Evaluación de métodos de determinación automática de una transcripción fonética

Alejandra Olivier
is095462@mail.udlap.mx
Laboratorio Tlatoa, CENTIA[1]
Universidad de las Américas-Puebla
72820 Sta Catarina Martir, Puebla
México

Ingrid Kirschning
ingrid@mail.udlap.mx
Laboratorio Tlatoa, CENTIA
Universidad de las Américas-Puebla
72820 Sta Catarina Martir, Puebla
México

Abstract

Para la investigación y el desarrollo de aplicaciones de tecnologías de voz uno de los principales requerimientos es una colección de datos de voz (corpus) correctamente etiquetados. Este proceso de etiquetado es tedioso y tiende a provocar errores inherentes a las tareas largas, tediosas y repetitivas. Además, la determinación de los límites fonéticos de manera manual puede estar sujeta a decisiones subjetivas. Esto provoca que las etiquetas generadas por un usuario pueden diferir fuertemente de las generadas por otra persona. Es por estas razones que se reconoce la necesidad de una herramienta confiable que pueda realizar esta tarea de manera automática. En este artículo se describe un trabajo de tesis, el cual consistió en el desarrollo y evaluación de una herramienta basada en un proceso llamado *forced alignment*, la cual permite determinar automáticamente limites fonéticos. La evaluación de la herramienta de etiquetado automático muestra muy buenos resultados y se ha empezado a usar en los proyectos del grupo de investigación en tecnologías de voz TLATOA de la Universidad de las Américas-Puebla, ya que agiliza la creación de sistemas de reconocimiento de voz. La documentación de este proceso se encuentra disponible en http://info.pue.udlap.mx/~sistemas/tlatoa/howto/labelfa.html

1 Introducción

La alineación de una señal de voz con su correspondiente transcripción fonética es un proceso esencial en la investigación del habla. De ahí que la alineación fonética y reconocimiento automático de voz (ASR) sean tareas muy relacionadas. En reconocimiento de voz, dada una grabación, nos interesa identificar lo que se ha dicho, sin importar donde empieza cada segmento individual (palabra, sílaba o fonema). Por otro lado en el caso de la alineación fonética, se supone que se conoce lo que se dijo en la grabación y únicamente nos interesa saber cuáles son los puntos de inicio y fin de cada segmento individual [Rapp, 1995].

Determinar los puntos de inicio y fin de cada segmento individual se conoce como etiquetado de una señal, y la creación de cualquier reconocedor de voz requiere de un corpus grande etiquetado a nivel de fonemas. Además, con una suficiente cantidad de datos alineados en tiempo (etiquetados), se pueden cuantificar las propiedades de los segmentos fonéticos y describir como sus características se modifican por el contexto, lo cual conduciría a un mejor modelo de la producción de voz y el desarrollo de mejores técnicas para el reconocimiento de voz.

Tradicionalmente, el alineamiento de límites fonéticos o etiquetado es hecho manualmente por un experto que escucha la señal de voz y utilizando alguna herramienta, visualmente examina la señal y coloca las etiquetas. Este proceso tiene varias desventajas:

- Consume mucho tiempo. Etiquetar unos segundos de habla puede tomar varios minutos.
- Se requiere de habilidad y conocimiento específico para identificar correctamente la porción de la señal que corresponde a cierto símbolo en el texto.
- Hay falta de consistencia y reproducción de resultados.
- El etiquetado involucra decisiones subjetivas.
- Existe el problema de error humano asociado con tareas tediosas y repetitivas.

Debido a los problemas asociados con el etiquetado manual, junto con la necesidad de un corpus grande etiquetado correctamente, surge la necesidad de un sistema capaz de realizar el proceso de manera automática. El propósito de este artículo es presentar el desarrollo de nuevas herramientas que permiten etiquetar automáticamente una base de datos de voz, y la comparación de los resultados obtenidos por estas herramientas con el proceso manual.

