**UNIVERSIDAD POLITÉCNICA SALESIANA**

**SEDE CUENCA**

**CARRERA: INGENIERÍA DE SISTEMAS**

# DISEÑO Y DESARROLLO DE UN MÓDULO PROTOTIPO INTELIGENTE DE LECTURA DE CUENTOS (TTS) PARA APLICACIONES DE APRENDIZAJE PARA NIÑOS

DESIGN AND DEVELOPMENT OF AN INTELLIGENT PROTOTYPE FOR STORYTELLING (TTS) FOR APPLICATIONS TO CHILDREN'S LEARNING.

**TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE: INGENIERO DE SISTEMAS**


Autores:

Tania Elizabeth Flores Tapia.

Celia Elena Ordoñez Arce.

Tutor:

Ing. Vladimir Espartaco Robles Bykbaev.


Cuenca, Septiembre de 2016

## CESIÓN DE DERECHOS DE AUTOR

Nosotras Tania Elizabeth Flores Tapia con C.I. 010488791-4 y Celia Elena Ordoñez Arce con C.I. 070604240-5 manifestamos nuestra voluntad y cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autoras del trabajo de grado intitulado:" DISEÑO Y DESARROLLO DE UN MÓDULO PROTOTIPO INTELIGENTE DE LECTURA DE CUENTOS (TTS) PARA APLICACIONES DE APRENDIZAJE PARA NIÑOS", mismo que ha sido desarrollado para optar por el título de: Ingeniera de Sistemas, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En aplicación a lo determinado en la Ley de Propiedad Intelectual, en la condición de autoras nos reservamos los derechos morales de la obra antes citada. En concordancia, suscribimos este documento en el momento que hacemos entrega del trabajo final en formato impreso y digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, Septiembre de 2016

Tania Elizabeth Flores Tapia

C.I.:010488791-4

Celia Elena Ordoñez Arce

C.I.:070604240-5

**DECLARATORIA DE RESPONSABILIDAD Y AUDITORÍA.**

Nosotras Tania Elizabeth Flores Tapia con C.I. 010488791-4 y Celia Elena Ordoñez Arce con C.I. 070604240-5, autores del trabajo de titulación "DISEÑO Y DESARROLLO DE UN MÓDULO PROTOTIPO INTELIGENTE DE LECTURA DE CUENTOS (TTS) PARA APLICACIONES DE APRENDIZAJE PARA NIÑOS", certificamos que el total contenido de este Proyecto Técnico es de nuestra exclusiva responsabilidad y auditoría.

Cuenca, Septiembre de 2016

Tania Elizabeth Flores Tapia

C.I.:010488791-4

Celia Elena Ordoñez Arce

C.I.:070604240-5

**CERTIFICACIÓN.**

En calidad de TUTOR DEL TRABAJO DE TITULACIÓN “Diseño y desarrollo de un módulo prototipo inteligente de lectura de cuentos (TTS) para aplicaciones de aprendizaje para niños”, elaborado por Tania Elizabeth Flores Tapia y Celia Elena Ordoñez Arce, de claro y certifico la aprobación del presente trabajo de titulación basándose en la supervisión y revisión de su contenido.

Cuenca, Septiembre de 2016

Ing. Vladimir Espartaco Robles Bykbaev.

Tutor del trabajo de titulación

**Agradecimientos.**

Con una meta más cumplida del largo camino que deberemos recorrer, queremos expresar nuestros más sinceros agradecimientos a nuestro tutor de proyecto técnico Ing. Vladimir Robles por su paciencia, apoyo, guía y sobre todo amistad que nos ha brindado, a la Lcda. Gabriela Ortiz y a todos quienes conforman el Grupo de Investigación en Inteligencia Artificial y Tecnologías de Asistencia (GIIATA) de la Universidad Politécnica Salesiana – Sede Cuenca por brindarnos su amistad, consejos y apoyo desinteresado y sincero antes y durante el desarrollo de este proyecto. Además de manera muy especial agradecemos al Ing. Hernán Fajardo por su apoyo incondicional.

**Dedicatoria.**

*Dedico este trabajo de titulación, primero a Dios por estar conmigo en cada uno de los pasos que doy, por darme la fé y fortaleza necesaria para culminar mis estudios, además agradezco a mis padres Julio Flores y Narcisa Tapia, por haberme apoyado incondicionalmente todo este tiempo, por brindarme sus consejos y nunca dejarme caer, porque gracias a ellos soy una persona de bien, porque todo el sacrificio que realizaron ahora se ve recompensado y simplemente porque son los mejores padres.*

*A mis hermanas porque ellas sufrieron conmigo cada tropiezo y porque dejaron de lado sus ocupaciones por ayudarme. Finalmente, a mi esposo que estuvo dándome aliento desde el inicio de la carrera, porque vivió junto a mí cada uno de mis logros y fracasos, porque siempre tuvo las palabras correctas para levantarme.*

**Tania Elizabeth Flores Tapia.**

*Quiero dedicar esta tesis a mi familia, amigos y profesores quienes fueron una guía muy importante para la culminación d este proyecto como lo es Ing. Vladimir Robles. A mis amigos Hernán Fajardo y Elizabeth Andrade x brindarme su ayuda incondicional, a mi Esposo por su atención y ayuda continua.*

*De manera muy especial va dedicada a mi querida hija Rosy, por darme la fuerza para seguir adelante y a mi querida madre por su dedicación y ejemplo de lucha.*

**Celia Elena Ordoñez Arce.**

## Índice de Contenidos

## Índice de figuras



## Índice de tablas

## 1. Resumen

En la actualidad la tecnología se ha convertido en una herramienta muy importante al momento de conducir actividades para la recreación de los niños, ya que, por medio de estas herramientas y el apoyo de los padres de familia, se pueden ayudar a que sus hijos aprendan de una manera entretenida y divertida.

Desde hace muchos años la síntesis de voz ha jugado un papel crucial tanto en la comunicación como en el aprendizaje y entretenimiento de los niños, jóvenes y adultos. Día a día se busca crear una voz lo más natural/emocional posible, ya que esto ayudaría a mantener la atención y no cansaría al usuario.

En el presente proyecto se diseñó y desarrollo un módulo para la lectura y narración de cuentos que apoye el aprendizaje de los niños. Para el diseño de este módulo nos hemos planteado implementar un TTS prototipo desde cero, siguiendo la arquitectura de la herramienta FreeTTS y creando para ello un diccionario/base de datos de sílabas y palabras, mismas que serán grabadas por una locutora.

Dicha base de datos está diseñada para brindar una mejora en el sonido del motor de síntesis de voz, lo que posibilitará obtener una voz con más naturalidad e inteligibilidad para que sea agradable para los niños y sobretodo, para un TTS adaptado al dialecto de idioma español que se maneja en Ecuador. El aporte fundamental de este proyecto se centra en el análisis fonético y lingüístico que permitirá realizar la construcción de la palabra si ésta no existe en la base de datos.

Como siguiente punto se procederá a realizar el módulo para la lectura de cuentos, los mismos que serán analizados fonéticamente para poder ser relacionados con una base de datos, lo que constituirá un cuenta cuentos multimedia.

La idea es no perder la buena costumbre de leer cuentos a los niños, ya que esto permite ampliar su vocabulario y mejorar su lenguaje, pero con una voz que esté de acuerdo a nuestro entorno y relacionándole con material multimedia como imágenes. Esto permitirá ayudar aún más al niño a desarrollar su vocabulario relacionando las palabras o frases con imágenes/pictogramas.

En esta tesis se describe de forma detallada el proceso de diseño y construcción del sistema prototipo. En la primera parte se presenta una breve introducción acerca de la realidad actual de los sistemas para contar cuentos. En la segunda parte se describe el marco metodológico en el que se detalla el diseño y construcción del módulo para contar cuentos, y que también permite recomendar un cuento en base a la edad, sexo, nivel de vocabulario fonológico y de compresión y expresión, y el módulo TTS que es el complemento del sistema para contar cuentos a través de diagramas. Para esto, se emplean diversas capturas de pantalla del funcionamiento de los mismos e interfaces gráficas. Seguidamente, se describen las pruebas y resultados realizados con la herramienta. Finalmente, se presentan las conclusiones y los trabajos futuros que se podrían implementar.

## 2. Abstract

Nowadays, the technology has become one of the most important tools to improve activities for recreation of children, since through these tools and with appropriate parental support it is possible for children learning from an entertaining and a funny way.

For many years the speech synthesis has played key roles both in communication, learning and entertaining of children, youth and adults. Every day the researchers try to create most natural and emotional voices to be integrated to speech synthesis systems. This is be done with the aim of providing tools able to interact with humans in a natural way.

In this thesis we present the different stages that we have conducted to design and implement a Text To Speech (TTS) system to support the children's learning process through storytelling. In order to develop a tool to work with children, we propose a design that relies on two elements: the architecture of the FreeTTS system and the creation of a database of syllabes and words. Each syllabe and word will be recorded by a locutor. This base is designed to provide a better sound to the speech synthesis system, that would enable to get a voice with more naturally and intelligibility possible, to make more enjoyable for the kids and to get a speech synthesis system like a Spanish language that is used in Ecuador. The main contribution of this project focuses on the phonetic and linguistic analysis that will be able to build a word if it does not exist in the data base.

After this will be developed a module for storytelling, the same that will be analyzed phonetically and to be related to a database that contains the story, the images and an analysis with Freeling tools, and this will be perform a multimedia storytelling.

The idea is that it shouldn't lost the good habits of reading stories to the kids, this will improve their vocabulary and they will get better language skills, but with a natural voice and more like our enviroment and connecting with a multimedia images. This will helps the children vocabulary, keeping words linked with images or pictograms.

This thesis will give a the detailed description of the design and development of the system. The first part shows a short introduction about the reality of our days of the storytelling systems. The second part describes and details the process of the design and development of the storytelling module that also will suggest an specific story according the age, sex, vocabulary, and the levels of phonological, comprehension and expression, and a TTS module that complements the system. To show this are used some screenshots of the running system and graphical user interfaces. Following describes the testing and results. Finally, are being presented the conclusions and future work that will be implemented.

## 3. Introducción

En la actualidad los múltiples tipos de tecnologías existentes (tabletas, ordenadores, celulares inteligentes, etc.), juegan un importante papel e incluso ha permitido cambiar la forma de llevar nuestra vida diaria. Actualmente muchos maestros se apoyan en estos avances para facilitar su trabajo a través de material multimedia, como por ejemplo videos, audios, etc.; así mismo, el Ministerio de Educación del Ecuador incentiva más a esta tendencia estableciendo a las TICs como "*instrumentos para mejorar su tarea pedagógica a través de la aplicación de estrategias con el diseño de clases interactivas que incentivan el aprendizaje de los estudiantes en el aula* [1]."

Desde tiempos remotos la narración de cuentos, historias y leyendas ha sido una forma muy común y eficaz de entretenimiento; pero según [2], la narración ayuda a desarrollar el lenguaje expresivo de los niños como también su vocabulario, y además es considerado un método eficaz para aprovechar la imaginación de los niños autistas, según se explica en [7].

La lectura de cuentos mejora la creatividad, imaginación y ayuda a mejorar la concentración de los niños, además mejora el pensamiento crítico y habilidades de escucha [8]. El cómo se narra la historia, es decir, la expresividad, variedad del tono de voz, la relación con objetos o imágenes, imitaciones, etc. determinará la atención y el aprendizaje que el niño adquiera.

La narración de cuentos ha evolucionado a una lectura y narración digital gracias a la grabación en dispositivos tecnológicos y algo más novedoso aún, los sistemas de conversión de texto a voz (TTS, *Text to Speech*, por sus siglas en inglés). Estos sistemas permiten que un texto dado como entrada se convierta en voz.

Esta narración es una actividad de enseñanza-aprendizaje que muchos maestros aplican para enseñar lectura, escritura, gramática y matemáticas, o para presentar conceptos [6], ya que a más de la narración se puede integrar sonidos, imágenes, videos, etc. aportando diferentes modalidades de refuerzo para que un niño pueda aprender de mejor manera.

Los sistemas de conversión de texto a voz son de gran ayuda para un sinnúmero de aplicaciones desde el entretenimiento hasta el soporte para personas que tienen dificultad para comunicarse a través de su voz siguiendo el contexto de la narración. Como ejemplo de esto, en Edimburgo se desarrolló una aplicación móvil que narra las historias de la ciudad según su locación, este sistema se basa en la posición geográfica donde se encuentra el usuario y utiliza el API Georeferencing de Google [3].

En [4] se desarrolló una interfaz de juego de mesa que permite la interacción social mediante la narración interactiva empleando técnicas de inteligencia artificial. Esta aplicación permite que los niños creen sus propias historias.

