

Figura B.6: Resumen nueva máquina virtual

- Al hacer click en siguiente accederemos al resumen de nuestra configuración de la nueva máquina virtual como se muestra en la figura B.6

- Hacemos click en terminar y volveremos a la ventana inicial, donde vemos nuestra máquina virtual instalada correctamente, figura B.7

B.1.3 Configuración de la red

- El siguiente paso es configurar la red, para ello haremos click en Configuración accediendo a la figura B.8

- En la ventana de red, seleccione la opción “conectado a: adaptador puente”. Para hacer un puente y que la máquina virtual sea vista como un equipo más de la red.

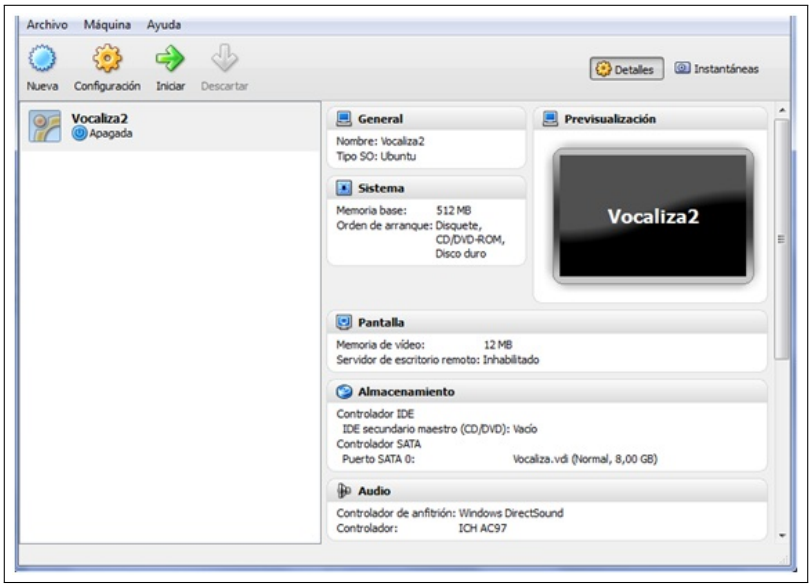

Figura B.7: Ventana inicio Oracle VM, con máquina virtual instalada

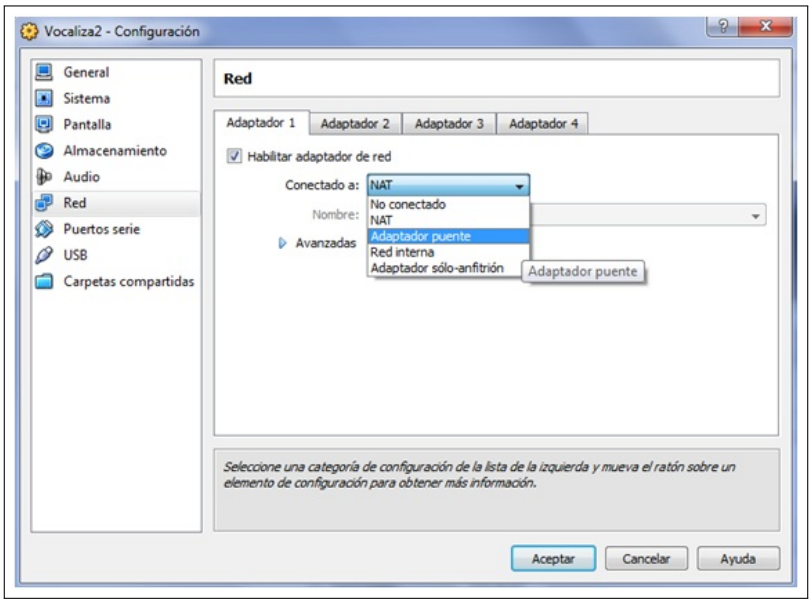

Figura B.8: Configuración de la red

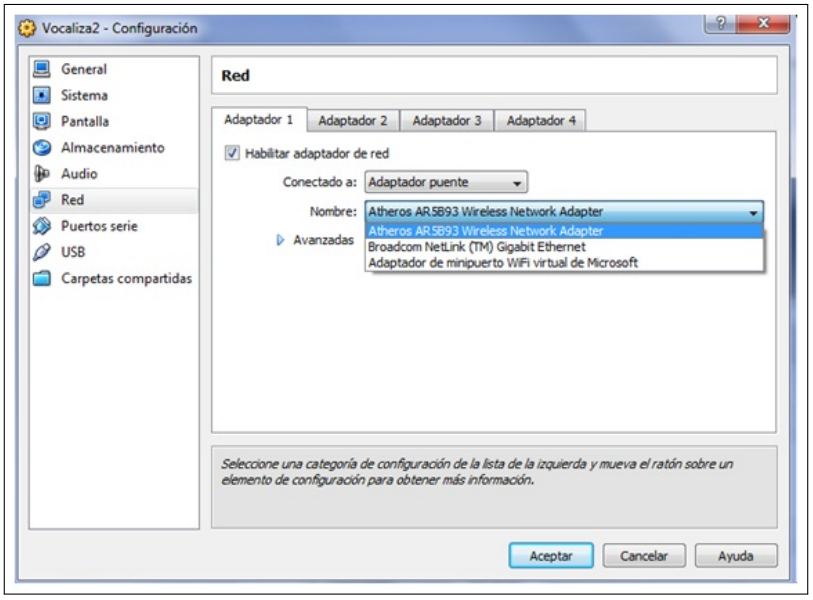

Figura B.9: Selección tarjeta de red

- En la opción "nombre" seleccione la tarjeta de red que utiliza para conectarse a internet. Figura B.9
- Por último ejecute su máquina virtual y abrá un terminal, figura B.10
- Ejecute el comando ifconfig. La dirección correspondiente a la que está marcada en rojo("Direccion inet: " en la figura B.11), será la dirección de la máquina virtual en su red.
- Deje corriendo la máquina virtual en el equipo y desde cualquier ordenador de su red entre a: http://sudireccionEnrojo/vocaliza y aparecerá la página principal de la aplicación.

Enhorabuena, la aplicación ha sido instalada correctamente.

El usuario avanzado que quiera realizar otro tipo de instalación tiene a su disposición el vocaliza.war pero requisito indispensable es una instalación mysql con una base de datos vocaliza, en el puerto 3306 de la misma máquina.