Las actividades comprendidas en el desarrollo de este trabajo son las siguientes:

[1] Centro de investigación en tecnologías de información y automatización.

- Desarrollo de software que facilite el proceso para determinar limites fonéticos.
- Documentación del proceso.
- Pruebas de evaluación.

La organización de este artículo sigue la estructura de las actividades comprendidas en este trabajo, mencionándose primero los componentes básicos de sistemas de reconocimiento de voz y la tecnología usada para su creación.

2 Reconocimiento de voz y tendencias actuales

La meta de las investigaciones en reconocimiento de voz es desarrollar nuevas técnicas y sistemas que permitan a la computadora aceptar la voz como entrada. La voz como dispositivo de entrada ofrece a la computadora muchas ventajas, pues para el humano es un medio natural, rápido y flexible, ya que permite tener las manos y ojos libres, y la persona puede localizarse en cualquier parte.

Algunas de las principales aplicaciones de los reconocedores de voz son en servicios de telefonía y servicios de información, como por ejemplo conmutadores automáticos, balances financieros, horarios de cine e información de vuelos.

Por otra parte, la última década se ha caracterizado con el surgimiento de una nueva especie de interfaces hombre – computadora que combinan varias tecnologías de lenguaje humano, y permiten accesar información y realizar procesos de transacción usando diálogos. Este nuevo tipo de interfaces se conocen como sistemas conversacionales y surgen de la necesidad de brindar un "acceso universal" a la información, a cualquier hora, en cualquier lugar. La solución a este problema es proveer a la máquina capacidades humanas a través de estos sistemas, para poder entender, hablar y escribir tal como los usuarios con los que interactua. La meta de los sistemas conversacionales es el manejo de diálogos con una mezcla de iniciativa, en los cuales ambos, el usuario y la computadora participen activamente para resolver un problema usando un paradigma coversacional. [Zue, 1997].

Los sistemas conversacionales constituyen una de las principales metas de las investigaciones que se están iniciando actualmente y marcan la tendencia de los desarrollos futuros en esta área. Sin embargo, un componente clave de estos sistemas lo conforman los reconocedores de voz. Sin un buen reconocimiento no se puede intentar interpretar ni responder a lo que el usuario dijo.

2.1 Reconocimiento de voz utilizando redes neuronales.

El CSLU toolkit, desarrollado por el Center for Spoken Language Understanding (CSLU) del Oregon Graduate Institute (OGI), es un es ambiente de software para investigación y desarrollo, que provee una herramienta poderosa y flexible para crear y usar sistemas de lenguaje hablado. Uno de los ambientes de desarrollo de reconocedores es CSLU-NN, una herramienta que nos permite crear reconocedores de voz basados en redes neuronales [Cosi et al., 1998].

Los reconocedores son desarrollados usando muestras de la señal tomadas cada t unidades de tiempo, lo que llamamos marcos o frames, junto con una red neuronal para estimar las probabilidades posteriormente. Los pasos realizados durante el reconocimiento se muestran en la figura 1 y son los siguientes:

- Dividir la señal en muestras (marcos);
- Extraer las características de cada marco. Estas características describen el espectro que envuelve a la señal en ese marco y en un pequeño número de marcos vecinos.
- Las características de cada muestra son clasificadas en categorías fonéticas usando una red neuronal. Las salidas de la red neuronal son usadas como estimaciones de probabilidad para cada categoría fonética de la muestra actual.
- Determinar la(s) palabra(s) más parecidas a través de la búsqueda Viterbi [Serridge, 1997]. Esta búsqueda usa una matriz de probabilidades (las salidas de la red neuronal) y un conjunto de modelos de pronunciaciones.

