CAPÍTULO

4

Desarrollo de un corpus de voz de niños

Un corpus de voz es un conjunto de grabaciones con formato de audio, así como un conjunto de comentarios y documentos referentes a los datos de voz. Dichos datos deben estar almacenados con un formato estándar para ser utilizados por el equipo de cómputo. Las colecciones o corpora de voz son utilizadas para estudiar la variabilidad de la señal, características de la producción de voz, métodos para mejorar el reconocimiento automático (entrenamiento de sistemas estadísticos o de redes neuronales), estudios para desarrollar sintetizadores de voz que suenen de manera mas natural, etc.

4.1 Corpora de Voz

Para hacer investigación en el área de la lingüística, fonética y reconocimiento de voz en Español es necesario hacer un análisis de datos de voz y experimentar con esos datos; para eso es necesario contar con un corpus de voz y un conjunto de herramientas que apoyen el trabajo de investigación. Como ya mencioné anteriormente, un corpora de voz es una base de datos de voces con sus respectivas anotaciones, dicha recolección las podemos dividir en tres categorías principales:

De tipo análisis / diagnóstico: para el estudio del habla;
De propósito general: incluyen vocabularios no - específicos, útiles para diversas aplicaciones;
De propósito específica: son colecciones de habla con características especiales para algún tipo de aplicación (por ejemplo, colecciones de grabaciones de dígitos o de nombres de ciudades)
El habla grabada en estos corpora de voz puede ser espontánea (natural) o leída ("de laboratorio").
También las grabaciones pueden estar compuestas por:
Fonemas aislados: utilizados para comparar las pronunciaciones y modos de articulación e diferentes fonemas sin ser afectados por otros fonemas.
Palabras aisladas: puede incluir palabras que no tienen sentido, sólo por construir combinaciones de fonemas, para estudiar o captar el fenómeno de coarticulación.
Frases completas: producen ejemplares de voz mas naturales que las palabras aisladas debido a que contienen fragmentos de habla continua.
Fragmentos de textos: son generalmente grupos de frases relacionadas semánticamente.
Habla semi-espontánea: se controla estrictamente el vocabulario, la sintaxis y la pronunciación. Se utilizan para aplicaciones comerciales, como por ejemplo el anuncio de las salidas de los autobuses en una terminal.
Habla espontánea sobre un tema predeterminado: el habla espontánea no esta restringida, sin embargo, en este caso se controla el tema del que se habla.

Habla espontánea: el locutor escoge el tema sobre el cual va a hablar, así como el vocabulario, conversando de manera natural.

4.1.1 Algunos aspectos prácticos para la construcción de corpora

Es importante resaltar algunos puntos importantes para la construcción de una base de datos de voces, puesto que engloba aspectos más allá de los relacionados con la parte técnica.

- ?? Visibilidad
- ?? Datos
- ?? Ambiente
- ?? Control de la interacción
- ?? Monitoreo

Visibilidad: se refiere a la decisión de dar a conocer al locutor que su voz esta siendo grabada.
En algunos casos es mejor mantenerlo en secreto para evitar que el locutor se ponga nervioso y hable de manera no-natural, manifieste estrés, tartamudeo, muletillas, etc.

Datos: se puede grabar con un micrófono o varios, inclusive con un laringógrafo, el cual capta la frecuencia fundamental, dependiendo de la información que se requiere.

Ambiente: un ambiente de grabación puede ser desde una cabina de grabaciones o un lugar público. La decisión de cual ambiente ocupar para grabar afecta por un lado la calidad del sonido (ruido, eco, etc.) . Un lugar especial puede garantizar una excelente calidad del sonido, sin embargo para muchas aplicaciones se necesita un ambiente más natural, de hecho, se busca que sea lo más cercano al ambiente en el cual funcionará más tarde el sistema (si se trata de reconocimiento). Por otro lado el ambiente puede, al igual que en el punto de visibilidad, afectar el estado del locutor, hacerlo sentir cómodo o situarlo en un lugar que lo pone nervioso.

Control de la interacción: Siempre se debe llevar un control de las sesiones de grabación, anotando fecha, duración, tipo de habla grabada, etc.

Monitorea: se pueden modificar las características fonéticas (pronunciación, velocidad del habla, estilo) y técnicas (nivel de volumen, canal de grabación, etc.) durante las sesiones de grabación. Esto asegura que las grabaciones contengan exactamente lo que se programó, evitando errores, estornudos, risas, etc., pero no representan situaciones reales, en las que estos eventos son naturales.

Algunos proyectos que son fuentes de Corpora de voz son los siguientes: ESPRIT en Europa, SQUALE, SUNDIAL, SUNSTAR, y actividades coordinadas por DARPA en EUA, GRECO en Francia, u organizaciones de estándares como NIST en EUA, NPL en Reino Unido y AFNOR en Francia. A través del proyecto NSF-ITR Reading Tutor Project (http://cslr.colorado.edu/beginweb/research.html) que tiene la UDLA-PUE con el CSLR de la Universidad de Colorado se esta promoviendo la creación de corpora de voces de niños en Español Mexicano.

Algunas de las bases de datos conocidas: TIMIT y ATIS

La variabilidad inherente a la señal de voz puede provenir de diferentes fuentes:

- ?? Coarticulación
- ?? Dialectos
- ?? Contexto social
- ?? Características individuales
- ?? Ambiente (ruido, condiciones de grabaciones, micrófonos, calidad del medio de grabación, etc.)

