Capitulo 4 Desarrollo del Protocolo de Grabación

4.1 Protocolo de grabación

El protocolo de grabación es una de las partes más importantes del proceso para crear un corpus para síntesis de voz, debido a que este documento establece las frases y palabras que se van a grabar con el fin de obtener el mejor resultado (véase Apendice C y D). En este caso, se busca realizarlo con la cantidad necesaria de letras, números, palabras y frases que abarquen casi todos los posibles sonidos del lenguaje español de México, con el fin de reproducir cualquier documento para niños de 6 a 12 años.

4.2 Grabación

La grabación de los archivos de sonido, es la segunda fase en la construcción del corpus de voz, el objetivo es grabar dentro de un directorio el protocolo como archivos de sonido. Dicha fase se realiza usando un sistema de grabación hecho con el CSLU Toolkit, como se muestra en la figura 4.1.

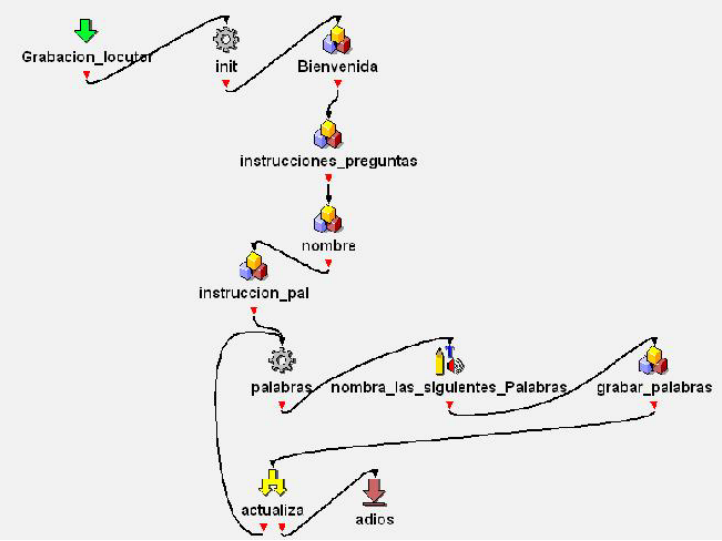


Figura 4.1 Diagrama del sistema de grabación de voz

EL CSLU Toolkit es un conjunto de herramientas que provee un ambiente poderoso y flexible para el desarrollo de sistemas de voz, está diseñado para facilitar el desarrollo rápido de sistemas del lenguaje hablado para una variedad de aplicaciones, así como también para proveer una plataforma para realizar investigaciones en tecnologías de voz [Serridge, 1998].

Por esa razón se realizó el sistema para la grabación del locutor, el cual está compuesto de herramientas que permiten hacer una mejor grabación de audio. Se utilizaron documentos de texto (.txt), un micrófono Plantronics y un cuarto de grabación para aislar el sonido. En la pantalla se logró un mejor ambiente, para que el locutor se encuentre cómodo y tenga una mejor atención al leer, como se muestra en la figura 4.4.

El sistema graba la voz del locutor transformándolo a un archivo (.wav) con un formato de audio PCM de un solo canal (mono), una calidad de 256 Kbps (Bit Rate) y muestreo 16 bits.

El sistema ésta hecho con el fin de facilitar la grabación del locutor ya que los documentos contienen la información necesaria y así nos olvidemos de las hojas de papel, asimismo, evitar el ruido que se puede grabar externo a la voz, como los ruidos ambientales. Automáticamente, el programa se encarga de mostrar lo que se quiere grabar, lo único que tiene que hacer el locutor es leer lo que aparece en la pantalla. En la tabla 4.1 se da una pequeña explicación de lo hace cada uno de los componentes para realizar la grabación.

	Inicio
	INIT es donde damos todas las variables y direcciones que necesitaremos para la grabación
	Da un mensaje de bienvenida
	Da las instrucciones necesarias para poder comenzar la grabación
	Graba el nombre del locutor
	Da las instrucciones necesarias para grabar las palabras
	Indica el siguiente renglón del protocolo de grabación o si es todo lo que contenía el documento
	Se encarga de mostrar el contenido de cada uno de los renglones que contiene el protocolo
	Graba automáticamente lo que dice el locutor en el formato .wav
	Es una decisión si es todo o continuamos grabando el protocolo
	Fin de la grabación

Tabla 4.1 Muestra la fase de grabación del protocolo.