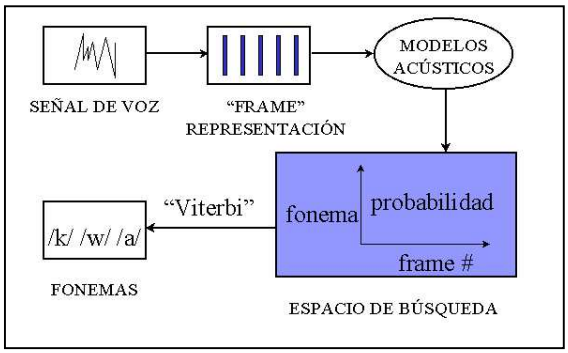


Figura 1. Sistemas de reconocimiento de voz basado en redes neuronales.

3 Desarrollo del sistema de etiquetado automático

Para entrenar un reconocedor se necesita contar con una grande porción de datos de voz correctamente etiquetados a nivel de fonema. En esta sección se presentan las características de las diferentes etiquetas, el método de alineación y un diagrama funcional de la herramienta desarrollada.

3.1 Tipos de transcripciones

Las etiquetas son símbolos que nos permiten identificar una porción de la señal de voz. Existen diferentes tipos de etiquetas, tales como etiquetas que identifican frases, palabras, sílabas o fonemas. En general se puede usar cualquier identificador de un segmento individual de la señal. De acuerdo con las convenciones del CSLU, se manejan tres niveles de transcripciones: a nivel de texto, a nivel de palabra y a nivel fonema [Lander, 1996].

Transcripciones ortográficas o a nivel de texto

Los etiquetas a nivel de texto nos indican el contenido de una pronunciación, es decir son una transcripción textual de lo que se dijo, sin indicar el tiempo de inicio y de fin. Se

representan con la ortografía estándar y no se hace una distinción entre las partes de la señal con voz o sin voz.

Generalmente las transcripciones a nivel de texto son creadas en un editor de texto y su contenido es como el de cualquier texto pero sin puntuación, ni distinción entre letras mayúsculas y minúsculas o indentación. Por ejemplo, la transcripción de a nivel texto de la frase "9 2" sería NUEVE DOS. Otra manera de obtener la transcripción a nivel de texto es a partir de una de las herramientas del CSLU Toolkit, un script que revisa los archivos de sonido, los reproduce y nos permite teclear el texto correspondiente. Véase figura 2. La transcripción no necesariamente tiene que ser correcta ortográficamente, pero sí debe representar lo que dijo el locutor.

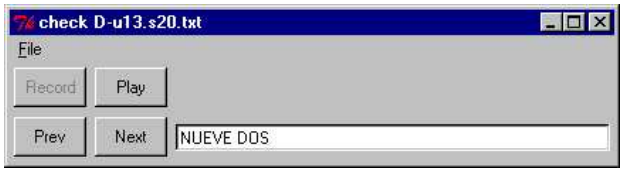


Figura 2. Transcripción a nivel de palabra de la grabación "9 2"

Transcripción a nivel de palabras

La transcripción a nivel de palabra, también se representa con la ortografía estándar. Se distingue entre los que son segmentos de voz y los que no lo son. Además las etiquetas son alineadas en tiempo, indicando así el punto de inicio y fin de cada palabra. La figura 3 muestra el ejemplo anterior con transcripciones a nivel de palabra alineadas en tiempo.

Transcripción fonética o a nivel fonema

Las transcripciones fonéticas, representan el contenido fonético de una pronunciación en cierto nivel de detalle. Existen distintos tipos de convenciones para transcripciones fonéticas. Como referencia básica del conjunto de fonemas que usamos para etiquetar español mexicano se considera el *Worldbet* [Cole et al, 1994], el cual define un conjunto de símbolos fonéticos ASCII que intentan representar todos los fonemas de cualquier idioma del mundo.