El resultado de esta variabilidad afecta todos los niveles de la comunicación hablada, desde los sonidos hasta la gramática y el vocabulario.

En esencia, cada producción de voz es única y esto hace que el proceso de decodificación sea (natural o automático) bastante difícil.

La variabilidad se ha clasificado de diferentes formas, una de ellas considera las diferencias intra-locutor y las que son inter-locutores.

Variabilidad Intra-locutor: Son las diferencias que se dan en el habla de una misma persona a causa de estados de ánimo, de salud, (tono, volumen, velocidad y las condiciones en las que habla

Variabilidad Inter-locutor: Se refiere a las diferencias entre las diferentes personas. Estas variaciones son causadas por las diferencias en la anatomía de las cuerdas vocales, pulmones, cavidades bucal y nasal, etc. También afecta el medio ambiente (acústica, eco, dispositivos de grabación, ruido, etc.)

Los objetivos generales que se plantearon en este apartado fueron**:**

?? Diseñar el protocolo de grabación y la transcripción del corpus.
?? Recolectar la voz de los niños en diferentes escuelas.
?? Transcribir la voz de los niños en el menor tiempo posible.

El desarrollo de un corpus se divide en las siguientes etapas:

?? *Diseño.* En esta etapa se plantea el contenido del corpus.

?? *Grabación.* Proceso de recopilación de datos de voz.

?? *Trascripción.* Etapa en la que se etiquetan o transcriben los datos de voz, de manera que estén disponibles para su utilización

4.2 Diseño del corpus

Un corpus de voz, como ya lo mencioné anteriormente es una base de datos en computadora que contiene una colección de grabaciones de habla y sus respectivas anotaciones y transcripciones. Esta documentación de los datos es muy importante, ya que describe las técnicas utilizadas en el diseño y grabación de los datos, el número y tipo de locutores, contenido lingüístico, etc. Además de una transcripción textual de cada grabación, en muchas ocasiones se incluyen etiquetas a nivel palabra y a nivel fonema. Estas etiquetas son archivos de texto que mencionan las posiciones o fronteras en milisegundos, en donde empieza y termina cada palabra y cada fonema hablado.

El contenido del corpus se diseñó tomando en cuenta al grupo al que va dirigido este trabajo de tesis, el cual incluye niños de nivel primaria de segundo a quinto grado. También se abarcaron aspectos como el balanceo en el número de cada género (masculino y femenino) y la inclusión de los diferentes estratos sociales.

4.2.1 La aplicación del corpus

Un corpus de voz se diseña para un propósito específico. En este trabajo se realizó un tutor animado interactivo para la enseñanza del lenguaje dirigido a niños de nivel primaria.
La aplicación consiste en enseñar al alumno un tema acorde a su nivel educativo y evaluarlo de manera automática, las respuestas del usuario son de forma hablada y reconocidas por el sistema para poder evaluarlas. En el capítulo VI de este documento se muestra una figura del sistema desarrollado.

4.2.2 Contenido lingüístico del corpus

Para determinar el contenido del corpus es necesario saber para que va a ser utilizado, y así poder seleccionar los datos necesarios y adecuados que integrarán dicho corpus.

Para diseñar el corpus fue necesario entrevistar a profesores de los diferentes grados de nivel primaria, revisar los libros de texto de cada grado, así como de los planes de estudio que ofrece la Secretaría de Educación Pública y otras dependencias no gubernamentales.

El objetivo que se persigue con esto, es detectar algunas de las frases o el tipo de vocabulario ya sea de tipo matemático, comandos para computadora, o enunciados que contienen cierto grado de dificultad en la pronunciación, tales como refranes o adivinanzas que emplea un niño dependiendo del grado escolar que cursa y con ello determinar el contenido lingüístico del corpus.

El contenido lingüístico del corpus esta basado en oraciones formadas con frases o palabras que caen dentro de alguna de las siguientes categorías:

Tipo	Categoría	Comentarios	Número Estimado
Habla Directa	Sonidos	Fonemas y/o otros sonidos que tiene diferentes duraciones u otras características. Por ejemplo, /mmmmm/	39 fonos
Habla Directa	Letras y números	Letras y números comunes	400 palabras
Habla Directa	Palabras Aisladas	Balanceadas fonéticamente. Todas las palabras pueden ser pronunciadas por niños de primaria. Se incluyen los colores.	600 palabras en total
Habla Directa	Comandos	Comandos comunes	600 palabras y frases

		en interacción hombre-computadora; cada niño dice exclusivamente aquellas palabras de su grado	
Habla Directa	Palabras Matemáticas	Palabras y fracciones comunes en matemáticas. Específicas del año escolar	24 palabras
Habla Directa	Oraciones	Oraciones de historias de niños.	100 frases por conjunto, 300 en total
Lectura	Palabras	Palabras de Historias. Las mismas palabras que en frases para lectura. Estas son específicas del año escolar.	200 palabras por conjunto, 600 en total.
Lectura	Rimas Comunes	Frases con rima comunes	300 frases en total
Lectura	Frases para Lectura	Oraciones tomadas de historias de niños	100 frases por conjunto, 300 oraciones en total
Habla espontánea	Nombres	Incluye el nombre de sus amigos y la ciudad de origen.	4 palabras por niño
Habla espontánea	Monólogo	Monólogo de uno o varios temas, incluyendo sus	1 historia por niño

		vacaciones favoritas o lo que hicieron el fin de semana	

Tabla 4.1: El contenido lingüístico del corpus

4.2.3 Número y tipo de locutores

El número de locutores es una de las características más importantes de un corpus de lenguaje hablado. De acuerdo al número de locutores un corpus se puede dividir en tres clases [Gibbon et al., 1997]:

- *Corpus grabado con un grupo pequeño de locutores.* Se considera un grupo pequeño de locutores al que está formado por un conjunto menor a 5 personas. Un corpus de voz grabado con un grupo pequeño de locutores se utiliza frecuentemente para desarrollar sistemas de reconocimiento de voz monolocutor y para síntesis de voz.