En [5] se puede encontrar otro ejemplo en el que se presenta una narración de cuentos interactivos como una nueva forma de entrenamiento masivo digital que puede ser cargado en diversos dispositivos como tabletas, televisores, ordenadores, etc.

Por estas razones, los sistemas de conversión de texto a voz (TTS) han ido evolucionando diversos aspectos técnicos como pasar de una voz sintética a una voz con altos grados de naturalidad e inteligibilidad; siendo estas dos características muy importantes en el desarrollo o aplicación de estos sistemas. Como primera característica, la naturalidad es la cualidad por la que una onda sintética se parece a la voz humana (en factores como la entonación, el ritmo y la intensidad del habla) [9]. Y la inteligibilidad es la cualidad de poder entender la onda de voz sintética [9].

El uso de estos sistemas para la narración de cuentos/leyendas para los niños es menos expresiva, sensible, dulce, espontánea, etc., por el hecho que en el habla convencional intervienen no sólo músculos sino también órganos que permiten emitir el sonido; por esta razón se ha venido trabajando desde hace muchos años atrás en el desarrollo de un lector/narrador lo más humano posible apoyándose en métodos de prosodia, entonación y ritmo en la pronunciación, tratando de conseguir una voz más humana y evitando una voz robótica.

## 4. Objetivos

### 4.1. General

Diseñar y desarrollar un módulo prototipo inteligente de lectura de cuentos (TTS) para aplicaciones de aprendizaje para niños.

### 4.2. Específicos

- Diseñar un sistema de aprendizaje para niños sustentado en relato de cuentos, monitoreo y evaluación de comprensión.
- Conocer y estudiar las técnicas, y librerías que permiten realizar la conversión de texto a voz.
- Estudiar las técnicas de prosodia, entonación y conceptos relacionados, a fin de establecer las bases de los sistemas TTS.
- Diseñar y desarrollar un módulo prototipo que empleen técnicas y/o librerías de conversión de texto a voz a fin de realizar lectura de cuentos o historias para niños.
- Diseñar y desarrollar una interfaz gráfica para realizar el proceso de experimentación con el modulo prototipo.
- Diseñar y ejecutar un plan de experimentación que permita analizar y validar las funcionalidades provistas por el módulo prototipo.

## 5. Marco metodológico

Imitar la voz humana es una tarea muy compleja porque esta es una onda de presión acústica producida por el aparato fonador. La voz humana se compone del flujo del aire que sale de los pulmones pasando por unas cuerdas que vibran provocando distintos sonidos. Estas ondas son muy diversas ya que varían la frecuencia de la onda de sonido (voz grave o aguda), el timbre o la entonación dependiendo de la persona que lo emite [9]. Por esta razón, el centro de este trabajo es el desarrollo de un sistema de conversión de texto a voz para el idioma español que permita la narración de cuentos para niños, pero con una voz lo más cercana posible al habla convencional, ya que este proyecto tiene fines pedagógicos y terapéuticos. Todo esto ya que al usar voces sintéticas causan una percepción negativa del cuento generando hasta cierto punto temor en los niños evitando el fin de la herramienta.

Actualmente existen varias herramientas ya desarrolladas dentro de este campo, pero lamentablemente las que ofrecen un buen resultado en cuanto a la entonación de voz no son de distribución libre ni de código abierto, lo que dificulta su utilización para los propósitos generales de proyectos de este tipo y cuya finalidad es generar una herramienta de software libre para uso en instituciones educativas.

Se ha realizado un análisis de herramientas TTS de libre distribución, en cuanto a estas se ha podido constatar que son herramientas muy potentes pero en su mayoría tienen voces desarrolladas para idiomas como el Inglés, mismas que no son aplicables al sistema ya que al usar textos en español no solo se tiene que la voz sea poco natural sino que la entonación puede llegar a ser hasta incomprensible, en algunos casos se ha podido usar voces en idioma español y de las que se ha probado tienen una entonación con acento español.

Por las razones planteadas el centro de este trabajo es la creación de un sistema prototipo de conversión de texto a voz para el idioma español con acento ecuatoriano, que permita la narración de cuentos para niños, pero con una voz lo más cercana posible al habla convencional.

### 5.1. Diseño y desarrollo del sistema lector de cuentos.

Este módulo se encarga de integrar el TTS, los recursos multimedia y los contenidos de cada cuento, contiene la interfaz gráfica con la que interactúan los usuarios finales, esta es una aplicación de escritorio desarrollado en Java SE con tecnología swing, integra los siguientes componentes:

#### 5.1.1. Repositorio de recursos

- **Cuentos**: archivos de texto con el contenido de cada cuento, los mismos que serán leídos con el módulo TTS.
- **Recursos Multimedia**: imágenes y sonidos que se mostrarán de acuerdo al contexto del cuento, estos recursos se presentan por palabras clave del contenido del cuento, relacionándolos al texto a través de un análisis lingüístico. Para ello, se obtiene la raíz de cada palabra, es decir, cada vez

que se asocie una palabra a un recurso multimedia se reproduce o visualiza para que la experiencia del uso del sistema sea más interactiva.

### 5.1.2. Analizador Lingüístico

- **WordNet:** base de datos que permite relacionar palabras con sus sinónimos, antónimos o por sus posibles significados, esta base de datos cuyo esquema podemos ver en la ***Figura 1***, está siendo reutilizada de una tesis anterior presentada en [10].

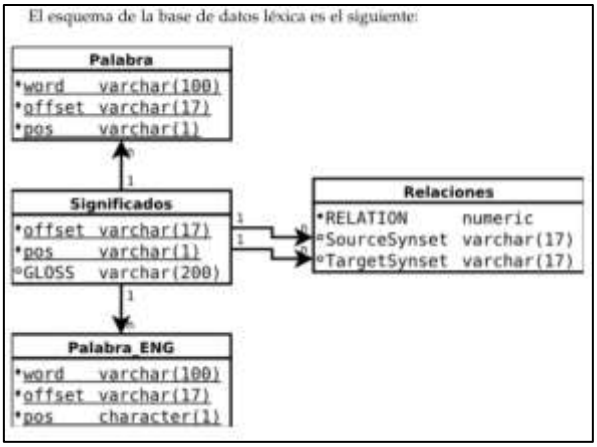


Figura 1. Esquema base de datos WordNet [10].

- **Freeling:** es una herramienta de Procesamiento del Lenguaje Natural, que permite llevar a cabo diversas tareas como el etiquetado gramatical, obtener los lemas o raíces de una palabra, análisis morfológico, detección de nombres, etc.

### 5.1.3. Funcionamiento

El software a través de diferentes medios (audio, imágenes, texto) se encarga de transmitir el contenido de un cuento, para esto se realiza lo siguiente:

1. **Carga de recursos:** carga la información de un cuento de un repositorio, esto incluye las imágenes, los audios y el texto como tal, en esta parte se tiene un archivo descriptor de recursos en formato XML que indica el tipo de recurso, la ruta en donde está almacenado y a qué palabra está relacionado como se observa en la ***Figura 2***.

```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<multimediaBD>
    <palabra>
        <id>1</id>
        <imagen>
            <id>1</id>
            <ruta>/Imagenes/Caperucita Roja.png</ruta>
        </imagen>
        <palabra>caperucita</palabra>
        <significado></significado>
        <sinonimos>
            <palabra>niña</palabra>
            <palabra> caperuza</palabra>
        </sinonimos>
    </palabra>
    <palabra>
        <id>2</id>
        <imagen>
            <id>2</id>
            <ruta>/Imagenes/caperuza.png</ruta>
        </imagen>
        <palabra>caperuza</palabra>
        <significado>chompa de color rojo</significado>
        <sinonimos>
            <palabra></palabra>
        </sinonimos>
    </palabra>
    <palabra>
        <id>3</id>
        <imagen>
            <id>3</id>
            <ruta>/Imagenes/abuela.png</ruta>
        </imagen>
        <palabra>abuelo</palabra>
        <significado>señora o señor de edad mayor</significado>
```

Figura 2. Archivo XML descriptor de recursos para la carga de información.

Cada recurso está ligado a una palabra de relevancia de un cuento, dicha palabra debe estar expresada en su forma de raíz en el archivo descriptor de recursos, ya que el sistema busca de manera automática en base a la raíz más no directamente a la palabra del texto (véase en la ***Tabla 1***):

| **Fragmento del texto de un cuento** | **Palabras raíces obtenidas** (se busca recursos multimedia en base a estas palabras) |
|---|---|
| iba caminando por el bosque | ir caminar por el bosque |

Tabla 1. Funcionamiento de la herramienta Freeling.

2. **Tokenización:** con el objetivo de sincronizar el TTS con la reproducción de recursos multimedia, se realizó una separación por frases del contenido del cuento, estas se van procesando una por una, de cada frase se realiza un análisis

para obtener los recursos multimedia a presentarse y se inicia el TTS para cada frase.

3. **Reproducción interactiva:** en esta parte para dar mayor dinamismo e interactividad al aplicativo se realizan 3 procesos en paralelo por cada frase tokenizada que son:
    a. **Iniciar TTS:** una vez obtenida la frase que se procesa, se inicia el TTS ya sea a través de una herramienta nativa o usando el Runtime del sistema para herramientas externas.
    b. **Análisis léxico y uso de multimedia:** de manera paralela a la ejecución del TTS se hace un análisis léxico combinando las dos herramientas mencionadas anteriormente (carga y tokenización), con la finalidad de asociar el texto de cada cuento con los recursos multimedia disponibles para ese cuento. Esto se hace analizando cada palabra leída por el TTS, obteniendo sus sinónimos con WordNet y la raíz de cada una de las posibles palabras usando Freeling, para que con ese dato se asocie a los recursos multimedia. Cada recurso multimedia está indexado por una palabra clave que coincide con una de las raíces de las palabras analizadas, en caso de que se obtenga alguna coincidencia entre el contenido del cuento con algún recurso, se visualiza la imagen o se reproduce un sonido dependiendo del caso.
    c. **Animación del texto:** El texto del cuento se presenta en una ventana usando HTML y CSS, conforme la lectura avanza se altera el estilo CSS de cada palabra en la que se encuentra para dar un efecto de animación resaltando la palabra actual, esto con el fin de poder seguir el texto conforme avanza la lectura.

### 5.2. Diseño y desarrollo del sistema de conversión de texto a voz.

Este módulo será el encargado de realizar la síntesis de voz para los cuentos, tomando en consideración que el proyecto tiene fines pedagógicos y terapéuticos, se requiere que la voz de este módulo sea lo más natural posible, con la finalidad de que los niños o cualquier persona que use el sistema tenga una mejor percepción del contenido de los cuentos, ya que al emplear voces sintéticas o robotizadas causan una percepción negativa del cuento, generando hasta cierto punto temor en los niños y afectando el proceso de enseñanza.

El proceso de desarrollo se inicia con el análisis de posibles soluciones al trabajo a realizar, por lo cual se ha investigado la arquitectura de un sistema TTS Figura 3, además de un análisis de herramientas de libre distribución y así poder saber qué parámetros debemos considerar a fin de obtener resultados satisfactorios.

#### 5.2.1. Síntesis de voz.

"*La síntesis de voz consiste en la creación de ondas de sonidos artificiales semejantes al habla humana*" [9]. Este tipo de sistemas son componentes cruciales para diferentes

aplicaciones, pero especialmente para aquellas enfocadas a la tecnología de asistencia, véase ***Figura 3*** [11].

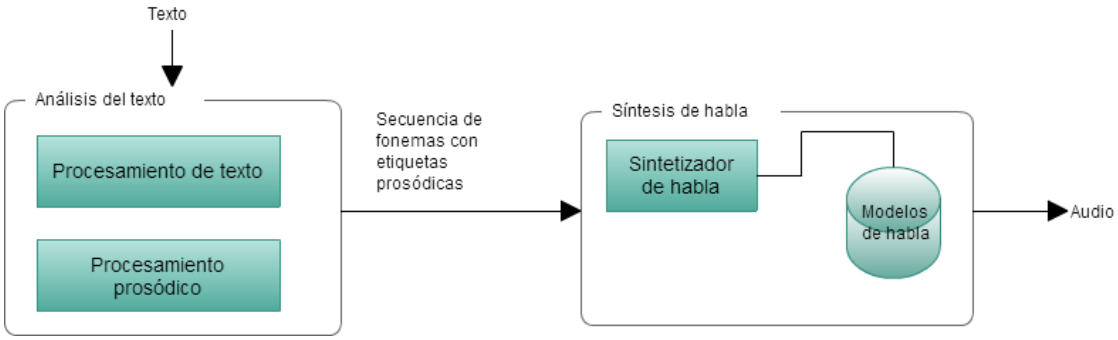


Figura 3. Arquitectura de un sistema TTS [11].