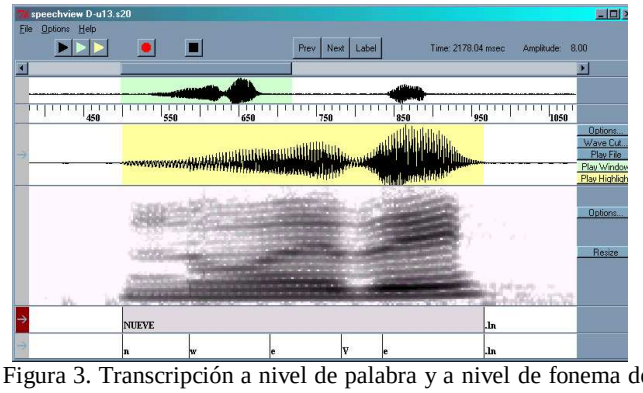


Figura 3. Transcripción a nivel de palabra y a nivel de fonema de la frase "NUEVE". En la parte superior de la figura se muestra la señal, seguida por su espectrograma, la transcripción a nivel de palabra y la transcripción a nivel de fonema.

La figura 3 muestra la obtención de transcripciones fonéticas de la señal. Después de haber obtenido la palabra que representa esa porción de la señal, podemos obtener los símbolos fonéticos correspondientes a esa palabra, utilizando un script que convierte cada palabra a su correspondiente representación fonética, usando un conjunto de duraciones y reglas definidas previamente. Después se alinean los símbolos fonéticos adecuadamente. Véase el ejemplo 1. A este nivel se empieza a distinguir las diferentes pronunciaciones de letras. Así, se aprecia por ejemplo, que "NUEVE" se transcribe utilizando la "w" para el sonido particular de la letra u. También existe una gran cantidad de información acústica en la señal, como en este ejemplo la etiqueta ".ln" (*line noise*) que identifica ruido en la línea de telefóno Sin embargo no toda la información acústica es importante, de hecho mucha es completamente ignorada por el sistema de percepción humana. La meta de obtener transcripciones fonéticas, es representar el contenido fonético de una pronunciación con cierto detalle, o al menos la información fonética más importante para fines de procesamiento automático de esas señales de voz. Para la representación detallada del contenido fonético de una pronunciación, se usan diacríticos. Estos son un conjunto de transcripciones que usan un subrayado "_" como símbolo de ligado. Por ejemplo, la vocal a con sonido nasal se transcribe como a_n.

```
MillisecondsPerFrame: 1.0
END OF HEADER
0 500 .pau
500 586 n
586 690 w
690 782 e
782 835 V
835 966 e
966 1429 .ln
```
Ejemplo 1. Archivo con etiquetas a nivel de fonema, alineadas en tiempo de la frase "NUEVE".

3.2 *Forced alignment*

El planteamiento del problema de este trabajo hace clara la necesidad de contar con una herramienta automática que permita determinar las transcripciones fonéticas. La solución esta basada en un proceso llamado *forced aligment*. Este proceso nos permite determinar limites fonéticos a partir de la transcripción a nivel de texto y el archivo de sonido de una señal de voz. Sin embargo, éste no es un proceso automático y no ha sido probado sobre el idioma español. La automatización del proceso manual de alineación requiere revisar cada paso del proceso de *forced alignment*. El resultado de aplicar *forced alignment* es la obtención de categorías de un corpus del cual sólo se tenían transcripciones a nivel de texto. Las categorías fonéticas son los contextos en los que se presenta un fonema, por ejemplo `o>s,` describe que el fonema `o` precede al fonemas.

Para determinar limites fonéticos usando el protocolo de *forced alignment* se requiere de las transcripciones a nivel de texto del corpus y un reconocedor. Los limites fonéticos se obtienen siguiendo los siguientes pasos.

- Con el texto correspondiente a las grabaciones se crea el vocabulario.

- Se crea una gramática restringida a dar el resultado correcto a partir del vocabulario.
- Se usa el reconocedor para evaluar las características obtenidas de la señal previamente.
- Se realiza la "búsqueda" restringida por la gramática que sólo acepta la secuencia de símbolos correcta. Así, de la búsqueda se obtiene la secuencia correcta de símbolos y su duración. La duración está dada en múltiplos de 10ms ya que las muestras de la señal se obtienen cada 10 milisegundos.