- *Corpus grabado con un grupo mediano de locutores.* Se considera un grupo mediano de locutores al que está formado por un conjunto de 5 a 50 personas. Generalmente se usa un grupo mediano de locutores para tener un conjunto numeroso de repeticiones del fenómeno lingüístico que se desea investigar y poder realizar procesamientos y modelados de tipo estadístico. En este punto es importante determinar, de acuerdo al número de locutores, la cantidad mínima de pronunciaciones que se requiere obtener por cada hablante.

- *Corpus grabado con un grupo grande de locutores.* Se considera un grupo grande de locutores al que está formado por un conjunto de más de 50 personas. Para entrenar y probar adecuadamente un sistema de reconocimiento de voz independiente del locutor es necesario un corpus grabado con un grupo grande de locutores.

La selección del tipo de locutores (hombres, mujeres, niños) depende de la aplicación. Para entrenar y probar sistemas de reconocimiento de voz se debe seleccionar una muestra representativa de diferentes locutores. Si los usuarios del reconocedor serán mujeres entonces se debe considerar un conjunto de mujeres para el diseño y grabación de los datos de voz. En general, se deben incluir todos los posibles tipos de locutores en un corpus de voz, a menos que haya argumentos imperativos para excluir a grupos específicos. Algunas de las características que deben ser consideradas en la selección del grupo de locutores están asociadas a factores como sexo, edad y lugar de origen.

4.2.4 Grabación de los datos de voz

Además de los datos de voz se debe tener información adicional acerca de las características de los datos, el material, el equipo y las condiciones de la grabación. Es conveniente tener la información lo más detallada que sea posible. A continuación se presentan algunos puntos importantes en la documentación de un corpus de voz [Gibbon et al.,1997]:

- ***El tipo de señal.*** Se debe especificar si es una señal acústica de voz, una señal del laringógrafo, rayos X, etc.

- ***Análisis de los datos.*** Se debe indicar si se cuenta con datos de voz filtrados o procesados con métodos de Fourier o espectrales y si se han extraído algunas características como los formantes de la señal.

- ***Descripción de los datos de voz.*** Se deben describir las características de los locutores como sexo y edad, el tipo de datos de voz (palabras aisladas, oraciones, discursos, etc), el lugar de grabación (en un laboratorio, en una cabina de radio, etc.), si es vía telefónica o por micrófono (en el caso de usar micrófono se debe especificar el tipo de micrófono) y el equipo de grabación.

?? ***Comentarios.*** Se refiere a comentarios útiles sobre los datos de voz tales como el criterio utilizado para segmentar los datos de voz, el protocolo utilizado en la grabación, si hay errores de pronunciación, si existe un límite máximo en la duración de los datos de voz, etc.

?? ***Etiquetas.*** Se refiere a las transcripciones ortográficas, fonémicas, fonéticas, etc, así como a la descripción de los símbolos usados en las transcripciones o el uso de algún alfabeto fonético.

La porción de audio del corpus se grabó a la frecuencia de muestreo de 16 kHz almacenándose 16 bits por muestra y usando una tarjeta de sonido sound blaster. Las grabaciones se realizaron en oficinas y laboratorios de las diferentes escuelas utilizando un micrófono de diadema (noise canceling), dos computadoras Pentium III a 650 MHz y bocinas. El tiempo estimado de grabación por niño fue de 20 minutos.

La grabación de los datos de voz consistió de tres tipos de voz: respuestas a preguntas directas, leídas y espontáneas. Una cara animada se usó para hacer las preguntas directas. El texto que tuvieron que leer los niños se desplegó en la pantalla con letras grandes. Cada grabación se escribió en un archivo de audio con formato RIFF. Todos los archivos tienen un nombre para identificar el grado, género, sub-protocolo, niño y frase. Cada campo del nombre está separado por un guión. El formato que se utilizó es el siguiente:

CN-GG-S-NNNN-PP-UUUU-A.wav

Ejemplo:

CN-02-F0001-01-0001-A.wav

Donde:

CN indica las iniciales del corpus, en este caso el corpus de niños.

GG indica el grado en el que va el niño

S indica el género del niño

NNNN indica el número único del participante

PP indica el subprotocolo que está siendo grabado

UUUU indica el número de pronunciación

A indica el número de intento por grabar esa pronunciación

El procedimiento de grabación consistió en dar una pequeña explicación al niño (instrucciones) y pedirle que escoja desde el principio el tema para monólogo. Antes de comenzar la grabación, se realizó una calibración del micrófono.

Este tipo de grabación se utilizó para garantizar que el material de voz tuviera todas las unidades fonéticas del Español.

Un ejemplo de la aplicación para la recolección de los datos se muestra en la siguiente grafica, la cual es sólo una parte del sistema de recolección de voces de niños. Dicho sistema fue desarrollado en el Rapid Application Developer (RAD).