La interfaz que se presenta con este sistema de grabación es completamente amigable, debido a que no se requiere de que el locutor memorice todo lo que tiene que grabar, ni recurrir a las hojas de papel y pensar "¿En cual me quede?", ¿Cual sigue?, ¡Ya me equivoque!, etc., por eso nos muestra una ventana donde le indica al locutor si esta grabando o no y otra donde muestra lo que tiene que leer.

En la figura siguiente se muestra claramente lo que hace cada una de las ventanas:

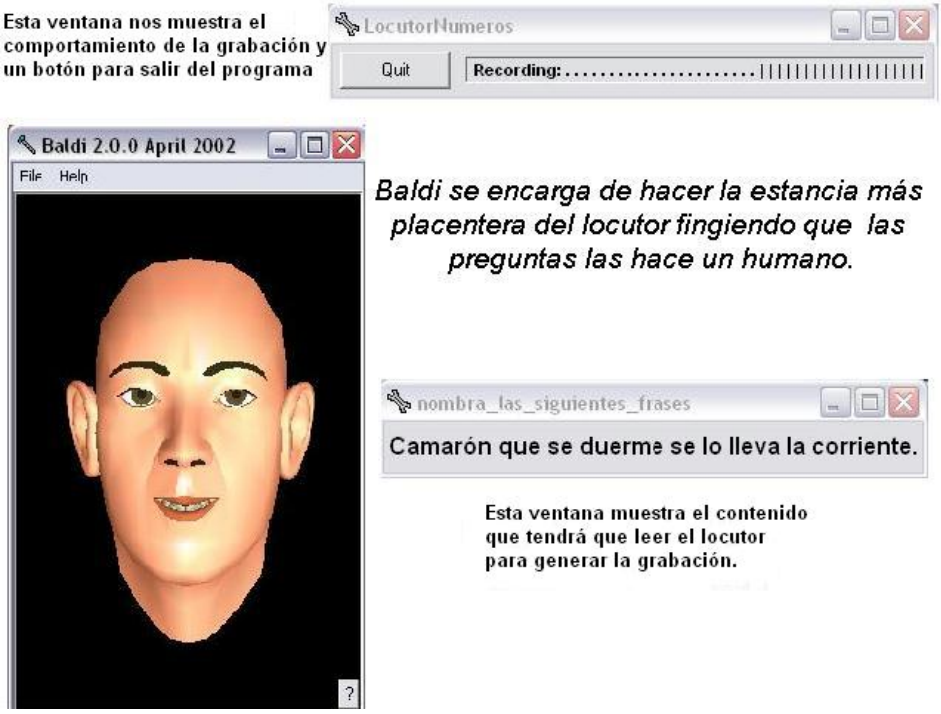
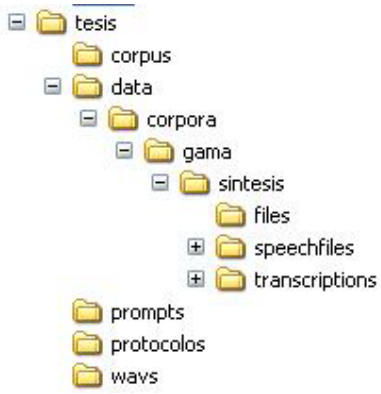


Figura 4.2. Muestra la interfaz del sistema de grabación

Para el almacenamiento se tiene un fólder llamado "tesis" el que contiene todos los folders, archivos y documentos que se requieran para la grabación. Este folder está compuesto de cuatro folders llamados corpus, data, prompts y protocols.



En el folder *corpus* se guarda toda la información escrita de las grabaciones en formato de texto, para posteriormente usarla en la etiquetación a nivel texto, que se verá en la siguiente sección. El folder *data* contiene otro folder llamado *corpora* en dónde se encuentra el folder *gama* que agrupa todo el corpus. En el fólder *gama* se encuentra otro llamado *síntesis,* que es para separar los documentos importantes para sintetizar, en éste existe otro llamado *speechfiles*, el sistema anteriormente mencionado genera automáticamente un fólder nuevo por sesión junto con una bitácora de grabación en la que nos describe los tiempos y días en que se generó cada una de las grabaciones. En el folder *prompts* se guardan todos los mensajes que dirá Baldi con el fin de darle la comodidad al locutor y dar las instrucciones debidas de lo que tiene que hacer y cómo lo tiene que hacer. Esto se hace con grabaciones pregrabadas con el fin de ser más eficiente. Por último, en el folder de *protocolos* que es donde se guarda toda la información de lo que el locutor tiene que leer para lograr la grabación total.