#### 5.2.1.1. Técnicas de entonación y prosodia

Según [26] la prosodia es la "*parte de la gramática que enseña la correcta pronunciación y acentuación*". Por ello, es una de las partes más importantes para un sintetizador de voz, dado que provee naturalidad e inteligibilidad al mismo. Para esto es necesario identificar características como género, edad, lugar de procedencia incluso identificar los estados de ánimo de la persona hablante, el estudio de estas técnicas servirán para aplicarlos en el TTS desarrollado, y así el sistema para contar cuentos permita que los niños al momento que escuchen la voz emitida no la sientan tan sintética, sino más bien, más natural.

Dentro de la prosodia lingüística intervienen los siguientes elementos como el acento, el tono, la entonación, las pausas, el ritmo, la velocidad de elocución[1] y la calidad de la voz. Uno de los aspectos fonéticos más directamente ligados a la entonación es la variación melódica que se consigue al ir cambiando la frecuencia de vibración de los pliegues o cuerdas vocales mientras se habla. Estas variaciones de tono aportan información lingüística. Aunque en las lenguas entonativas, como es el caso del español, las variaciones melódicas no permiten los contrastes léxicos o morfológicos que se observan en las lenguas tonales, sí aportan información lingüística en el ámbito del enunciado [23].

"*La entonación depende de las variaciones melódicas efectuadas al pronunciar los enunciados, que van asociados a cambios en la duración y en la intensidad de los sonidos emitidos. Aunque cada hablante posee una frecuencia media que le resulta cómoda y que depende de sus características articulatorias –en especial del grosor y la longitud de los pliegues vocales–, puede también, dentro de sus propios límites, modificar voluntariamente la altura del tono generado en la laringe. El tono de la voz puede resultar especialmente grave al final de una orden o muy agudo al final de una pregunta*" [23].

[1] Elocución: manera de hablar para expresar los conceptos.

Para identificar de mejor manera los cambios melódicos en la ***Figura 4*** se presenta un gráfico del oscilograma, espectrograma y contorno de F0 [2] de la secuencia auditiva [3]"*La mirada de Manolo la encandilaba emitida por una voz femenina*" [23]

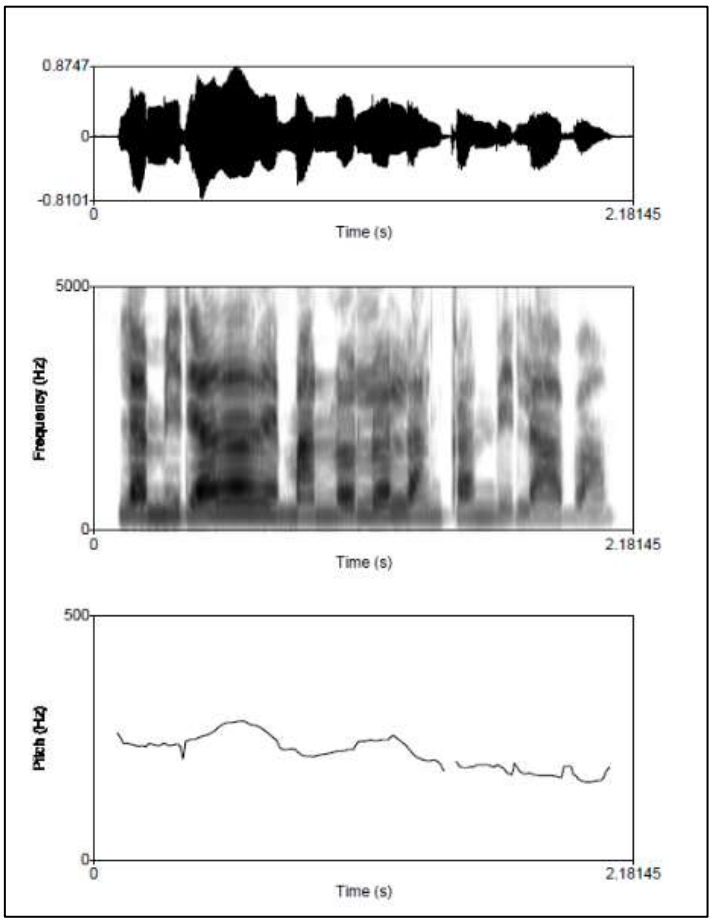


Figura 4. Oscilograma, espectrograma y contorno de F0 de la secuencia La mirada de Manolo la encandilaba emitida por una voz femenina [23].

La agudeza del oído humano para percibir cambios en los parámetros acústicos varía en función de la banda de frecuencias y del contexto, así como también la existencia de causas fisiológicas que provocan que la variación de algunos parámetros físicos sea consecuencia de la acción de otros.

**Métodos didácticos**: con la integración de las Nuevas Tecnologías de la Información y Comunicación (NTIC) se evidencian problemas pendientes de la práctica educativa como son definición de los objetivos, determinación del valor y sentido de los medios y delimitación de las prácticas consecuentes. Para realizar un análisis melódico de la voz se utilizan programas que ayudan a identificar la frecuencia del habla [23].

**Percepción Fonética y la inferencia emocional del mensaje**: "*el sonido producido por nuestro complejo aparato fonador llega al oído del receptor y se convierte, al final de*

[2] Contorno F0: representación de la melodía o curva melódica en los correlatos acústicos a lo largo del tiempo.
[3] Secuencia auditiva: detalle de secuencia de información auditiva.

*un largo proceso, en un mensaje lingüístico, es lo que denominamos fonética perceptiva. La audición pasa a percepción y de ahí a comprensión a través de unas etapas bien definidas que se integran proporcionando el sentido deseado en el receptor de los sonidos emitidos por el emisor*". [24]

**Proceso de la Percepción fonética**: el proceso acústico llega al oído externo (proceso mecánico) a través de determinadas vibraciones. Luego de ello, se transforma en un proceso hidráulico (medio acuoso) cuando llega a la cóclea y finalmente, en el proceso más complejo se transmuta en una información electroquímica a través del órgano de Corti cuando llega (mediante los complejos mecanismos de comunicación neuronal) a la corteza cerebral que realizará la decodificación de los sonidos percibidos, "clasificará" las percepciones y proporcionará una comprensión pretendidamente coherente de los signos recibidos [24].

En el proceso se pueden identificar los siguientes tipos prosódicos detectados que se consideran en el proceso de información/formación [24]:

- Neutro
- Alegría, agrado
- Interrogación, duda
- Burla
- Continuidad
- Asombro, admiración
- Enfado
- Tristeza
- Desagrado

**Modelado y estimación de la prosodia**

La investigación en este campo se centró inicialmente en conseguir el mayor grado de inteligibilidad posible. Conseguir un audio de mayor naturalidad, pudiendo así emular "*la riqueza del habla humana que es intrínsecamente expresiva, ya que posee la capacidad de complementar la información verbal con una intención, actitud o estado emocional determinados*". En este contexto, la mejora de la expresividad de los sistemas TTS se debe a avances en el modelado de la prosodia y la generación de la señal de voz de una alta calidad [25].

Como podemos observar en la ***Figura 5*** los parámetros que determinan la prosodia de un texto hablado son esencialmente la duración e intensidad, segmentar el posicionamiento además de la duración de las pausas y el contorno de F0. En el ámbito de los sistemas

TTS, la literatura en modelado prosódico es muy extensa. La curva de entonación es el parámetro prosódico con una mayor presencia, distinguiéndose entre métodos cuantitativos y métodos cualitativos. Para el modelado de la duración se han utilizado métodos basados en reglas, o métodos estadísticos tales como redes neuronales o árboles de clasificación y regresión [25].

| Etiqueta | Atributo | Tipo |
|---|---|---|
| F0 | Fonema anterior | D |
| F1 | Fonema actual | D |
| F2 | Fonema siguiente | D |
| ACENTUADO | Fonema acentuado | B |
| GA-en-GE | Posición de GA en GE | D |
| FON-en-GE | Posición de FON en GE | D |
| **DURACION** | Duración del fonema en $ms$ | N |
| Etiqueta | Atributo | Tipo |
| F1 | Fonema actual | D |
| ACENTUADO | Fonema acentuado | B |
| GA-en-GE | Posición de GA en GE | D |
| FON-en-GA | Posición de FON en GA | D |
| FON-en-GE | Posición de FON en GE | D |
| **INTENSIDAD** | Intensidad *rms* | N |
| Etiqueta | Atributo | Tipo |
| TIPO-GE: | Tipo of GE | D |
| GA-en-GE | Posición de GA en GE | D |
| ACENTO | Posición de la sílaba tónica | D |
| GA-en-FRA | Posición del GA en la frase | D |
| NUM-SIL | Número de sílabas del GA | N |
| $a_0, a_1, \ldots a_n$ | Coeficientes del polinomio | A |

Figura 5. Parámetros que determinan la prosodia [25].

Para el desarrollo del módulo se ha considerado analizar los siguientes sistemas TTS:

- **FreeTTS.-** "*Es un sistema de síntesis de voz escrito completamente en lenguaje de programación Java*" [12]. Soporta solo el idioma inglés. En la ***Figura 6*** se describe el esquema de la arquitectura de FreeTTS.

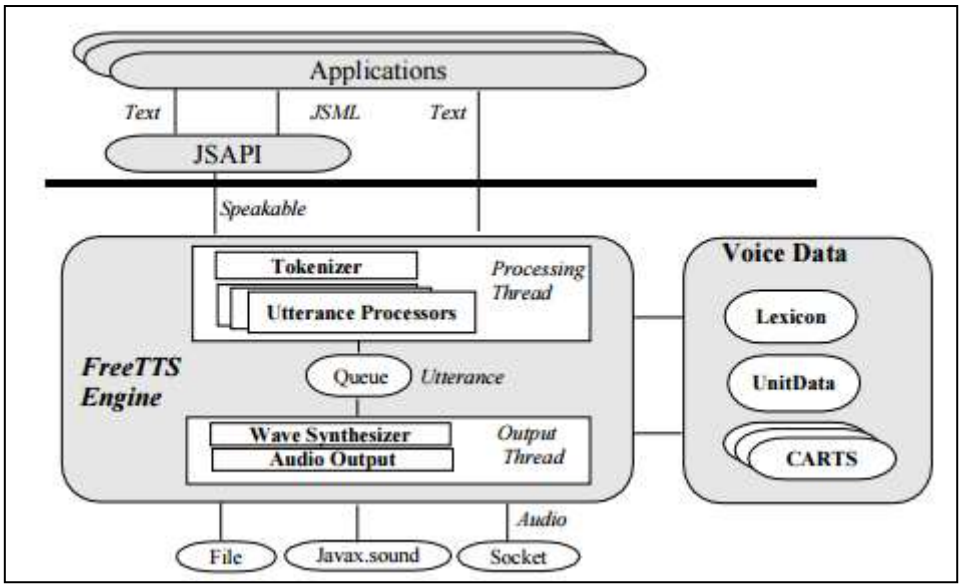


Figura 6. Arquitectura FreeTTS. [13]

- **MarysTTS.-** "*Es un sistema de código abierto, multilingüe de síntesis de texto a voz escrito en Java*" [14]. Soporta los idiomas inglés, telugú, turco y ruso. Funciona bajo un esquema cliente-servidor, la ***Figura 7*** describe la arquitectura del sistema antes mencionado

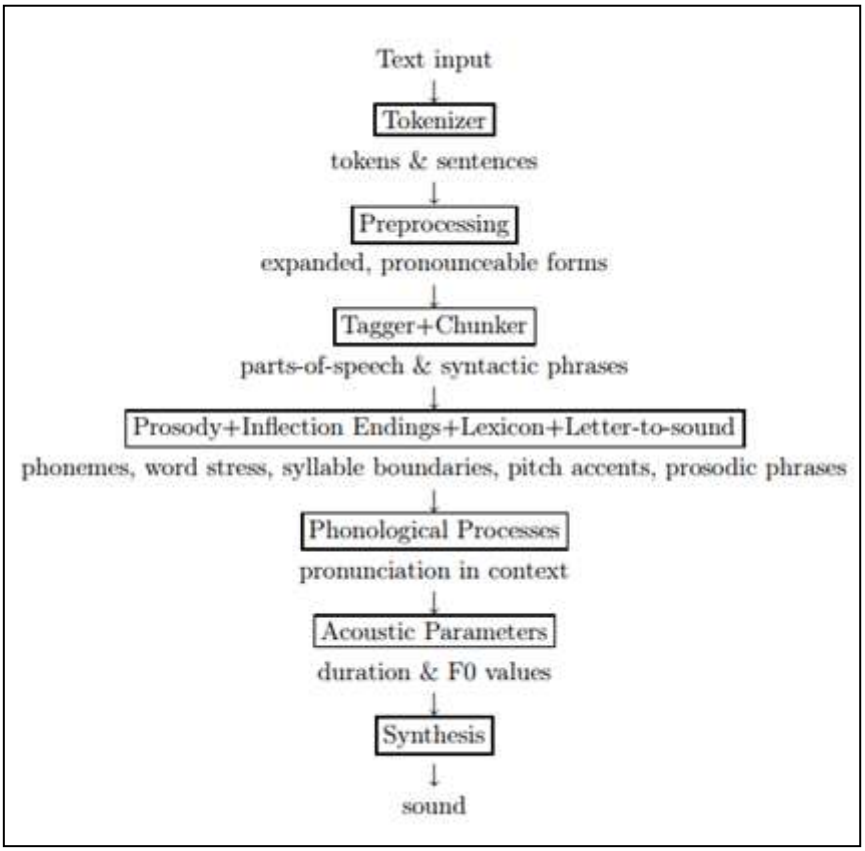


Figura 7. Arquitectura MaryTTS. [14]

- **Festival TTS.-** *"Es un sistema de síntesis de voz de propósito general para múltiples lenguajes, escrito en lenguaje C++"* [15]. Soporta inglés y castellano.