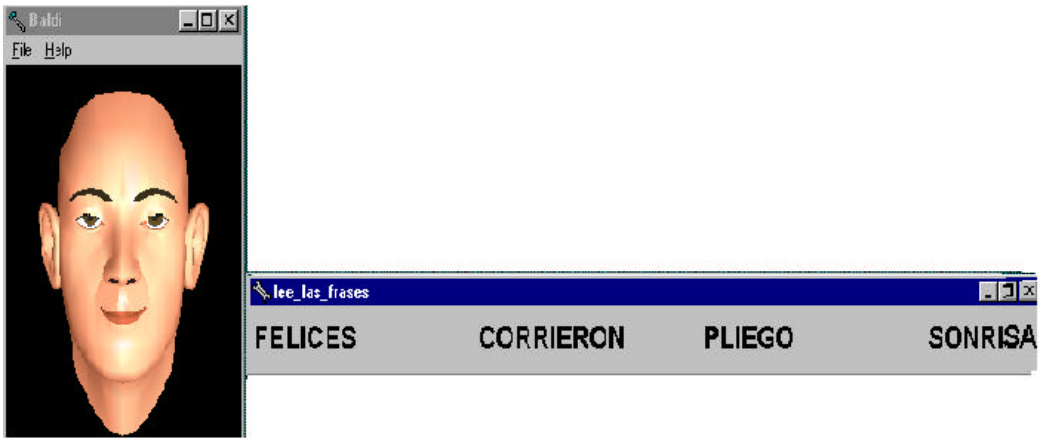

Figura 4.1: Aplicación para la Recolección de Datos

4.3 Transcripción del corpus de voz

En este apartado daré una introducción a algunos conceptos lingüísticos útiles para entender las siguientes secciones. Se describe como se representa lingüísticamente un corpus. Finalmente, describiré los niveles y tipos de trascripción de datos de voz.

4.3.1 Fonética y Fonología

La fonología y la fonética son disciplinas lingüísticas que estudian los sonidos, pero con diferentes fines. La fonología estudia la función de los sonidos en una lengua. La fonética estudia los sonidos desde el punto de vista físico y fisiológico [Real Academia, 1974] .

La fonología clasifica a los sonidos en un sistema, basándose en sus características articulatorias y en la distribución de estos sonidos en la cadena sonora del habla. Los fonemas son las unidades de sonido establecidas por la fonología y se representan gráficamente con símbolos lingüísticos entre barras (/ /). Por ejemplo, el fonema /s/ representa el sonido producido al pronunciar las letras s y z en las palabras casa y caza, respectivamente.

La fonética distingue los sonidos por sus particularidades contextuales y sus propiedades sonoras en el habla. Las unidades de sonido establecidas por la fonética son las unidades fonéticas. A cada uno de los sonidos producidos al articular un fonema de distintas formas le corresponde una unidad fonética. Las unidades fonéticas se representan por símbolos fonéticos. Por ejemplo, las unidades fonéticas correspondientes a la pronunciación del fonema /n/ en las palabras mango y mano se representan con los símbolos fonéticos N y n, respectivamente.

4.3.2 Sistema fonológico del lenguaje Español

Cada lengua presenta en su código, un número limitado y exacto de fonemas que constituyen un sistema fonológico. El sistema fonológico del Español consta de 23 fonemas:

- ***18 consonantes.*** (/p/ , /b/, /t /, /d/ , /k/ , /g/ , /ch/ , /f/ , /s/ , /j/ , /y/ , /m/ , /n/ , /ñ/ , /l/ , /ll/ , / r/ , /rr/) y

- ***5 vocales.*** (/a/ , /e/ , /i/ , /o/ , /u/).

4.3.3 Correspondencia entre fonemas y letras.

Las letras son los signos gráficos que utilizamos para representar a los fonemas en la escritura y cuyo conjunto constituye el alfabeto de una lengua.
En Español existen 29 letras para representar a los 23 fonemas de que consta el lenguaje.

4.4 Órganos Articulatorios

Reciben el nombre de órganos articulatorios los órganos que intervienen en la fonación, proceso mediante el cual se produce la emisión de los sonidos del lenguaje. Mediante la fonación, el aire procedente de los pulmones hace vibrar las cuerdas vocales, que son dos músculos paralelos y elásticos, situados en la laringe, que pueden tenderse o aflojarse, aproximadamente entre sí o separarse más o menos, dejando entre ellos una abertura triangular llamada glotis. Con ello se emiten normalmente los sonidos. Si la abertura es muy ancha, el aire pasa libremente, sin que las cuerdas vocales vibren; si es estrecha, el paso del aire por ella determina las vibraciones sonoras que producen el sonido que llamamos voz [Onieva, 1995].

Los órganos articulatorios se pueden clasificar en tres grupos bien delimitados: cavidades infraglóticas, cavidad laringea u órgano fonador y cavidades supraglóticas.

- ***Las cavidades infraglóticas.*** Están formadas por los órganos propios de la respiración: pulmones, bronquios y tráquea.

- ***La cavidad laríngea.*** Está situada inmediatamente por encima de la tráquea y está constituida por una serie de cartílagos que envuelven las cuerdas vocales.

- ***Las cavidades supraglóticas.*** Son la cavidad bucal y las fosas nasales.