Al comenzar las pruebas de grabaciones, se observó que la cantidad óptima de grabaciones por sesión era de 250 archivos de audio (.wav). Esto es para que el locutor no se canse y su voz no comience a variar demasiado.

De esta forma, se realizaron siete sesiones, en las que se grabaron todas las frases. Éstas se revisaron, el 93% resultaron tener buena calidad y el 7% se grabaron con errores de pronunciación, distracciones, fallos de lectura y confusiones. Para solucionar este problema se realizó otra sesión con las grabaciones malas para cumplir

con el balance que el programa de comparación con los archivos de texto que contienen lo que se grabó, con el archivo de silabas en las que se busca agrupar todos y cada uno de los sonidos hablados en México. En los apéndices C y D se muestra la cantidad de grabaciones.

Así logramos tener un corpus completo y amplio. Después de grabar se revisaron minuciosamente archivo por archivo para garantizar la mejor calidad de voz posible.

4.3 Etiquetado del corpus

En el artículo escrito por Alejandra Olivier nos comenta [Olivier, 2000] sobre el etiquetado de los archivos de audio, que es la parte mas importante para lograr la concatenación de estos segmentos de audio para que la voz del sintetizador sea continua y clara, por lo tanto, el objetivo es etiquetar el corpus *GAMA*, el cual consta de 2 locutores. Se deberá etiquetar las frases, palabras y la gramática numérica (1400 archivos de audio .wav) utilizando las herramientas del CSLU Toolkit. Etiquetar un archivo es colocar la palabra o fonema alineadas a las ondas de voz que generan las grabaciones de tipo wav donde ocurre cada palabra y cada fonema, la información que es necesaria para crear el sintetizador.

Para minimizar la cantidad de trabajo, esta tarea se divide en 5 partes:

1.- Crear las etiquetas a nivel de texto.

- Crear las transcripciones del habla espontánea (preguntas y monólogo).
- 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.

4.3.1Etiquetado del corpus manualmente

Se utiliza la ventana del tcsh donde se puede obtener la etiquetación realizando los siguientes pasos:

Figura 4.3 Shell de UNIX utilizado por Tlatoa para la etiquetación.

1.- Verificar las transcripciones no alineadas

- La verificación se puede hacer usando los siguientes programas que residen en el directorio **/src/tcl/.**
- Se crea el archivo de referencia .files para cada locutor (folder), el cual se encuentra ubicado dentro de la ruta c:/tesis/data/corpora/gama/síntesis/files, en este caso será realizado un archivo para la voz del hombre y otro para la voz de la mujer. Utilizando la siguiente línea se logra crear el archivo.

  >tcl **mk_file_children.tcl** –user **0** –corpus **gama** –escuela **síntesis<**

  Donde "0" es el locutor, gama es el corpus y síntesis es un sub-folder donde se almacena la información necesaria para el sintetizador.

- Se tienen que generar automáticamente las transcripciones a nivel de texto (archivos .txt) para cada locutor de acuerdo al protocolo que grabó.

  >tcl **crear_txt_children.tcl** –user **0** –protocol **1.txt** –escuela **síntesis<**

  El folder *protocol* contiene los archivos de texto que agrupa todas las frases y palabras que se grabaron como archivo de audio.

- Estas transcripciones se almacenan en archivos de texto con la extensión .txt, los cuales contienen las palabras y frases que se escucharon en el archivo de audio. A menudo estos archivos pueden ser creados automáticamente, porque se sabe de antemano lo que la persona iba a decir. En ese caso, las transcripciones deben ser verificadas a mano, porque en ocasiones la persona no dirá lo que se supone. Esto se puede hacer eficientemente utilizando el programa check_txt_files.tk:

  >tk **check_txt_files.tk** –files **/tesis/data/corpora/gama/sintesis/files/0.files<**

  Este script muestra una ventana donde se puede reproducir el archivo de audio (.wav) y escribir lo que el locutor dijo para lograr obtener los archivos .txt como se muestra en la figura 4.4.