A continuación se detalla dos propuestas iniciales que se realizaron pero no se obtuvo los resultados esperados:

### 5.2.2. Análisis de mejora de voz en un TTS Libre

Como primera alternativa se analizó la factibilidad de mejorar una voz de un TTS libre, se ha optado por analizar 3 herramientas las cuales han sido especificadas anteriormente en la parte donde se describen los sistemas TTS.

Con las herramientas antes mencionadas se realizó un análisis de cuál sería el instrumento más viable para poder mejorar la voz que dicho sistema ya implementa, para lo cual se escogió Festival TTS, por medio de comandos se modificó la voz, tratando de retardar el tiempo de lectura. Sin embargo, no se obtuvieron resultados precisos, ya que a medida que se fueron realizando pruebas se llegó a la conclusión que este sería un proceso demasiado extenso y que llevaría más tiempo realizarlo, alejándonos así de los objetivos previamente establecidos para el proyecto.

### 5.2.3. Análisis de reutilización de código

Como primer paso se debió escoger una herramienta y se optó por reutilizar el código del sistema FreeTTS, porque es multiplataforma, desarrollado en Java, por su arquitectura y facilidad de adaptación en cuanto a código. Entre los pasos que se realizaron fue la descarga del código fuente y binario de dicho sistema, en el cual se manipularon y crearon algunas clases para hacer las respectivas pruebas, este sistema permite importar voces en inglés creando así una pequeña complicación por lo que no podríamos agregar una nueva voz, como resultado se logró cambiar el idioma de la voz del sistema a español. Sin embargo, el inconveniente fue que no sonaban con acento latino y habría complicaciones a futuro con fonemas como (ll, ñ, z).

#### 5.2.3.1. Solución propuesta

Siguiendo la arquitectura global del sistema de síntesis de voz FreeTTS se propone realizar un TTS desde cero, por lo cual se ha planteado el siguiente módulo (véase ***Figura 8***).

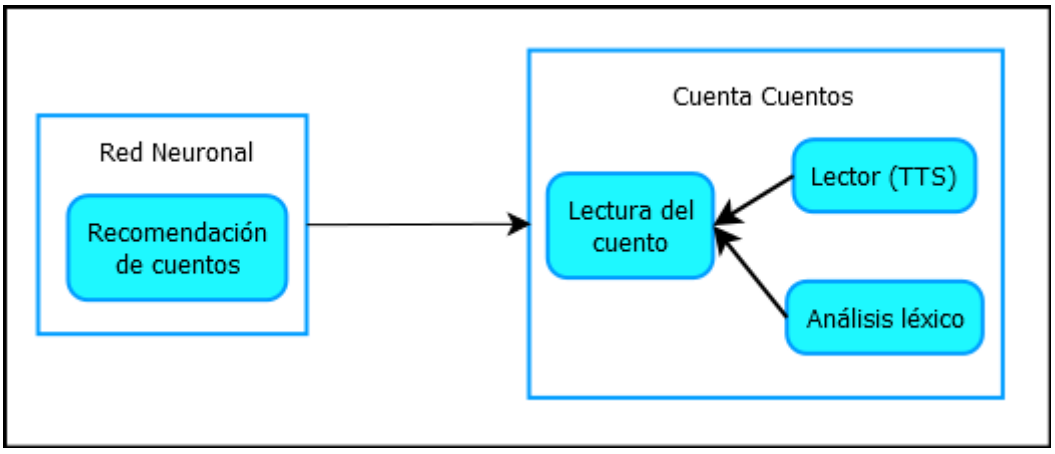


Figura 8. Solución propuesta

#### 5.2.3.1.1. Módulo TTS

El módulo TTS se desarrolla en base a características tomadas de otros TTS, obteniendo el diseño que se puede observar en la ¡Error! No se encuentra el origen de la referencia.:

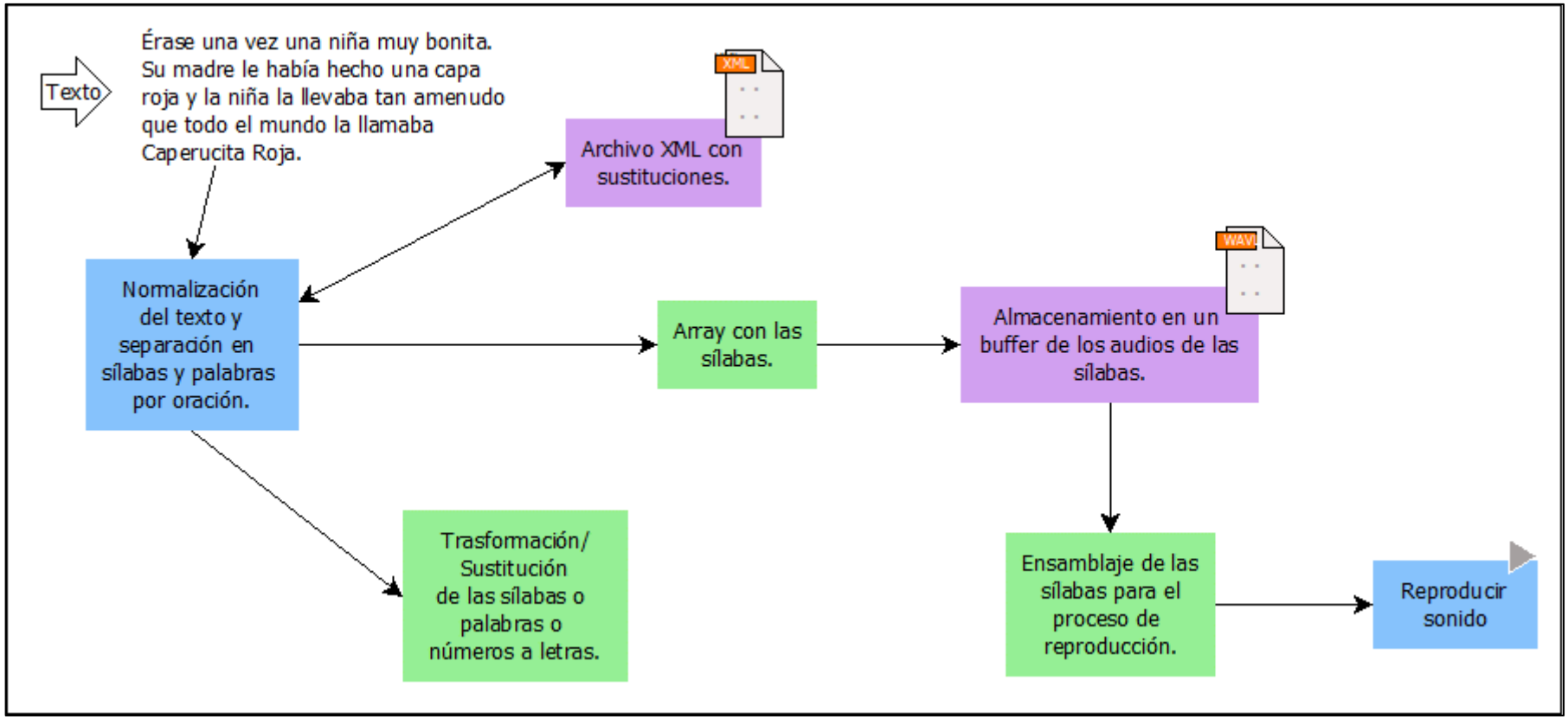


Figura 9. Descripción del funcionamiento del módulo TTS.

#### 5.2.3.1.1.1. Análisis o Normalización del texto.

Este proceso no es trivial, ya que los textos están llenos de homógrafos, números y abreviaturas que tienen que ser transformados en una representación fonética.