Cuando el sonido es oral, la única gran cavidad que encontramos es la bucal. La parte superior de la cavidad bucal está constituida por el paladar, la parte inferior de la boca está

ocupada por la lengua. Cerrando la cavidad bucal por su parte anterior se encuentran los dientes o incisivos superiores y los incisivos inferiores. Entre los incisivos superiores y el comienzo del paladar existe una zona de transición que recibe el nombre de alvéolos.

4.4.1 Clasificación de los fonemas

La clasificación de los fonemas del Español está en función de la forma y el lugar de articulación [Onieva, 1995]. Se denomina forma de articulación a la manera en que se realiza el contacto o aproximación de los órganos articulatorios. Atendiendo a la forma de articulación, las vocales son:

- ***Abiertas.*** (/a/)
- ***Medias.*** (/e/ , /o/)
- ***Cerradas.*** (/i/ , /u/)

Las consonantes según la forma de articulación, se clasifican en:

- ***Oclusivas.*** Caracterizadas por una interrupción en el paso del aire motivada por el cierre completo entre dos órganos articulatorios. Deshecha súbitamente la interrupción u oclusión, el aire espirado se precipita hacia fuera con una breve explosión. En Español, los fonemas oclusivos son:

    1. Orales: /p/ , /t /, /k/ , /b/, /d/ , /g/
    2. Nasales: /m/ , /n/ , /ñ/.

- ***Fricativas.*** Caracterizadas por un estrechamiento del canal bucal que no llega hasta el cierre completo, determinación de un roce o fricación del aire, con ruido más o menos fuerte. Los fonemas fricativos son: /f/ , /s/ , /y/ , /j/.

- ***Africadas.*** Caracterizadas por un cierre momentáneo de la salida del aire (momento oclusivo) simultáneo a un estrechamiento del canal bucal (momento fricativo),

realizado por los mismos órganos articulatorios. En Español, el único fonema africado es /ch/.

- ***Líquidas.*** Las consonantes líquidas sin dejar de ser consonantes poseen algunos rasgos propios de las vocales. Dentro de este grupo se distinguen dos clases:

    1. Laterales. En cuya emisión el aire sale por un lado o por ambos de la cavidad bucal (/l/ , /ll/);
    2. Vibrantes. Que se producen por una o varias vibraciones del ápice o punta de la lengua (/r/ , /rr/).

Se denomina lugar de articulación el lugar de la cavidad bucal donde un órgano articulatorio activo se aproxima a otro, activo o pasivo. Atendiendo a lugar de articulación, las consonantes pueden ser:

- ***Bilabiales.*** El labio inferior se aplica contra el superior: /p/ , /b/ , /m/.

- ***Labiodental.*** El labio inferior se apoya contra los incisivos: /f/.

- ***Dentales.*** La punta de la lengua se aplica sobre la cara interna de los incisivos superiores: /s/ , /n/ , /l/ , /r/ , /rr/.

- ***Palatales***. El predorso de la lengua se sitúa al nivel del paladar duro: /y/ , /ch/ , /ñ/ , /ll/ .

- ***Velares.*** El postdorso de la lengua se aplica al velo del paladar: /k/ , /g/ , /j/.

Según vibren o no las cuerdas vocales, los fonemas se clasifican, respectivamente, en sonoros y sordos.

?? ***Sonoros.*** Son todas las vocales y las consonantes /b/, /m/ , /n/ , /l/ , /d/ , / r/ , /rr/ , /y/ , /ll/ , /ñ/ , /g/ .

?? ***Sordos.*** Son las consonantes /p/ , /t/ , /k/ , /ch/ , /s/ , /j/ , /f/.

El velo del paladar puede abrir o cerrar el paso del aire a través de la cavidad nasal. Según la posición del velo del paladar, los fonemas pueden ser orales o nasales.

?? ***Nasales.*** Se originan cuando el velo del paladar está separado de la pared faríngea, saliendo el aire solamente a través de la cavidad nasal. Los fonemas nasales en Español son: /m/ , /n/ , /ñ/. Los demás fonemas del Español son orales.

?? ***Orales.*** Se producen cuando el velo del paladar se encuentra en contacto con la pared faríngea y el aire sale completamente a través de la cavidad bucal.

4.5 Ejemplo de una herramienta para generar las transcripciones y las etiquetas (CSLU Toolkit)

El trabajo de etiquetado es tedioso y lento. Además de ser poco placentero, este tipo de tarea sufre de errores ya que muchas veces la decisión de fronteras es un tanto subjetiva, y una persona cansada cae más fácilmente en errores.
Para generar las etiquetas a nivel palabra y fonema existen ya algunas herramientas automáticas, sin embargo estas requieren de un sistema de reconocimiento de voz que ya funcione. Por ello, al principio es necesario hacerlo manualmente.

El objetivo de este apartado fue etiquetar el corpus ***children,*** que consta de 450 locutores aproximadamente. Etiquetar una frase significa poner etiquetas donde ocurre cada fonema y cada palabra, información que es necesaria para entrenar un reconocedor. Esta tarea se divide en 5 partes:

1. Crear las etiquetas a nivel de texto
    a. Crear las transcripciones del habla espontánea (preguntas y monólogo)
    b. Corregir errores en las transcripciones de las frases
2. Crear automáticamente las etiquetas a nivel de palabra.
3. Ajustar manualmente las etiquetas a nivel de palabra.
4. Crear automáticamente las etiquetas a nivel de fonema.
5. Ajustar manualmente las etiquetas a nivel de fonema.