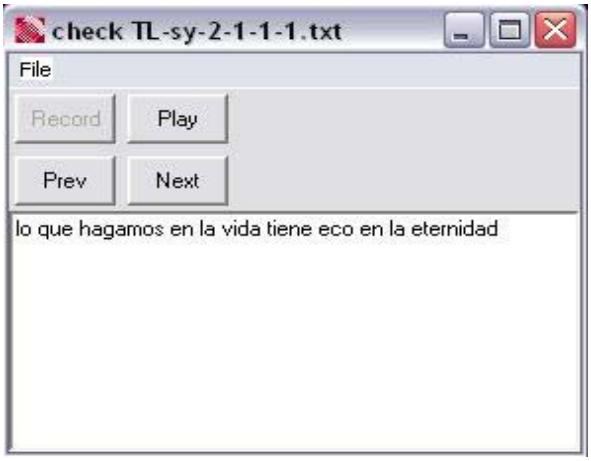


Figura 4.4 Ventana que reproduce el audio que se quiere etiquetar a nivel texto.

- Para sustituir acentos, signos de puntuación en las trascripciones será utilizado clean_txt.tcl el cual se encarga de colocar los acentos de la manera apropiada por ejemplo **árbol**- la forma correcta es **a'rbol.**

  >tcl **clean_txt.tcl** –files **/tesis/data/corpora/gama/sintesis /files/0.files<**

2. Crear las etiquetas a nivel de palabra alineadas en tiempo

El siguiente script crea los archivos .wrd a partir de los .txt, que ya se han revisado anteriormente. Para crearlos, se usa un algoritmo 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 **/tesis/data/corpora/gama/sintesis/files/0.files** –vocab **exceptions.vocab** –rules **tts_rules.txt** –dur **phones.dur<**

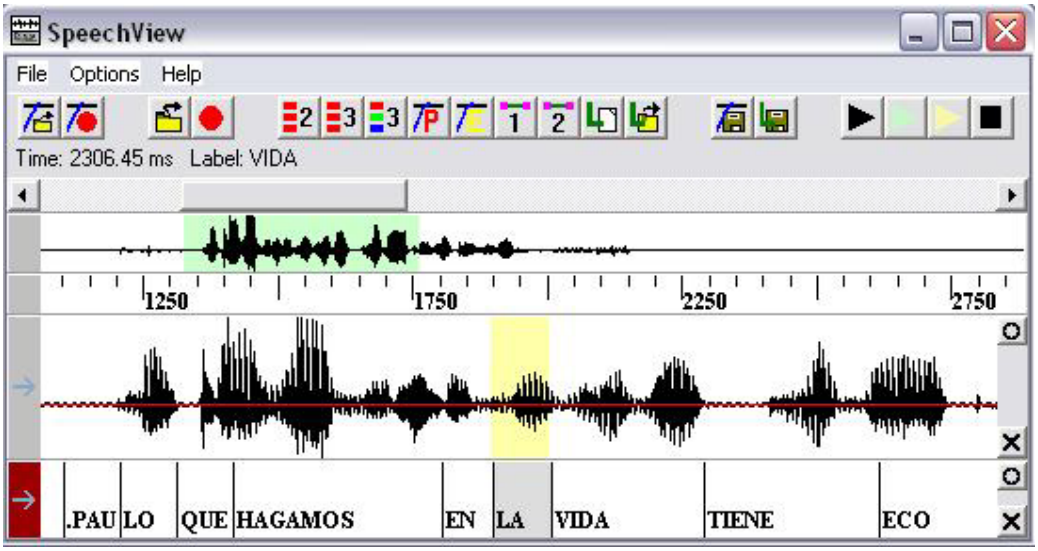
3. Ajustar manualmente las trascripciones a nivel de palabra, creadas en el paso anterior.

Para ajustar las etiquetas, se utilizara la herramienta SpeechView, la cual es parte del CSLU Toolkit .

a) **Se abre el Speech View.**

b) **Se ingresa el archivo .wrd**

c) **Se ajustan las fronteras de cada palabra manualmente de acuerdo con la información de la señal.**



Nota: Esto se hace con cada uno de los archivos de audio (.wav)

Figura 4.5 SpeechView alineación de etiquetas de tipo palabra con audio.