Para ello es necesario separar el texto en frases y estas en palabras, para posteriormente transformar las expresiones numéricas, abreviaturas, siglas y acrónimos a sus correspondientes transcripciones ortográficas, de acuerdo a una base de datos almacenada en un archivo XML (véase ***Figura 10***). Si dicha expresión existe la reemplazará, caso contrario dejará la original. Además, los signos de puntuación no deseados (“” () [] {}) son eliminados para evitar confusión en el momento de la lectura.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<sustituciones>
    <sustitucion>
        <original>av</original>
        <reemplazo>ab</reemplazo>
    </sustitucion>
    <sustitucion>
        <original>avs</original>
        <reemplazo>abs</reemplazo>
    </sustitucion>
    <sustitucion>
        <original>avz</original>
        <reemplazo>abs</reemplazo>
    </sustitucion>
    <sustitucion>
        <original>abz</original>
        <reemplazo>abs</reemplazo>
    </sustitucion>
    <sustitucion>
        <original>ak</original>
        <reemplazo>ac</reemplazo>
    </sustitucion>
```

Figura 10. Archivo XML de sustituciones.

Como primera alternativa se separó el texto en oraciones, estas en palabras y las palabras en fonemas, pero esta opción no fue la más apropiada, dado que la unión de dos o más fonemas es más complicado por cómo suena el fonema, es decir, es más fácil pronunciar “be” que “b”. Por esta razón se tomó la decisión de separar las palabras en sílabas para su correspondencia con un archivo de audio o directamente su correspondencia de la palabra con el archivo de audio, claro que con esta segmentación la base de datos de audios sería mucho más amplia, pero el resultado de la voz es mucho más agradable y a la vez claro. Por ejemplo, si tenemos un párrafo, este será dividido en oraciones, dichas oraciones en frases, estas frases en palabras y a su vez estas palabras en sílabas o palabras si existiesen en la base de datos, como se ve en la ¡Error! No se ncuentra el origen de la referencia..

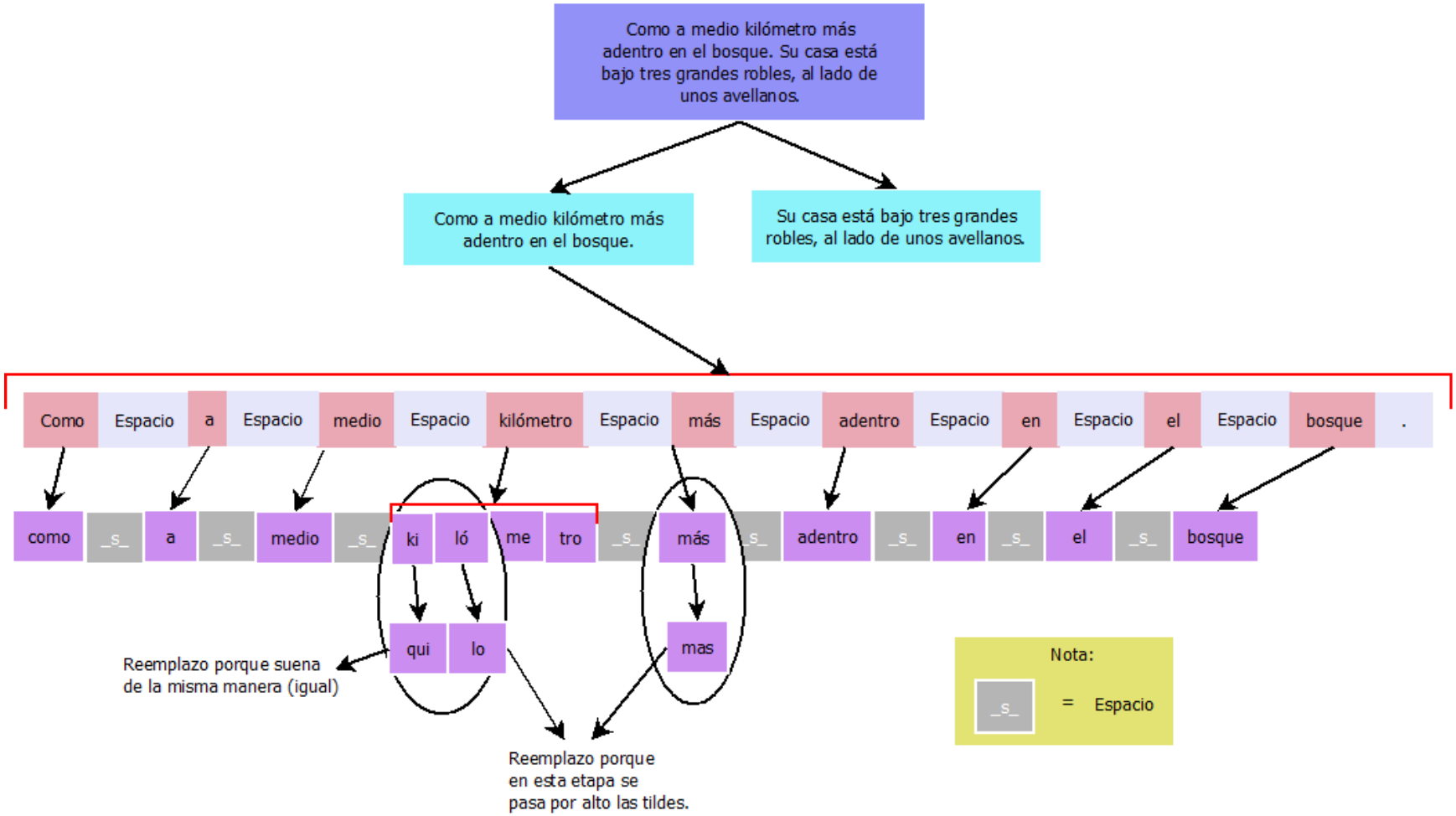


Figura 11. Ejemplo de normalización y transformación del cuento a frases, estas a oraciones y estas a sílabas o palabras (síntesis concatenativa por selección de unidades con base de datos de sílabas y palabras).

#### 5.2.3.1.1.1.Síntesis.

Existen dos tipos de síntesis para la generación de la voz del TTS [9]:

- Síntesis concatenativa: se basa en la unión de segmentos de voz grabados, además este método produce una síntesis más natural y no tiene un modelo matemático. Existen tres métodos para realizar este tipo de síntesis:
  - **Síntesis por selección de unidades**.- utiliza una base de datos en la que se encuentran grabaciones de voz tanto de fonemas, sílabas, palabras, frases y oraciones. La ventaja de este método es que produce un sonido más natural, pero las bases de datos suelen ser muy grandes.
  - **Síntesis por difonemas.-** utiliza una base de datos mínima en la que se encuentran grabaciones de voz de difonemas. La desventaja de este método es que produce una poco natural y parcialmente inteligible.
  - **Síntesis específica para un dominio.-** se tiene una base de datos con palabras y oraciones grabadas. Utilizado para ámbitos muy limitados como por ejemplo un sistema de sonido que de las horas.
- Síntesis de formantes: Este método utiliza un modelo acústico, por lo que se crea una onda de habla artificial. No usa muestras de habla humana.

El presente proyecto utiliza la síntesis concatenativa por selección de unidades, ya que como ya se ha mencionado el objetivo es que la voz del sistema sea lo más natural posible y este método es el más apegado a nuestro objetivo.

#### 5.2.3.1.1.2.Creación de base de datos

- **Grabación de un corpus de voz en español.**

Esta fase consiste en la preparación de un corpus en español, y la grabación del mismo. Como la opción más viable se realizó la grabación de todos los fonemas, pero debido a la complejidad que tiene expresar o modular un fonema para grabarlo y luego manipularlo, se ha decidido usar sílabas y palabras. Internamente el software realiza una separación de estructuras semánticas en frases, oraciones y sílabas. Se trató de abarcar todas sílabas y palabras respecto al ámbito de cuentos infantiles grabando así aproximadamente 1800 sílabas y 1585 palabras.

En nuestro dialecto existen algunos fonemas que se pronuncian de la misma manera, según [17], son:

- La z por la s: “en toda Hispanoamérica representa el sonido predorsal fricativo sordo /s/: zapato [sapáto]. Este fenómeno recibe el nombre de «seseo»”.
- La v por la b: “representa el sonido consonántico bilabial sonoro /b/”.
- La y con la i: “en posición inicial de palabra o de sílaba representa el sonido consonántico palatal central sonoro /y/. Este mismo sonido puede representar el

grupo gráfico hi en posición inicial de palabra seguido de e, o la letra i en esta misma posición, seguida de a, o".

- La ll con y: "Además, en casi todo el mundo hispánico el dígrafo ll se pronuncia como /y/, fenómeno que se conoce con el nombre de «yeísmo»".
- La j con la g: "representa también la letra g ante e, i".
- La c con la s: "Cuando precede a las vocales a, o, u (casa, comer, cuerdo), va ante consonante (cráneo, acción, acné) o está en posición final de palabra (frac, vivac, chic), representa el sonido velar oclusivo sordo /k/. En toda Hispanoamérica representa el sonido predorsal fricativo sordo /s/ (cena [séna], aciago [asiágo])".
- La q con la k: "ante las vocales e, i, un dígrafo que representa el sonido velar oclusivo sordo /k/; la u no se pronuncia en estos casos: queso [késo], esquina [eskína].".

Para el desarrollo de esta fase se utilizó el programa Adobe Audition[4] y una cabina de grabación profesional de la Universidad Politécnica Salesiana, a fin de disminuir el ruido exterior y cualquier otra señal extraña a la voz, permitiendo así que ésta se escuche con mayor claridad. De igual forma, es necesario definir los siguientes parámetros:

- **La frecuencia de muestreo.-** *número de muestras por unidad de tiempo que se toma de una señal continúa para producir una señal discreta, durante el proceso para convertir una señal de analógica a digital*.[18]
- **Los bits de resolución.-** *número de bits utilizados para almacenar cada muestra* [19].
- **Número de canales.-** *número de pistas muestreados. Puede ser monofónico(un solo canal) o estereofónico(dos canales)*
- **Formato.-** *extensión en que el archivo será guardado ya se comprimiendo o no la información* [20].

La grabación la realizó una locutora que debía mantener el volumen y tono de voz durante la grabación para conseguir una clara pronunciación. La frecuencia de muestreo de grabación fue de 16 KHz y 16 bits de resolución en canal estéreo, en formato WAV[5] que no comprime la información por lo tanto no existirá pérdida de información. Con el objetivo de mejorar la calidad de la grabación, se utilizó el programa Adobe Audition y se aplicaron una serie técnicas como:

- **Reducción de ruidos.-** eliminar sonidos de fondo, murmullos, viento, etc.
- **Normalización de ondas del audio.-** establecer el porcentaje del pico más alto relativo a la amplitud máxima posible [16].

[4] Adobe Audition: programa orientado al mundo profesional para la grabación y edición de audios.
[5] WAV: formato de audio mono canal, que no comprime la información, recomendado para grabar la voz humana, y formato que tienen los audios usados en este proyecto.

- **Aplicación de efecto** como por ejemplo amplificación[6], compresor multibanda[7], etc.

- **Extracción de audio y etiquetado de las sílabas y palabra.**

En el proceso de etiquetado se delimitan las fronteras para cada unidad fonética y se le asigna un nombre, esto se lo realizó en un archivo XML, como se ve en la ***Figura 12*****Figura 1**, mismo que se usó para la extracción de cada una de las sílabas y palabras. Para la extracción de las sílabas y palabras se usó la librería FFMPEG[8], esta librería recorta el audio de acuerdo a la etiqueta <tiempoInicio> <tiempoFin> y asigna un nombre en función de la etiqueta <fonema>. De esta manera se construyó la base de datos de 1800 sílabas y 1585 palabras, como se puede apreciar en la ***Figura 13***.