La siguiente figura ilustra la transcripción a nivel palabra de la palabra “cinco”

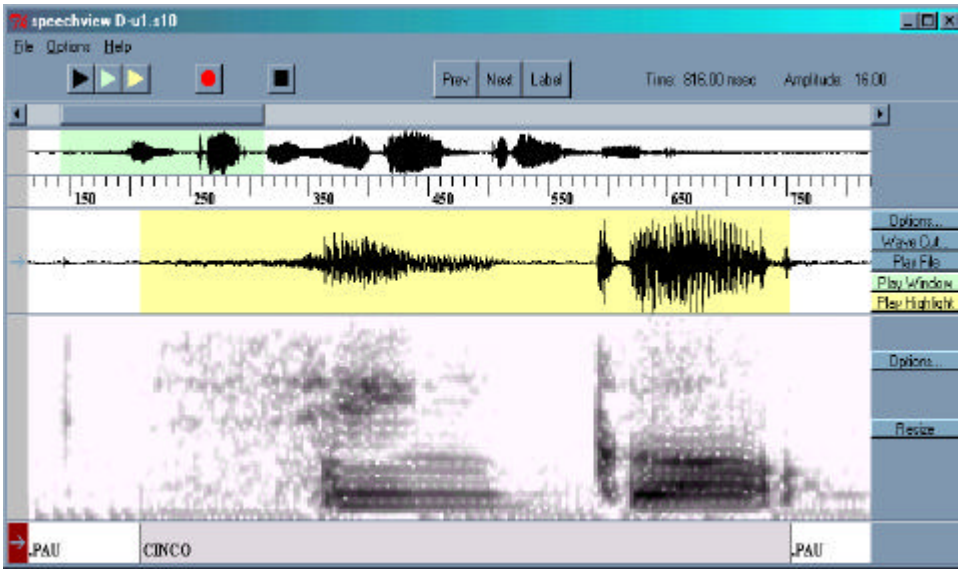

Figura 4.2: Transcripción a nivel de palabra

La siguiente figura ilustra la transcripción a nivel fonema de la palabra “Efrén”.

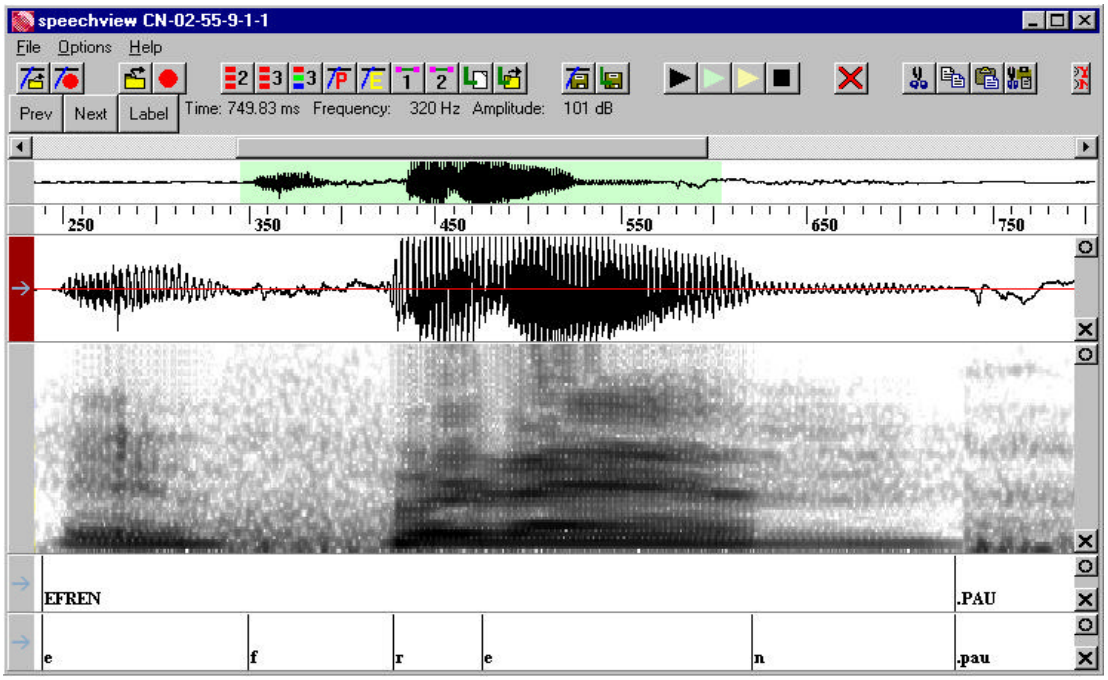

Figura 4.3: Transcripción a nivel fonema

Crear las etiquetas a nivel de texto

La primera etapa es crear las transcripciones del habla espontánea, deben escuchar y transcribir la respuesta a las preguntas:

¿Cómo te llamas?

¿Cuántos años tienes?

¿En que año vas?

¿Qué color te gusta más?

¿Cómo se llama tu mejor amigo?

¿En que ciudad naciste?

Danos el nombre de algún compañero o compañera que tengas.

Monólogo

Después hay que corregir las transcripciones hasta que corresponda con lo que dijo el niño. Se debe escribir lo que el niño intentó decir y poner una etiqueta especial si está mal

pronunciado. Es necesario porque los niños a veces no leyeron o no dijeron lo que deberían haber leído o dicho.