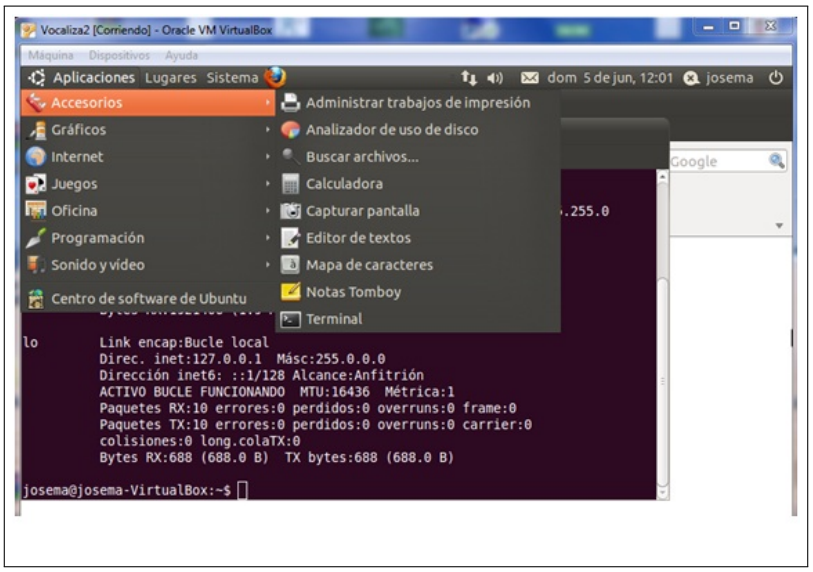

Figura B.10: Máquina virtual iniciada, ejecutar Terminal

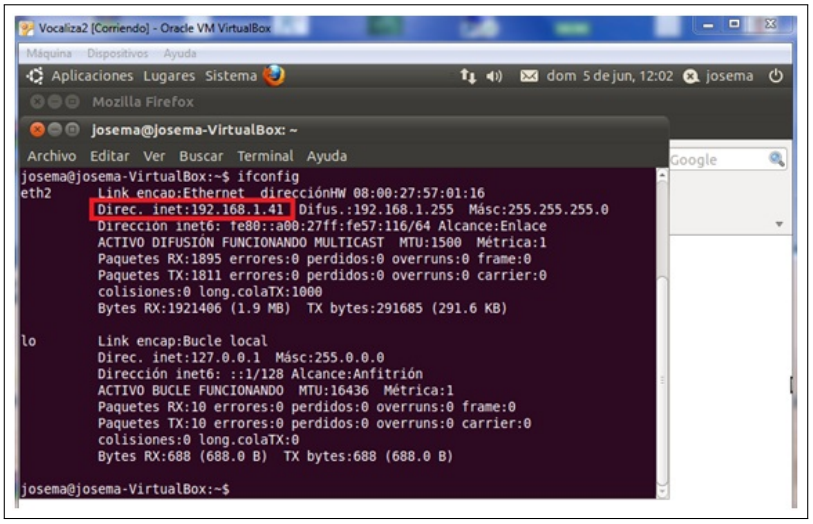

Figura B.11: Obtener dirección IP de la máquina virtual

Bibliografía

[1] C. Vaquero, O. Saz, E. Lleida, J.-M. Marcos, and C. Canalís, "Vocaliza: An application for computer-aided speech therapy in spanish language," in *Proceedings of the IV Jornadas en Tecnologías del Habla*, Zaragoza, Spain, November 2006.

[2] C. Vaquero, "Reconocedor de comandos orales para eliminar barreras de comunicación y movilidad en personas con discapacidades motrices y de comunicación," Proyecto Fin de Carrera, Departamento de Ingeniería Electrónica y Comunicaciones, University of Zaragoza, Zaragoza, Spain, 2006, Dirigido por O. Saz (Ponente E. Lleida).

[3] W.-R. Rodríguez, C. Vaquero, O. Saz, and E. Lleida, "Aplicación de las tecnologías del habla al desarrollo del prelenguaje y el lenguaje," in *Proceedings of the 2007 Congreso Latinoamericano de Ingeniería Biomédica (CLAIB)*, Isla Margarita, Venezuela, June 2007.

[4] A. Escartín, "Gestión de comunica: Conjunto de herramientas para la logopedia y ampliación de sus herramientas a los niveles semántico y pragmático del lenguaje," Proyecto Fin de Carrera, Departamento de Ingeniería Electrónica y Comunicaciones, Universidad de Zaragoza, España, 2008.

[5] Oscar Saz, José Enrique García, and Eduardo Lleida, *Accesibilidad en la web mediante voz a ciegos y personas con deficiencias visuales*, 2010.

[6] Aragón Radio 2, "http://www.aragonradio2.com/," 12 Agosto 2011.

[7] S. Öhgren, "Experiment with adaptation and vocal tract length normalization at automatic speech recognition of children's speech," M.S. thesis, Royal Institute of Technology, Stockholm, 2007.

[8] Spring Security Core, "http://grails-plugins.github.com/grails-spring-security-core/docs/manual/," 15 agosto 2011.

[9] ARASAAC, "http://www.catedu.es/arasaac/," 22 Agosto 2011.

[10] E. Lleida and R.-C. Rose, "Utterance verification in continuous speechrecognition: Decoding and training procedures," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 126–139, March 2000.

[11] Loquendo, *http://www.loquendo.com/es/productos/sintetizador-de-voz/perfiles-del-producto/*, 22 Agosto 2010.

[12] J. Holmes and Wendy Holmes, *Speech Synthesis and Recognition*, Taylor and Francis, second edition, 2001.

[13] H. Hon X. Huang, A. Acero, "Spoken language processing: A guide to theory, algorithm, and system development," *Journal of the Royal Statistical Society*, 2001.

[14] Q. Li and M. Russel, "An analysis of the causes of increased rates in children´s speech recognition," in *Proceeding of ICSLP 2002, Denver, páginas 2337-2340*, 2002.

[15] ViVoLab, "http://www.vivolab.es/demos/vivoapplet.html," 23 agosto 2011.

[16] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[17] J.-L. Gauvain and C.-H. Lee, "Map estimation of continuous density hmm: Theory and applications," *DARPA Speech and Natural Language Workshop*, 1996.