4. Quitar los silencios largos de las grabaciones.

Muchas veces, las grabaciones tienen 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 será reducido el tamaño de algunos archivos y el tiempo que se requiere para entrenar.

> tcl **wrd_trim_wav.tcl** –files **/tesis/data/corpora/gama/sintesis/files/0.files**

Antes de borrar los archivos .wav modificados marcados como .old, debemos verificar que las nuevos .wav, se encuentren en las condiciones que requieren para el proyecto.

5. Generar automáticamente las transcripciones a nivel de fonema.

A partir de los .wrd ya revisados, haremos lo mismo realizado a nivel de palabra y creamos los .phn.

> tcl **wrd2phn.tcl** –files **/tesis/data/corpora/gama/sintesis/files/0.files** –vocab **exceptions.vocab** –rules **tts_rules.txt** –dur **phones.dur<**

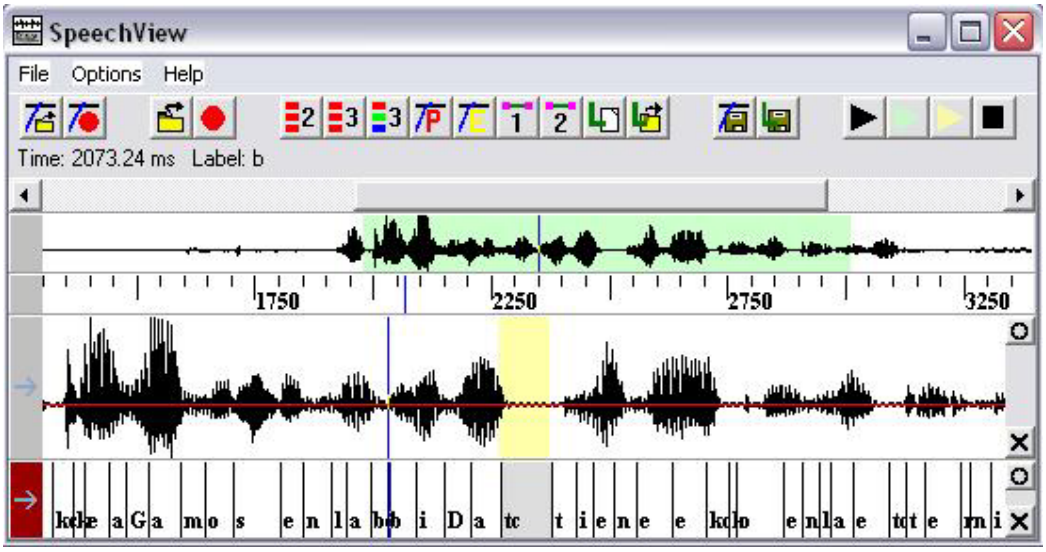
6. Ajustar manualmente las transcripciones alineadas a nivel de fonema.

Para ajustar las etiquetas se usa el mismo procedimiento 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.

a) **Se abre el Speech View.**

b) **Se ingresa el archivo .phn**

c) **Se ajustan las fronteras de cada fonema manualmente de acuerdo con la información de la señal.**



Nota: Esto se hace con cada uno de los archivos de audio (.wab)

Figura 4.6 SpeechView alineación de etiquetas de tipo fonema con audio.

7. 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 pueden descubrir automáticamente usando el siguiente script:

> tcl **detect_problems.tcl** –files **/tesis/data/corpora/gama/sintesis/files/0.files<**

8. Realinear las transcripciones 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

/tesis/data/corpora/gama/sintesis/files/0.files<

9. Crear nuevamente las transcripciones no alineadas (archivos .txt) a partir de las alineadas.

De la misma manera, si se cambia alguna palabra utilizando SpeechView, también hay que cambiar la palabra en los archivos no alineados (.txt).

>tcl **wrd2txt.tcl** –files **/tesis/data/corpora/gama/sintesis/files/0.files –noskip<**

4.3.2 Etiquetado del corpus Automáticamente (Forced Alignment)

Nos comenta Alejandra Olivier [Olivier, 2000] que el primer paso para hacer forced alignment, es organizar la estructura de directorios de la computadora. El CSLU Toolkit asume la estructura que se debe seguir para un corpus. Se sugiere tomar en cuenta las siguientes recomendaciones.