El nuevo sistema se basa en la automatización del protocolo de *forced alignment* y nos permitirá obtener las transcripciones fonéticas automáticamente.

3.3 Implantación de *forced alignment* automático

El diseño de la herramienta de etiquetado automático basado en el proceso de *forced aligment* se resume en la figura 4. Usando la señal de voz y su transcripción a nivel de texto como entradas, se puede obtener transcripciones a nivel de fonema alineadas en tiempo. Para este proceso se utiliza un reconocedor.

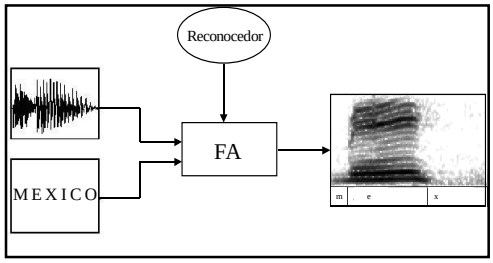


Figura 4. Diagrama funcional de la herramienta de etiquetado automático

Para la implantación de las herramientas que permiten la evaluación y el desarrollo de un sistema de etiquetado automático se utilizó Tcl [Ousterhout 1997], un script que permite mantener la consistencia en las herramientas del toolkit y manejar los módulos de sus tecnologías centrales (Reconocimiento de Voz, Síntesis de Voz, y Animación facial). De esta manera se implanta una extensión del toolkit sin tener que cambiarlo.

Durante el desarrollo del proceso automático se implantaron nuevos scripts en Tcl basados en la estructura del toolkit, en cuanto a la organización de los datos.
El proceso automático comprende los siguientes pasos.

- Crear un archivo con la ubicación del corpus
- Crear el vocabulario del corpus a etiquetar.
- Revisar que se tengan los archivos necesarios y la red en el mismo directorio.
- Generar las transcripciones a nivel de fonema del corpus.

Archivo de ubicación del corpus
El archivo de ubicación del corpus contiene la ruta de directorios donde se localizan los archivos necesarios para generar las etiquetas. Estos son los archivos correspondientes a cada señal de voz, el archivo de transcripciones a nivel de texto, y el lugar en donde estarán los archivos creados durante este proceso, los archivos de categorías y los archivos de transcripciones a nivel de fonema.

Vocabulario del corpus
El vocabulario del corpus se obtuvo creando un script que automáticamente revisa cada archivo de texto y extrae cada palabra distinta. Escribe una lista de las palabras encontradas y agrega su transcripción fonética, basada en un conjunto de reglas definidas anteriormente. En el ejemplo 2 se muestra el vocabulario del corpus de dígitos.

```
cero     {s e r o};
uno      {u n o};
dos      {dc d o s};
tres.....{tc t r e s};
cuatro   {kc k w a tc t r o};
cinco    {s i N kc k o};
seis     {s e i s};
siete    {s i e tc t e};
ocho     {o tSc tS o};
nueve    {n w e V e};
diez     {dc d i e s};
```

Ejemplo 2. Vocabulario para el corpus de dígitos.

Se deben colocar los siguientes archivos en un mismo directorio de experimento.

- La red neuronal
- El archivo descriptor
- El vocabulario
- Archivo de información
- Archivo de ubicación del corpus

La red neuronal es el reconocedor capaz de reconocer el texto que representa la señal. Puede usarse un reconocedor de propósito general o de propósito específico. Por ejemplo si se desea etiquetar ciudades y se tiene un reconocedor entrenado con datos de ciudades, éste será un reconocedor de propósito específico. La función del reconocedor es la evaluación de las características de la señal. Se entrega al reconocedor un conjunto de vectores de 130 características MFCC, tomadas de las muestras extraídas de la señal, cada 10 milisegundos. El reconocedor ya sabe a qué categoría corresponden esas características, ya que cuenta con el archivo de las categorías específicas para cada grabación; así que la salida es la secuencia correcta de categorías con sus tiempos de inicio y fin.