```
</indiceEntry>
<indiceEntry>
    <fonema>
        <fonema>ñ</fonema>
        <pronunciacion>normal</pronunciacion>
        <tipo>S</tipo>
    </fonema>
    <tiempoFin>1904353</tiempoFin>
    <tiempoInicio>1903992</tiempoInicio>
</indiceEntry>
<indiceEntry>
    <fonema>
        <fonema>n</fonema>
        <pronunciacion>normal</pronunciacion>
        <tipo>S</tipo>
    </fonema>
    <tiempoFin>1895544</tiempoFin>
    <tiempoInicio>1895196</tiempoInicio>
</indiceEntry>
<indiceEntry>
    <fonema>
        <fonema>m</fonema>
        <pronunciacion>normal</pronunciacion>
        <tipo>S</tipo>
    </fonema>
    <tiempoFin>1892281</tiempoFin>
    <tiempoInicio>1891920</tiempoInicio>
</indiceEntry>
```

Figura 12. Archivo XML para la extracción y etiquetado de las unidades fonéticas.

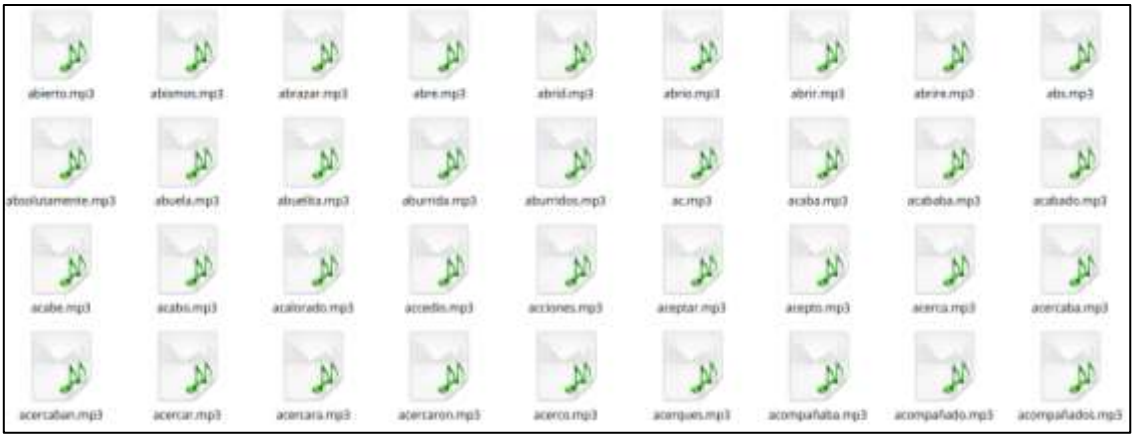


Figura 13. Ejemplo de base de datos de audio.

[6] Amplificación: aumenta o atenúa una señal de audio.
[7] Compresor Multibanda: permite comprimir cuatro bandas de frecuencia diferentes de forma independiente.
[8] Ffmpeg: software libre que permite grabar, convertir y hacer streaming de audio y video.

#### 5.2.3.1.2. Funcionamiento del programa

1. **Iniciación del Módulo**

El sistema está desarrollado usando paradigmas de programación orientada a objetos, está basado en capas modulares que se integran a través de interfaces, por ello cada funcionalidad puede ser actualizada con una nueva. Por ejemplo, para la reproducción de sonidos, el sistema maneja una interfaz con los métodos básicos de un reproductor de audio, por lo que se puede tener varios reproductores o componentes simultáneos permitiendo que el programa sea más escalable.

Tomando en cuenta lo anterior, se carga el reproductor que se haya parametrizado, la base de datos, estableciendo la ruta donde están los fonemas en formato MP3, se establece el listado de cambios que se realizarán en lo que se refiere a fonemas o palabras, todo esto en base a los archivos de configuración que se tienen.

2. **Separación**

Se realiza un proceso de separación en sílabas usando un algoritmo que las reconoce en base a la estructura de la palabra, identificando vocales y consonantes, además analizando si existen casos de hiatos o diptongos, dando como resultado un vector de sílabas con el texto ya normalizado.

El sistema tiene en cuenta que cada punto (.) es el final de una frase en la que se realizará una pausa, de la misma manera actuará al encontrar una coma (,), un punto y coma (;), dos puntos (:).

3. **Reemplazo de fonemas**

Cada una de las sílabas separadas, serán sustituidas por su correspondiente homónima si existe, es decir, por ejemplo \u201cva\u201d por \u201cba\u201d ya que las dos se pronuncian de la misma manera y tener dos audios diferentes de la sílaba sería innecesario.

4. **Asociación de sílabas a fonemas**

Los archivos de audio son copiados desde la carpeta donde se almacena todos los audios y almacenados en un buffer temporalmente hasta completar todos los archivos del texto para reproducir un solo sonido del texto completo como se muestra en la ***Figura 14***.

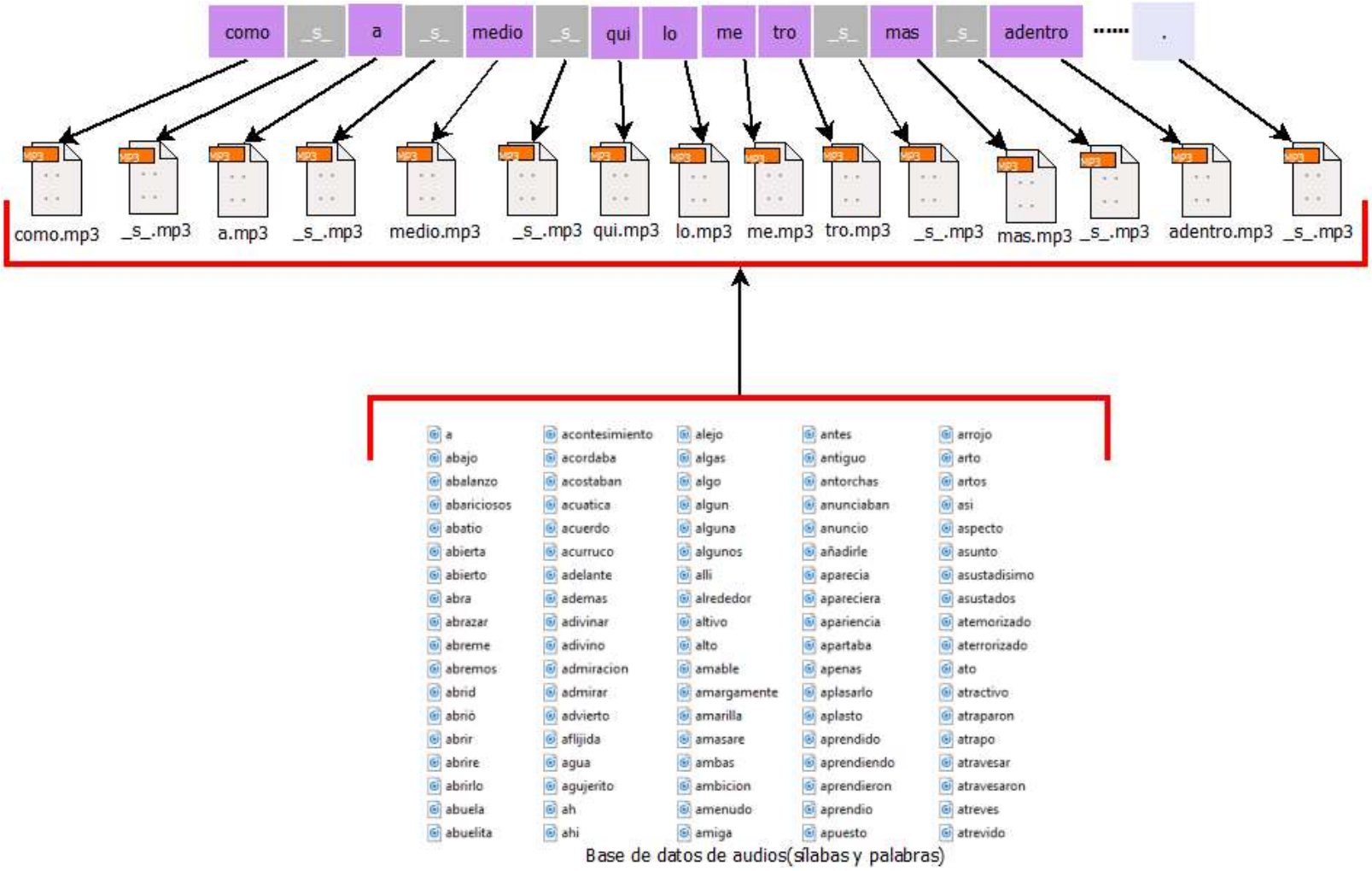


Figura 14. Ejemplo correspondencia de cada sílaba o palabra con su archivo de audio.

5. **Generar nuevos fonemas**

Si el sistema no encuentra el audio de la sílaba que lee, procede a armar un nuevo archivo siempre y cuando disponga de los fonemas base.

Como primer paso el sistema toma los fonemas de la palabra y forma la misma almacenándola a está en la base de datos, para que la próxima vez si se busca formar la misma palabra no se realice de nuevo el mismo proceso, sino que se lee directamente la palabra ya antes formada, permitiendo tener una base de datos combinada entres sílabas y palabras, garantizando un procesamiento más óptimo y con menos operaciones.

6. **Unir fonemas en un archivo de audio único**

En consideración a esto se toma un listado de todos los archivos de cada sonido generados en el paso 4, que al final de este proceso con un software conocido como FFMPEG se manda a unir todos los archivos del listado en un nuevo archivo de audio.

7. **Procesar y Filtrar audio**

Posterior a la unión con FFMPEG se hace un filtrado de sonidos en blanco o sonidos muertos ya q al utilizar el FFMPEG adiciona espacios en blanco entre los diferentes fonemas teniendo una duración promedio de hasta 5 milisegundos de sonidos muertos, se usa un software conocido como SOX-Sound eXchange[9] para realizar este proceso, una vez realizado esto se genera un nuevo archivo que contiene el audio final.

8. **Reproducir el sonido generado**

Para la reproducción de sonidos el sistema utiliza el API *basic player*[10] del paquete de librerías javazoom[11].

## 5.3. Diseño y desarrollo del sistema inteligente para el relato de cuentos.

Como parte del sistema inteligente, se ha decido implementar aprendizaje automático a través de una red neuronal artificial, la cual permite determinar, qué cuento es apto para el niño que usará la aplicación. Esta es una de las tareas más relevantes del proyecto, dado que gracias al aprendizaje automático y clasificación de cuentos, se ayudará y facilitará el uso de la misma, sin la supervisión de un adulto para que el niño escuche el cuento.

Se usó el programa de Minería de Datos Weka para crear la Red Neuronal, usando su librería escrita en código Java.

Weka es una herramienta Software Libre que cuenta con algoritmos de Machine Learning, y que permite, crear y entrenar una Red neuronal con algoritmos como el Perceptrón multicapa o el SMO el cual es el algoritmo que usa WEKA y el que se eligió como el algoritmo clasificador de la red neuronal.

---

[9] SOX - Sound eXchange: software multiplataforma que permite aplicar diversos efectos a archivos de audio mediante línea de comandos.

10 BasicPlayer: API de alto nivel basado en JavaSound para reproducir, detener, pausar, reanudar y buscar archivos de audio.

11 Javazoom: aplicación de reproducción de música desarrollada en la plataforma Java.

“SMO implementa el algoritmo de optimización secuencial mínima de John C. Platt para la formación de un vector clasificador de soporte usando una escala polinomio núcleos” [22].

“A través del algoritmo SMO weka implementa clasificadores SVM que transforman las salidas en probabilidades hacer el entrenamiento y clasificación en la red neuronal” [22], este algoritmo está inspirado en el cerebro humano y estaba basado en redes neuronales.

Para poder realizar las pruebas necesarias, y entrenar la red neuronal, es importante contar con información específica, la misma que gracias a la colaboración de instituciones educativas de la ciudad, se obtuvieron datos con ayuda de una especie de encuesta realizada a docentes de dichas instituciones, en dicho proceso se logró recoger información para elaborar el corpus que nos permitió realizar el primer entrenamiento de la red neuronal artificial.

La base de datos que se obtuvo está conformada de datos de niños de entre 6 a 11 años y de los que se recopilaron los siguientes parámetros:

6. Edad cronológica,
7. Sexo,
8. Nivel de vocabulario,
9. Nivel fonológico,
10. Nivel morfológico,
11. Nivel de expresión y comprensión,

Después de haber obtenido los datos necesarios se procedió a clasificar o asignar los cuentos de acuerdo a las variables antes mencionadas.

Para poder procesar los datos del corpus con Weka, se necesita que el archivo corpus tenga como extensión “.ARFF” que es la extensión principal de Weka y que tenga el siguiente formato como podemos observar en la ***Figura 15***.

```
% 7. Attribute Information:
%    1. sexo F o M
%    2. nivel de vocabulario 1 al 5
%    3. nivel fonologico 1 al 5
%    4. morfologia y sintaxis 1 a 5
%    5. nivel de expresion 1 a 5
%    6. nivel de comprension 1 a 5
%    5. cuentos {1 al 22}

@RELATION cuentos

@ATTRIBUTE edad {7,8,9,10,11,12}
@ATTRIBUTE sexo {F,M}
@ATTRIBUTE vocabulario     {1,2,3,4,5}
@ATTRIBUTE fonologia  {1,2,3,4,5}
@ATTRIBUTE sintaxis   {1,2,3,4,5}
@ATTRIBUTE expresion  {1,2,3,4,5}
@ATTRIBUTE comprension     {1,2,3,4,5}
@ATTRIBUTE cuentos    {Caperucita-roja,Tres-cerditos,Cisne-
orgulloso,Bella-princesa,Gallina-de-los-huevos-de-oro,El-leon-
y-el-raton,Barba-azul,El-espiritu-del-agua,Niño-
mago,Gallinita-roja}

@DATA
7,F,4,4,4,4,3,Niño-mago
7,F,5,5,5,5,5,Gallinita-roja
7,F,5,5,5,5,5,Gallinita-roja
7,F,2,2,1,1,1,Gallina-de-los-huevos-de-oro
7,F,5,5,5,5,5,Gallinita-roja
7,F,5,5,4,4,4,Niño-mago
```

Figura 15. Corpus para el entrenamiento de la red neuronal.

Se debe tener cuidado de que no existan espacios al final de cada registro del corpus, o se puede tener problemas con el formato cuando Weka lo desee procesar.

La red neuronal tiene como objetivo, es recomendar un cuento para el niño, de acuerdo a los parámetros del corpus antes mencionado. Luego de esto, el sistema realizará la lectura del cuento.

Para que nuestra red neuronal pueda empezar a clasificar; con ayuda de la herramienta Weka, se procedió a dividir el corpus, del cual del total de registros o casos, el 80% se usó para entrenamiento de la red y 20% para realizar las pruebas respectivas.

Como resultados de este entrenamiento, se obtuvo un porcentaje del 94% de precisión en la clasificación de cuentos, lo que es un muy alentador y que sirve para el propósito de la red neuronal artificial, podemos observar este resultado más a detalle en la siguiente ***Figura 16***:

```
Results
=======

Correctly Classified Instances         129               94.1606 %
Incorrectly Classified Instances         8                5.8394 %
Kappa statistic                          0.9339
Mean absolute error                      0.0161
Root mean squared error                  0.0918
Relative absolute error                  9.0793 %
Root relative squared error             30.8506 %
Total Number of Instances              137