- El corpora debe residir en la ruta **c:/tesis/data/corpora.**
- El archivo corpus debe estar en la ruta c**:/tesis/data/corpora/gama/sintesis**. Cada directorio del corpus incluye dos sub-directorios: **speechfiles y transcriptions**.
- Dentro de **speechfiles** hay un sub-directorio para cada locutor, el cual contiene los archivos de audio (.wav)
- Dentro de **transcriptions** hay sub-directorio para cada locutor, el cual contiene los archivos con las transcripciones a nivel de texto (.txt), palabra (.wrd) y fonema (.phn)

El toolkit debe residir en la ruta **c:/cslu**

En los siguientes pasos se describen como etiquetar un corpus automáticamente. Se asume que existen los archivos de voz en **speechfiles** (.wav) y las transcripciones a nivel de texto del corpus en **transcriptions** (.txt).

1. Crear el archivo **gama.files** usando script **make_all_files.tcl**. El archivo tipo files, sirve de referencia para generar el vocabulario y las etiquetas automáticas. En él se especifican los archivos que se desean etiquetar y el directorio donde se localizan.

<tcl **make_all_files.tcl** –files **gama.files** –corpus **gama** –cat>

-*files* indica el archivo de salida.

-*corpus* indica el nombre del corpus del que se desea generar el archivo tipo files.

-*cat* es una bandera que indica si se quiere incluir la localización del los archivos de categorías, las cuales serán ubicadas en el mismo directorio de transcripciones. Es necesario usar esta bandera, si se desea etiquetar a nivel de categorías.

2. Crear el archivo **test_digits.vocab**. Usando **generate_vocab.tcl** o manualmente.

Aquí está un ejemplo del archivo baseline.vocab. El vocabulario generado por este script es una lista de todas las palabras diferentes que incluye un corpus, seguidas de su transcripción fonética.

<tcl **generate_vocab.tcl** –files **gama.files** –rules **tts_rules.txt**

–vocab **exceptions.vocab** -out **baseline.vocab** –noskip>

NOTA: Revisar el archivo .vocab ya que es muy importante que las pronunciaciones estén incluidas en los fonemas. El vocabulario no necesariamente tiene que contener una gramática.

3. Revisar que se tengan los archivos necesarios y la red a usar en el directorio. Por ejemplo en el directorio baseline debe contener el reconocedor para etiquetar el corpus de **gama**. En este caso se usa el reconocedor de propósito general nnet.17.

- **baseline.vocab**
- **baseline.olddesc**
- **nnet.17**
- **gama.files**

4. Generar etiquetas usando **forced_alignment.tcl**

<**tcl forced_alignment.tcl** –files **/tesis/data/recognizers/gama/baseline/gama.files** –

name **baseline** –corpus **gama** –nnet **nnet.17** –level **pw**>

-*files* indica el archivo que contiene la ubicación de los archivos .wav, .phn, .cat, .txt y .wrd.

-*name* indica el nombre del directorio en dónde reside el experimento de etiquetado automático.

-*corpus* indica el nombre del corpus a etiquetar.

-*nnet* indica el nombre de la red que se usará para etiquetar.

-*level* indica el tipo de etiquetas que queremos generar, "p" para nivel fonético, "c" para nivel de categorías, y "w" para etiquetado a nivel de palabra.

NOTA: Es importante indicar el path completo para accesar el archivo .files.

La ejecución de **fa.tcl** creará un directorio nombrado con el nombre del experimento y la terminación fa. En este caso creará el directorio baselinefa donde residen los archivos de información y un readme de los pasos realizados.

5. Revisar etiquetas creadas utilizando la herramienta speechview.tcl o manualmente como se explico en la sección anterior:

<tk **/cslu/Toolkit/2.0/script/sview_1.0/speechview.tcl** –corpus **gama.files>**

Es importante conocer los dos tipos de etiquetados que existen, pero también es necesario de cualquier manera revisar la alineación de los archivos de audio (.wav) con los archivos de texto (.txt), de palabras (.wrd) y de fonemas (.phn), para crear una mejor calidad de voz. Por lo que se considera conveniente hacer el etiquetado manual y corregir los errores que no detecta el *forced alignment*, por esta razón este proyecto será etiquetado manualmente para lograr la calidad requerida del corpus.