Es muy importante saber que categorías identifica el reconocedor, para saber si nos será útil. Antes de generar categorías, se revisa el contenido de cada archivo de texto y se verifica que todos los contextos aparezcan en el archivo descriptor de contextos. El archivo descriptor contiene las definiciones de categorías que se pueden reconocer, es decir la lista de contextos en que se presenta cada fonema. El archivo descriptor contiene todos los contextos de los que aprendió la red.

Generar etiquetas automáticamente

Una vez realizado lo anterior, el proceso de etiquetado automático involucra los siguientes pasos:

- La validación de la existencia de los archivos `.phn`, los cuales contienen etiquetas generadas a nivel de fonema. Si existe algún archivo de etiquetas a nivel de

fonema el proceso se detiene y envía un mensaje de notificación. Esto impedirá que se sobre-escriban etiquetas anteriores. En este caso se reubican las etiquetas existentes para generar las nuevas.

- Después se generan categorías fonéticas usando el reconocedor, el vocabulario, el archivo descriptor, el texto de cada archivo a procesar y el archivo de la señal correspondiente. Con las transcripciones de la señal, se crea una gramática que sólo reconocerá esa secuencia de palabras, así la señal que se evalúa en la red corresponde a la secuencia correcta. La señal es procesada y dividida en marcos, de los cuales se obtienen vectores de características; estas características son evaluadas en la red y el resultado es la secuencia de categorías correcta y el tiempo en que ocurrieron. Con los tiempos de inicio y fin y las categorías conocidas, se escribe un archivo de transcripciones de categorías para cada archivo del corpus procesado.
- Finalmente, a partir de cada archivo de categorías, se genera un archivo de transcripciones fonéticas, que es el resultado buscado.

Resumiendo los pasos anteriores, dado un corpus con transcripciones a nivel de texto y utilizando un reconocedor, se pueden obtener transcripciones a nivel de fonema alineadas en tiempo. El proceso para generar estas transcripciones se basa en la automatización de los pasos que se siguen en el proceso de *forced alignment.*

4 Evaluación

El objetivo de esta sección es mostrar los resultados de desempeño y exactitud del sistema de alineación automática. Las pruebas y la evaluación se realizaron con el corpus test_digits

4.1 Características de los corpora de voz

`test_digits` consiste en grabaciones de series de dígitos vía telefónica. Se grabó usando el CSLU Toolkit, una Pentium Windows NT con tarjeta de teléfono Dialogic. La frecuencia de muestreo es de 8Khz y los archivos fueron guardados en formato RIFF. Un total de 20 participantes leyeron cada uno una lista de 20 series aleatorias de dígitos del 0 al 9, más el número 10. Cada serie puede estar formada de 2 a 9 dígitos.

4.2 Evaluación cuantitativa de las posiciones de las etiquetas

El primer experimento es evaluar las etiquetas generadas automáticamente comparándolas con las etiquetas manuales, las cuales se asumen como correctas.

Se cuentan con los siguientes conjuntos de etiquetas automáticas.

- Generadas utilizando el reconocedor de propósito general de español mexicano.
- Generadas utilizando el reconocedor de dígitos.

Para evaluar la diferencia entre las etiquetas automáticas y las manuales se obtiene un error global, calculando la diferencia promedio entre las fronteras de las etiquetas. Este error se calcula sumando el cuadrado de las diferencias entre las fronteras de inicio de cada fonema dividiendo esta suma entre el número total de etiquetas y obteniendo la raíz cuadrada de la división. Así, el error en los archivos evaluados se calculo usando la siguiente formula:

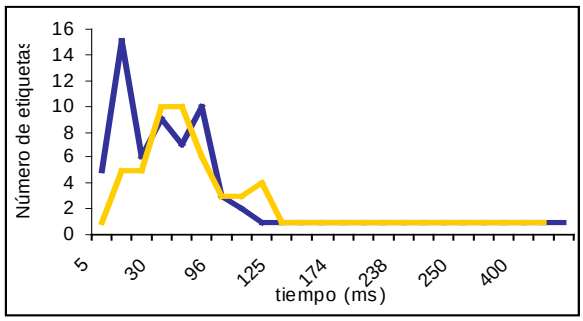
$$\sqrt{\frac{\Sigma(\text{manual} - \text{automático})^2}{N}}$$

Para hacer posible la evaluación de un error y la comparación entre las fronteras de las etiquetas, se seleccionaron sólo los archivos que tienen la misma secuencia de fonemas. Fueron utilizados 66 de 400 archivos del corpus etiquetado con el reconocedor de propósito general y 76 de 400 del corpus etiquetado con el reconocedor de dígitos. La tabla 1 muestra el error del total de archivos evaluados

Etiquetas automáticas del reconocedor de	Archivos Evaluados	Error total (milisegundos)
Propósito general	66/400	98.04
Dígitos	76/400	94.53

Tabla 1. Resultados de la evaluación del error de etiquetas

Otra comparación que se llevo a cabo fue la comparación de las posiciones de las etiquetas. La figura 5 resume los resultados de la posición en que fueron colocadas las fronteras con respecto al proceso manual. Se muestra el número de etiquetas y los milisegundos en que difieren con las etiquetas manuales.



Etiquetas generadas con el reconocedor de dígitos

Etiquetas generadas con el reconocedor de propósito general

Figura 5. Aquí se muestra una comparación entre las diferencias en milisegundos de las etiquetas automáticas y manuales para dos casos: utilizando el reconocedor de propósito general y utilizando el reconocedor de dígitos. Es evidente que las diferencias son menores para el último caso.

4.3 Evaluación del desempeño de los reconocedores

En este paso se evalúa el desempeño de los reconocedores entrenados con los corpora etiquetados automáticamente. Para evaluar su desempeño, se entrenó un reconocedor de dígitos para cada conjunto de etiquetas automáticas.

En la creación de un reconocedor los datos son divididos en datos para entrenamiento, desarrollo y prueba. Los datos de entrenamiento comprenden un 60% del total de archivos del corpus, los datos de desarrollo son el 20%

mientras que el 20% restante corresponde a los datos de prueba. Los datos de desarrollo son utilizados para determinar cuál fue el mejor reconocedor obtenido y en los datos prueba se evalúa el desempeño del mejor reconocedor. El desempeño del reconocedor se mide con el porcentaje de reconocimiento en palabras y series de dígitos de los datos de prueba.

Desempeño de los reconocedores de dígitos, creados con datos etiquetados con reconocedor		
Propósito general	Dígitos	Manualmente
93.98%	96.14%	93.01%

Tabla 2 *Resultados del desempeño de los reconocedores creados*

5 Conclusiones

En base a este trabajo se puede concluir que para etiquetar automáticamente un corpus, se necesita un reconocedor y el texto correspondiente a las señales de voz. Si el reconocedor usado es adecuado para los datos del corpus y tiene un buen desempeño, entonces la alineación de las etiquetas será más consistente que el proceso manual. Y una vez que tenemos un corpus etiquetado, puede usarse para crear un reconocedor de voz. Aún si el reconocedor usado para generar las etiquetas produce errores en el alineamiento, este proceso puede ser usado para determinar un conjunto inicial de datos etiquetados y usarlos para crear un reconocedor base y mejorarlo repitiendo el proceso anterior.

No existía un sistema como tal para español, ya que no se contaba con reconocedores de español con un buen desempeño y lo suficientemente completos que se pudieran usar en la alineación de limites fonéticos. La creación de los reconocedores usados dentro del grupo Tlatoa contribuyó para realizar la automatización del proceso de etiquetado.