10|0|0|0|0|0|0|0|0|0|
0|19|1|0|0|1|0|0|0|0|
0|0|23|0|0|0|0|1|0|0|
0|0|2|8|0|0|0|0|0|0|
0|0|0|0|6|0|0|0|1|0|
0|0|1|0|0|15|0|0|0|0|
0|0|0|0|0|0|11|0|0|0|
0|0|1|0|0|0|0|15|0|0|
0|0|0|0|0|0|0|0|11|0|
0|0|0|0|0|0|0|0|0|11|
```

Figura 16. Resultados del entrenamiento y matriz de confusión de la red neuronal.

### 5.4.Interfaces Gráficas.

En este aparto se explica el funcionamiento del sistema a través de las ventanas. En primer lugar, se presenta la pantalla de inicio en la que el tutor del niño deberá ingresar unos datos para que el sistema pueda recomendar el cuento para el niño como lo muestra la ***Figura 17***:


Figura 17. Ventana inicial para el ingreso de datos.

La siguiente ventana se nos presenta al presionar recomendar cuento, esta ventana es de tipo informativa de la red neuronal, como se ve en la ***Figura 18***:

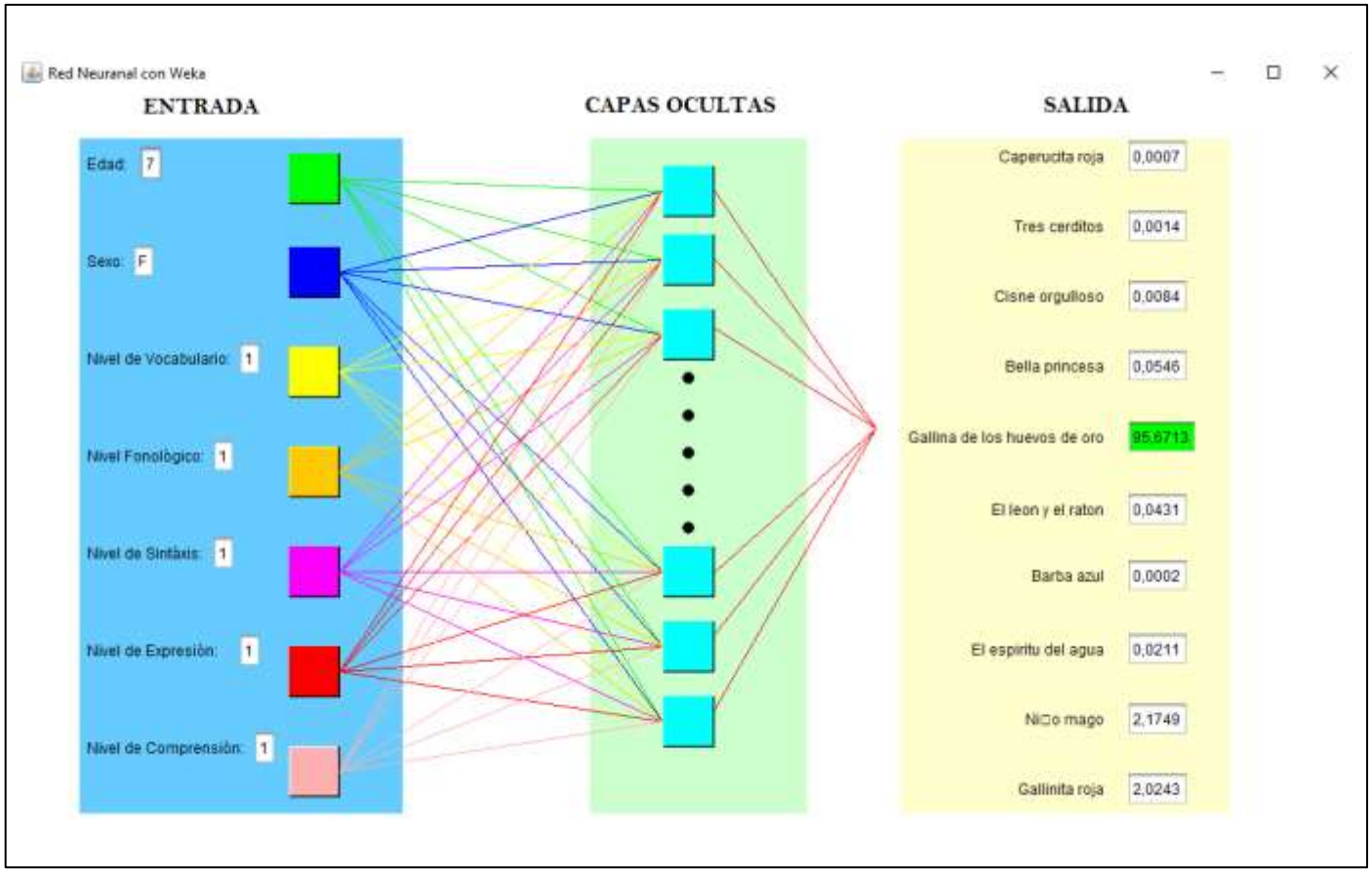


Figura 18. Explicación gráfica de la red neuronal y su resultado.

Al cerrar esta ventana el sistema activará la casilla para leer el cuento recomendado, como muestra la ***Figura 19***:


Figura 19. Ventana de lectura del cuento recomendado.

La ***Figura 20*** se presenta al presionar sobre el botón "contar", en esta ventana el sistema lee el cuento multimedia.

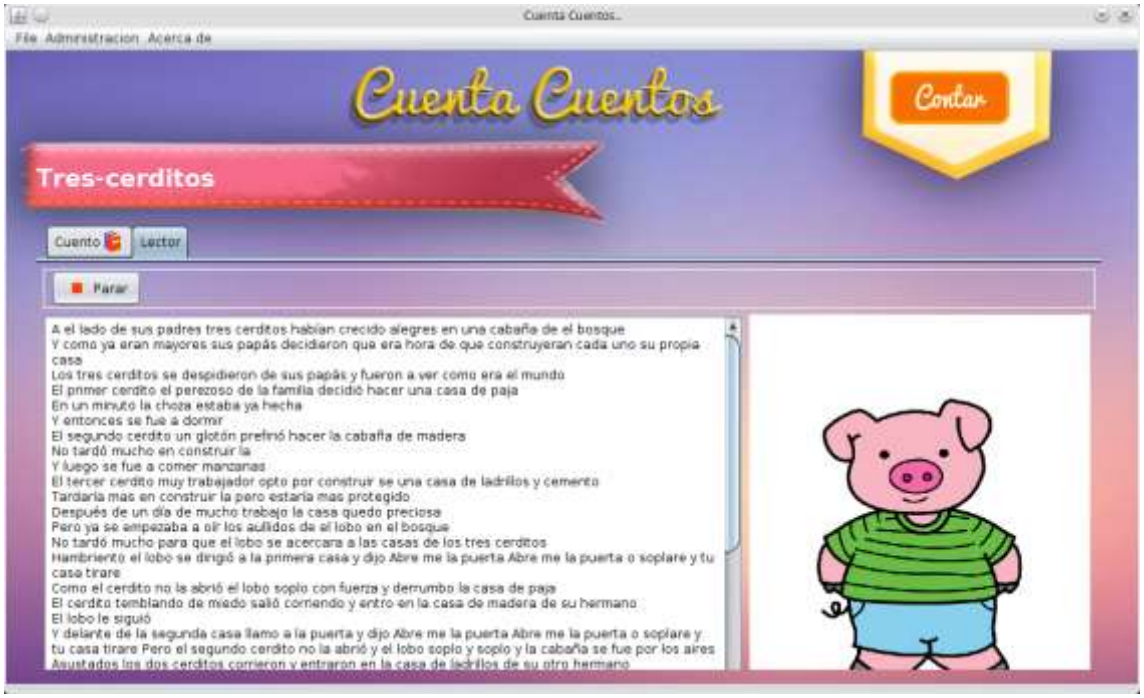


Figura 20. Ventana de lectura del cuento.

## 4. Experimentación y resultados

Con el objetivo de verificar la funcionalidad del prototipo se realiza una evaluación MOS[12], la inteligibilidad y naturalidad del módulo TTS.

### 4.4.Pruebas

Se realizaron pruebas de campo con la ayuda de 15 niños de entre 7 y 12 años de edad los mismos que calificaron con: 5(Excelente), 4(Bueno), 3(Regular), 2(Malo) y 1(Muy malo), la semejanza del sonido con la voz humana (véase ***Figura 22***) y cuan entendible es la lectura (véase ***Figura 21***), después de haber escuchado una fábula de 63 palabras. Las pruebas se realizaron con una base de datos de sílabas y palabras y con base de datos de solo sílabas evaluando de igual manera la la semejanza del sonido con la voz humana (véase ***Figura 22***) y cuan entendible es la lectura (véase ***Figura 23***) con el mismo texto.

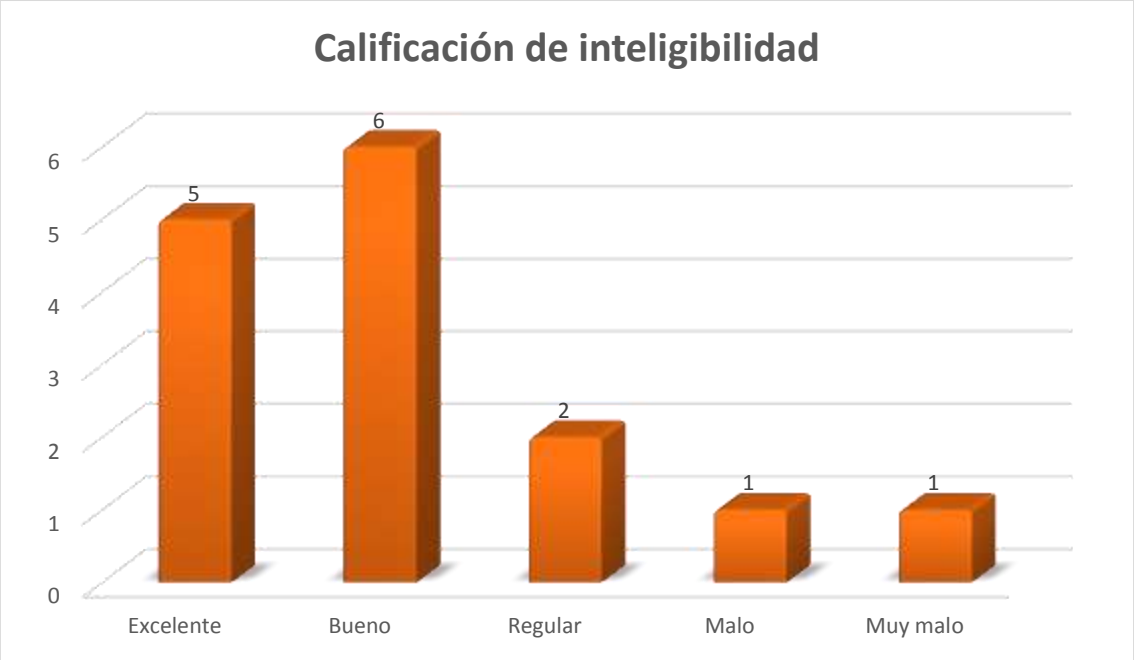


Figura 21. Pruebas de inteligibilidad con base de datos de sílabas y palabras.

[12] MOS (Mean Opinion Score), medida de referencia de calidad para las llamadas de voz.


Figura 22. Pruebas de naturalidad con base de datos de sílabas y palabras.

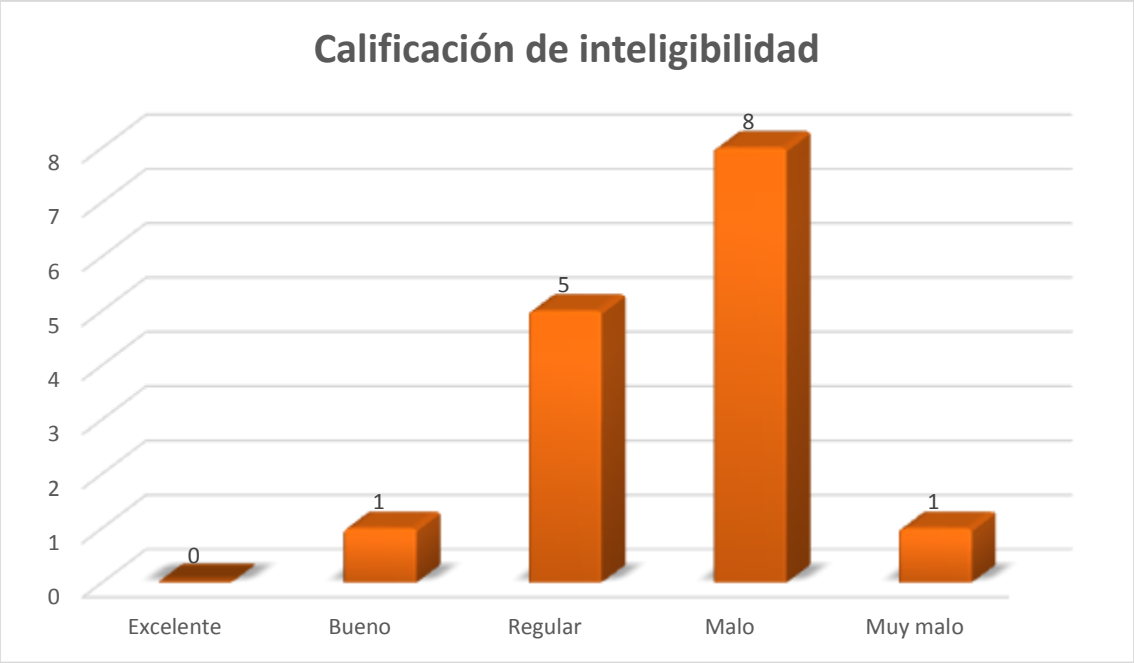


Figura 23. Pruebas de inteligibilidad con base de datos de sílabas.


Figura 24. Pruebas de naturalidad con base de datos de sílabas.

Además, como parte del experimento con la misma muestra ser procedió a analizar la lectura multimedia del cuento como se muestra en la ***Figura 25***, siguiendo la misma técnica que con la experimentación del módulo TTS.

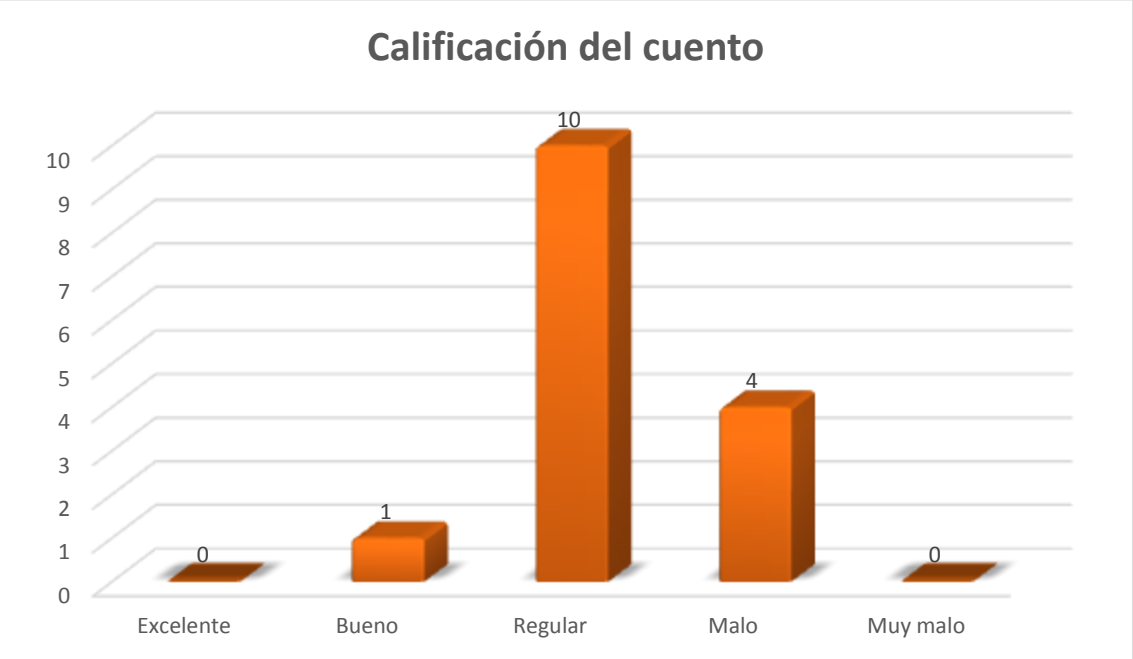


Figura 25. Pruebas de la lectura del cuento.

### 4.5.Resultados

Como se puede apreciar en la ***Figura 26***, el módulo TTS es mucho mejor cuando se realiza una síntesis concatenativa por selección de unidades con base de palabras ya que esto permite mantener mucho más la inteligibilidad del TTS aunque la naturalidad se ve afectada emitiendo un sonido robótico ya el sistema carece de prosodia.

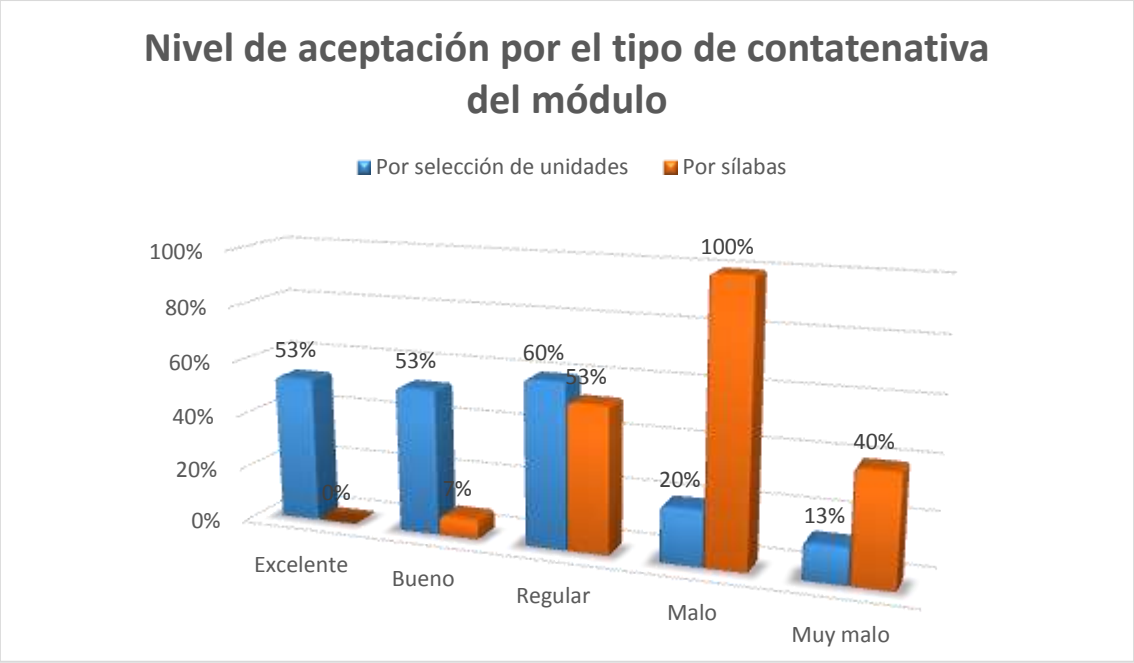


Figura 26. Nivel de aceptación dependiendo de la base de datos utilizada.

En cuanto a la lectura de cuentos multimedia es regular ya que la relación imagen-palabra es limita, es decir, por cada oración solo existe una imagen que describe la misma y esto se les hace aburrido a los niños.

## 5. Conclusiones

- Por medio de comandos se modificó la voz, tratando de retardar el tiempo de lectura, como consecuencia no se obtuvieron resultados precisos, a medida que se fueron realizando pruebas se llegó a la conclusión que este es un proceso demasiado extenso, complicado y que llevaría mucho más tiempo realizarlo, alejándonos así de los objetivos previamente establecidos para el proyecto.
- Con los experimentos realizados se ha determinado que una concatenativa por selección de unidades es un proceso más largo y costoso, pero su nivel de inteligibilidad es mucho más aceptable, aunque su naturalidad se ve afectada ya que carece de emociones, es decir, la lectura es plana. El sistema es capaz de utilizar cualquier corpus de voz siempre y cuando este correctamente etiquetado. Es necesario que el corpus sea grabado para un propósito específico para evitar la concatenación por sílabas ya que esto reduce el nivel de inteligibilidad y naturalidad del sistema.

## 6. Recomendaciones

- El módulo TTS es escalable ya que si se desea agregar más voces solo se necesitaría de la grabación de la voz.
    - Para esto se recomienda que la grabación se realice de manera profesional y con un locutor experimentado.
- Para el sistema cuenta cuentos se pueden agregar más historias ya que escalable, además la base de imágenes se puede agrandar solo añadiéndolas en la carpeta recursos del sistema.
- Se recomiendo ejecutar el sistema desarrollado en la plataforma Linux ya el mismo ha sido desarrollado específicamente para este sistema operativo.

## 7. Trabajo futuro

La siguiente etapa del proyecto se centra en la aplicación de prosodia para expresar emociones, dado que la aplicación es dirigida a niños. Esta propuesta se podría lograr con la manipulación de la señal de audio y el análisis lingüístico de cada oración, como por ejemplo el análisis de los signos de puntuación para determinar si la oración es una pregunta o contiene signos de admiración. Además de la ampliación de la base de datos.