Por ejemplo si el niño estaba leyendo letras del abecedario y al nombrar la "F" dice */fe/*, no se debe cambiar la letra de la transcripción, pero se debe poner la observación de que hay una pronunciación alternativa de la F. Si el niño intenta decir distinto pero sólo se entiende algo como */istinto/* se debe dejar la palabra que intento decir (distinto) y observar que se tiene una pronunciación alternativa.

Verificar las transcripciones no alineadas

La verificación se puede hacer usando los siguientes programas que residen en el directorio **/src/tcl/.** Desde una ventana de tcsh, se debe correr lo siguiente (substituyendo el número 19 con el número del locutor):

Crear el archivo de referencia .files para cada locutor (fólder)

>tcl **mk_file_children.tcl** –user **19** –corpus **children**–escuela **Alejandria**

Generar automáticamente las trascripciones a nivel de texto (archivos .txt) para cada locutor de acuerdo al protocolo que grabó.

>tcl **crear_txt_children.tcl** –user **19** –protocol **2.txt** –escuela **Alejandria**

Revisar las transcripciones generadas automáticamente y crear las de las preguntas.

>tk **check_txt_files.tk** –files **/data/corpora/children/Alejandria/files/19.files**

Sustituir acentos, signos de puntuación en las transcripciones

tcl **clean_txt.tcl** –files **/data/corpora/children/Alejandria/files/19.files**

Crear las etiquetas a nivel de palabra alineadas en tiempo

El siguiente script crea los archivos .wrd a partir de los .txt que ya revisamos. Para crearlos, usa un algoritmos que predice la duración de cada palabra, basándose en estadísticas sobre la duración promedio de cada fonema. El algoritmo sólo toma en cuenta la señal para

determinar donde empieza y donde termina el habla, pero no analiza la duración de las palabras.

> tcl **txt2wrd.tcl** –files **/data/corpora/children/Alejandria/files/19.files** –vocab **exceptions.vocab** –rules **tts_rules.txt** –dur **phones.dur**

Ajustar manualmente las trascripciones a nivel de palabra, creadas en el paso anterior.

Para ajustar las etiquetas, utilizamos la herramienta speechviewer, lo cual es parte del CSLU Toolkit y fue modificada por Tlatoa.

> cd /CSLU/Toolkit/2.0/script/sview_1.0/
> tk **speechview.tcl** –corpus **/data/corpora/children/Alejandria/files/19.files**

Ajustar las fronteras de cada palabra de acuerdo con la información en la señal.

Quitar los silencios largos de las grabaciones.

Muchas veces, las grabaciones tiene silencios largos al principio o al final de la grabación. Usando la información que ahora tenemos, de donde están las palabras, podemos quitar estos silencios y así reducir el tamaño de algunos archivos y el tiempo que se requiere para entrenar.

> cd C:/src/tcl/
> tcl **wrd_trim_wav.tcl** –files **/data/corpora/children/Alejandria/files/19.files**

Antes de borrar los archivos .wav modificados marcados como .old, debemos verificar que las nuevos .wav, estén bien.

Generar automáticamente las transcripciones a nivel de fonema.

A partir de los .wrd ya revisados, hacemos lo mismo que hicimos a nivel de palabra y creamos los .phn.

> tcl **wrd2phn.tcl** –files **/data/corpora/children/Alejandria/files/19.files** –vocab **exceptions.vocab** –rules **tts_rules.txt** –dur **phones.dur**

Ajustar manualmente las transcripciones alineadas a nivel de fonema.
Para ajustar las etiquetas se usa el mismo proceso que usamos a nivel de palabra. Puede haber casos en que la identidad del fonema tiene que cambiar, o se deben insertar o borrar fonemas de la transcripción.

> cd /CSLU/Toolkit/2.0/script/sview_1.0/
> tk **speechview.tcl** –corpus **/data/corpora/children/Alejandria/files/19.files**

Revisar las transcripciones en caso de problemas.
A veces, por accidente, se dejan espacios entre etiquetas o al principio o al final de la frase. Este tipo de problemas se puede descubrir automáticamente usando el siguiente script.

> cd /src/tcl
> tcl **detect_problems.tcl** –files **/data/corpora/children/Alejandria/files/19.files**

Realinear las trascripciones a nivel de palabra.
Puede ser que se ajustaron fronteras entre fonemas sin ajustar la frontera correspondiente entre palabras. El siguiente programa hace este tipo de ajuste automáticamente.

> tcl **adjust_wrd_boundaries.tcl** –files **/data/corpora/children/Alejandria/files/19.files**

Crear nuevamente las trascripciones no alineadas (archivos .txt) a partir de las alineadas.
De la misma manera, si cambiaste alguna palabra usando speechviewer, hay que cambiar la palabra en los archivos no alineados también.

>tcl **wrd2txt.tcl** –files **/data/corpora/children/Alejandria/files/19.files**

4.6 Consideraciones especiales de etiquetado a nivel de texto

Estas primeras consideraciones, se hacen automáticamente con el script que crea las transcripciones.

- Los acentos son marcados con una comilla sencilla después de la vocal acentuada. Por ejemplo en la palabra *esta'*
- La tilde se coloca después de la **n** como en la palabra *nin~o*
- No se usan letras mayúsculas en las transcripciones.
- La diéresis es marcada con dos puntos : después de la vocal. Por ejemplo pingu:ino.

- Los números se escriben como palabras:

19	**dieci nueve**
20	veinte
29	veinti nueve
79	setentai nueve
700	sete cientos
932	nove cientos treintai dos
500	quinientos
1978	mil novecientos setentai ocho

Tabla 4.2 Los números se escriben como palabras

- Se descartan los archivos en los que sólo hay ruido en la línea o en los que no hay nada útil que transcribir.