Como resultado de este trabajo se creó la documentación del proceso automático, la cual se encuentra disponible en http://info.pue.udlap.mx/~sistemas/tlatoa/howto/labelfa.html

Los resultados obtenidos y presentados en la sección 4, son muy buenos, por lo que podemos concluir que sin invertir mucho tiempo en el proceso de etiquetado, se obtiene un buen desempeño en reconocimiento. Otra ventaja es que utilizando este método siempre será consistente la alineación de la señal. Estamos dejando que la red determine los límites de la manera en que los aprendió que puede ser diferente de la alineación realizada manualmente. La creación de una herramienta de software de este tipo ayuda a estudiantes e investigadores en el desarrollo de reconocedores en menor tiempo.

La determinación de limites fonéticos mejoraría si las condiciones en que fueron recolectados los datos que se usaron para entrenar el reconocedor son similares o iguales a las características de los datos que se desean etiquetar. Los factores tales como la frecuencia de muestreo, condiciones de ruido, el medio de captura de los datos (teléfono, micrófono) y el tipo de información grabada, determinan las características de un reconocedor y pueden afectar su desempeño.

Sin embargo, el proceso usado para determinar limites fonéticos (*forced alignment)* tiene varias limitaciones, (destacadas por John Paul Hosom en su propuesta de tesis de Doctorado [Hosom, 1998]). La localización de las fronteras podría mejorarse tomando en cuenta más información acústico – fonética. Por otro lado, contando con mayor información acústico – fonética de la señal, también se podrían tomar en cuenta más categorías que ahora no se conocen.

Agradecimientos

La preparación de este artículo y el trabajo de tesis fue lo grado gracias a la colaboración del grupo Tlatoa, de la Dra. Ingrid Kirschning y del M.C. Benjamin Serridge.

Referencias

[Cole et al, 1994] Cole, R. A., B. T. Oshika, M. Noel, T. Lander and M. Fanty, *Labeler Agreement in Phonetic Labeling of Continuous Speech,* Proceedings of the 1994 International Conference on Spoken Language Processing, Yokohama, Japan, Septiembre, 18-22, 1994.

[Cosi et al., 1998] Cosi, P., Hosom, J. P., Shalkwik, Sutton S, and Cole R.A. *Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM-Based Recognizer,* In Proceedings, 4th IEEE Workshop on Interqctive Voice Technology for Telecomunications Application, Turin, Italy, septiembre 1998.

[Hosom, 1998] Hosom, J. P., Thesis Proposal: Accurate Determination of Phonetic Boundaries Using Acoustic-Phonetic Information in Forced Alignment, Center for Spoken Language Understanding. Oregon Graduate Instituteof Science & Technology, Portland, USA, Augosto 20, 1998.

[Lander, 1996] Lander, T. *The CSLU Labeling Guide*. Center for Spoken Language Understanding. Oregon Graduate Instituteof Science & Technology, Portland, USA, Junio, 1996

[Rapp, 1995] Rapp, S. *Automatic Phonemic Transcription and Linguistic Annotation from Known Text with Hidden Markov Models / An Aligner for German* Proceedings of ELSNET goes East and IMACS Workshop, "Integration of Language and Speech in Academia and Industry", Moscow, 1995.

[Olivier, 1999] Olivier, A., *Tesis de Licenciatura: Evaluación de métodos de determinación automática de una transcripción fonética.*, Universidad de las Américas Puebla, Sta. Catarina Martir, Mayo de 1999.

[Ousterhout, 1997] Ousterhout, Jonh k., *Tcl and the Tk Toolkit*, Addison-Wesley Publishing Company, 8th Printing Julio, 1997.

[Serridge, 1997] Serridge, B., Master's Thesis: *Context-Dependent Modeling in a Segment-Based Speech Recognition System*., MIT, 1997.

[Zue, 1997] Zue, Victor. *Conversational Interfaces: Advanced and Challenges. Proc,* EuroSpeech 97, p. KN-18, Rhodes, Greece, Septiembre 1997.