Con el objetivo de mejorar el sistema cuenta cuentos multimedia, la siguiente etapa se centra en mejorar la relación imagen-palabra con el objetivo de contar el cuento por oraciones para que cada una de las palabras esté relacionada con una imagen. Además, se podría mejorar la manera de relacionar la imagen con la palabra para que deje de depender de un archivo XML donde está especificado la palabra y la ruta de la imagen a la que esta relacionada.

## 8. Bibliografía


[1] Ministerio de Educación del Ecuador., «Cursos de TIC'S Y Herramientas para el Aula,» 2015. [En línea]. Available: http://educacion.gob.ec/cursos-de-tics-y-herramientas-para-el-aula-tic-2/

[2] Lee, S. Y., «Storytelling Supported by Technology: An Alternative for EFL Children with Learning Difficulties. Turkish Online Journal of Educational Technology-TOJET,» 2012. [En línea]. Available: http://files.eric.ed.gov/fulltext/EJ989221.pdf

[3] Zhu, Q., «Story Telling In Space and Time An Android Application for Ghost Tour in Edinburgh using Smart Phones,» 2013. [En línea]. Available: https://www.era.lib.ed.ac.uk/handle/1 842/8344

[4] Alofs, T., «The Interactive Storyteller,» 2012. [En línea]. Available: http://essay.utwente.nl/61 987/1 /AlofsThijsFinalThesis.pdf

[5] Camanho, M., Feijó, B., Furtado, A., Pozzer, C., & Ciarlini, A., «A Model for Stream-based Interactive Storytelling as a New Form of Massive Digital Entertainment. In XII Brazilian Symposium on Games and Digital Entertainment,» 2013. [En línea]. Available: http://www.sbgames.org/sbgames2013/proceedings/comp/13-full-paper.pdf

[6] Botturi, L., Bramani, C., & Corbino, S., «Digital storytelling for social and international development: from special education to vulnerable children. International Journal of Arts and Technology,» 2014. [En línea]. Available: http://www.sbgames.org/sbgames201 3/proceedings/comp/1 3-full-paper.pdf

[7] Dillon, G., & Underwood, J., «Computer mediated imaginative storytelling in children with autism. International Journal of Human-Computer Studies,» 2012. [En línea]. Available: https://www.researchgate.net/profile/Gayle_Dillon/publication/220108493_Computer_mediated_imaginative_storytelling_in_children_with_autism/links/0c960527a1b544d337000000.pdf

[8] Duveskog, M., Tedre, M., Sedano, C. I., & Sutinen, E., «Life Planning by Digital Storytelling in a Primary School in Rural Tanzania. Educational technology & society,» 2012. [En línea]. Available: http://www.ifets.info/journals/15_4/20.pdf

[9] Marín Plaza, P., «Síntesis de voz y reconocimiento del habla: implementación en el robot HOAP-3,» 2014. [En línea]. Available: http://repositoriocdpd.net:8080/bitstream/handle/123456789/587/Tes_MarinPlazaP_SintesisVozReconocimiento_2011.pdf?sequence=1

[10] Barrera Maura, M. F., & Fajardo Heras, N. H., «Diseño e implementación de una plataforma genérica para desarrollar y probar nuevas técnicas de detección de plagio en textos,» 2014. [En línea]. Available: http://dspace.ups.edu.ec/handle/123456789/8051

[11] Kertkeidkachorn, N., Chanjaradwichai, S., Punyabukkana, P., & Suchato, A., «CHULA TTS: A Modularized Text-To-Speech Framework,» 2014. [En línea]. Available: http://www.arts.chula.ac.th/~ling/paclic28/program/pdf/414.pdf

[12] FreeTTS, 2016. [En línea]. Available: http://freetts.sourceforge.net/

[13] Walker, W., Kwok, P., Lamere, P., «FreeTTS Open Source Speech Systhesis»

[14] MaryTTS, [En línea]. Available: http://mary.dfki.de/documentation/index.html

[15] Festival, [En línea]. Available: http://www.cstr.ed.ac.uk/projects/festival/

[16] Adobe Systems Software Ireland Ltd., [En línea]. Available: http://www.adobe.com/es/

[17] Real Academia Española., [En línea]. Available: http://www.rae.es/recursos/diccionarios/dpd

[18] Felardo, L. C., «Instalaciones de telefonía y comunicación interior,» 2014.

[19] Jiménez, I. R., De Vega, C. P., «Operaciones auxiliares con tecnología de la informática y la comunicación,» 2010.

[20] García Reyes, T. D., & Martinez Bautista, F. D. J., «Diseño de prácticas de la sonoridad en formatos de audio digital,» 2014. [En línea]. Available: http://tesis.ipn.mx/bitstream/handle/123456789/14882/TESIS_ESIME.pdf?sequence=1

[21] Pérez, C. A. «El Smartphone como Herramienta Pedagógica en la Producción Impreso, Radial y Audiovisual de la coordinación de recursos para el aprendizaje. Caso: Unidad Educativa" Los Andes" San Cristóbal, Estado Táchira,» 2014. [En línea]. Available: http://www.saber.ula.ve/bitstream/123456789/40122/1/articulo1.pdf

[22] Weka. [En línea]. Available: http://www.cs.tufts.edu/~ablumer/weka/doc/weka.classifiers.SMO.html

[23] Romero, A., Etxebarria, A., G, I., Garay, U. «El papel de la prosodia en la enseñanza de la L. Un aporte didáctico para el aula de educación infantil y de educación primaria,» 2015. [En línea]. Available: http://revistes.ub.edu/index.php/phonica/article/download/15359/18552

[24] De la Mota, C., «La enseñanza de la entonacion. docencia semipresencial y tecnologia,» [En línea]. Available: http://prosodia.upf.edu/membres/carmedelamota/arxiu/delaMota_entonacion_CIDUI04.pdf.

[25] Llacura, Morera, J., «Prosodia: modificación de la conducta a partir de las bases emocionales orales de la comunicación,» 2009. [En línea]. Available: http://comisionnacional.insht.es/InshtWeb/Contenidos/Documentacion/FichasTecnicas/NTP/Ficheros/821a921/845%20web.pdf

[26] Real Academia Española, 2016. [En línea]. Available: Obtenido de http://www.rae.es/