- En etiquetas a nivel de texto, los < > se usan para las etiquetas **sin voz**. Ver tabla 1. Nada debe poner nada entre estos símbolos excepto esas etiquetas.

✍✍ Los corchetes **[ ]** son usados para transcribir una aproximación de una pronunciación cortada o un falso inicio. Por ejemplo:

tengo ocho a[n~os]*

El locutor nunca dijo "n~os", pero se pude inferir el resto de la palabra que falta y se coloca entre corchetes seguidos de un *. No se debe intentar adivinar la palabra incompleta si no se puede.

✍✍ Todas las transcripciones siguen un formato. Así, cualquier palabra sólo puede tener los siguientes símbolos adicionales:

***[ ]palabra[ ]*< >**

Las palabras son los únicos elementos necesarios en una transcripción. No se necesitan los *'s o los [ ]'s, pero si se colocan [ ] debe ir un *. Las transcripciones validas son todas las siguientes:

escuela (palabra básica)
***escuela** (parte del sonido "e" está cortado)
escuela* (parte del sonido "a" está cortado)
***[es]cuela** ("es" está cortado)
escue[la]* ("la" está cortado)
escuela <ln> (hay ruido en la línea cuando se dice la palabra)
escue[la]*<ln> (hay ruido en la línea cuando se dice "escue", "la" está cortado)
***[es]cuela<ln>** (hay ruido en la línea cuando se dice "cuela", "es" está cortado)
escuela<ln><bn> (hay ruido en la línea y ruido de fondo cuando se dice la palabra)

✍✍ Cuando se transcriben letras, o se deletrea un nombre, éstas se transcriben sólo como la letra y separadas por un espacio.

Se escribe **"H S T"** en lugar de "ACHE ESE TE"

También porque hay varias maneras de decir las letras. Por ejemplo, hay muchos problemas con el nombre de la "v", "Y", y "S". Puede ser que la letra "Y" se pronunció de dos maneras:

- como los fonemas /j/ /e/
- como la palabra "igriega" (/i/ /g/ /r/ /i/ /e/ /g/ /a/)
- como la letra I (/i/)

En el primer caso, la transcripción debe quedarse como está. En el segundo caso se debe cambiar la letra "Y" por la palabra "IGRIEGA". Finalmente, en el último caso, se debe cambiar la letra "Y" por la letra "i"

Transcripción original	Pronunciación	Transcripción Nueva
A D Y T	**/a/ - /d/ /e/** - /j/ /e/ - **/t/ /e/**	**A D** Y **T**
A D Y T	**/a/ - /d/ /e/** - /i/ /g/ /r/ /i/ /e/ /g/ /a/ - **/t/ /e/**	**A D** IGRIEGA **T**
A D Y T	**/a/ - /d/ /e/** - /i/ - **/t/**	**A D** I **T**

Tabla 4.3: Formas de decir las letras

Los relleno de pausas se deben etiquetar como suenan, pero no para que se confundan con las letras del alfabeto, los rellenos de pausas más comunes son:

em pronunciado /m/ o /em/
eh pronuncioado como /e/
am pronunciado como /am/

✍✍ Ruido entres dos palabras. Por ejemplo si se escucha un ruido entre las palabras "muchos" y "autos", se transcribe como:

había muchos <bn> autos

✍✍ Si ruidos o actividades sin voz ocurren simultáneamente cuando una persona está hablando, entonces se conecta la etiqueta sin voz junto a la palabra. Por ejemplo, se escucha ruido cuando la persona dice "muchos" se transcribe como:

había muchos<bn> autos

✍✍ Si ruidos o actividades sin voz ocurren simultáneamente durante todo la grabación se transcribe colocando la etiqueta apropiada al inicio la transcripción. Por ejemplo:

(bn) nací en la ciudad de Puebla y estoy estudiando la preparatoria

4.7 Acerca de la conexión de etiquetas especiales

Las etiquetas especiales para el etiquetado a nivel de texto se dividen en tres categorías:

Conectadas a palabras	**NO deben ir con palabras**	**Pueden usarse de cualquier manera**
<long>	 	<bn>
<sp>	<burp>	<bs>
<glot>	<cough>	<fp>
	<ct>	<laugh>
	<ls>	<ln>
	<pau>	<nitl>

	<sneeze>	<unk>
	<tc>	

Tabla 4.4 Etiquetas especiales para el etiquetado a nivel de texto

Etiquetas especiales

Etiqueta	***Descripción***
*	Pronunciación cortada
<beep>	Sonido de "bip"
<bs>	Habla de fondo
<nitl>	No está en el idioma
<long>	Alargar una pronunciación
<fp>	Relleno de pausa genérico
<bn>	Ruido de fondo
 	Ruido de respiración
<burp>	Eructo
<ct>	Aclarando la garganta
<cough>	Tos
<laugh>	Risas
<ls> <ln>	Ruido con los labios Ruido en la línea
<pau>	Pausa
<sneeze>	Estornudo
<sniff>	Aspiración
<tc>	Ruido con la lengua
<unk>	No se pude identificar
<vb>	Leer en voz baja
<pm>	Pronunciar mal una palabra

<glot>	Glotalización

Tabla 4.5: Etiquetas sin voz para transcripciones a nivel de texto