Entrenamiento de una red neuronal, utilizando un corpus de voces infantiles

Autora: Ana María Hernández Zecuatl
Asesora Investigación: Mc Nancy Aguas García

6 de enero de 2009

Resumen

El presente trabajo describe de manera breve el proceso que se debe llevar a cabo para realizar un reconocedor de voz, en este caso en particular se habla sobre la realización de un reconocedor de voces infantiles. La creación del reconocedor de voces infantiles es necesaria para cubrir el módulo del lenguaje oral del sistema SAEL el cual está orientado a la enseñanza del lenguaje para niños.

1. Introducción

El Sistema de Apoyo a la Enseñanza del Lenguaje (SAEL) es un proyecto que propone el desarrollo de un sistema que sirva como apoyo en la enseñanza de niños con deficiencias auditivas y/o de lenguaje a través de la incorporación de un modelo de educación bilingüe basado en 2 módulos: uno de apoyo a la enseñanza del lenguaje oral y otro de apoyo a la enseñanza del lenguaje signado. SAEL utiliza estrategias y métodos de acuerdo a las necesidades de las escuelas y profesores de Educación Especial y hace uso de la tecnología de Reconocimiento de Voz.[Gar]

El módulo de enseñanza de lenguaje oral hace uso de un reconocedor de voz que fue entrenado con voces de jóvenes por lo que es inaplicable en el caso de los niños. Es por esto que se hace necesario contar con un reconocedor de voces infantiles para lograr que SAEL sea eficiente y funcional en la evaluación de las voces que recibirá.

Esta investigación propone llevar a cabo el desarrollo de un reconocedor de voces infantiles para lo cual será necesario realizar varias tareas tales como la definición del corpus, recolección de las voces, transcripción de las voces (etiquetado por palabra y fonemas), preparar los archivos para entrenar la red neuronal, entrenar la red neuronal y determinar la red de mejor desempeño; se persigue entrenar una red neuronal para optimizar la búsqueda que realizara el reconocedor de voz.

Los beneficiados directos son los niños con problemas del habla, ya que el módulo facilita el aprendizaje y el sistema será gratuito dado que se está trabajando con software libre.

El objetivo principal de esta investigación es entrenar una red neuronal mediante un corpus de voces infantiles; el corpus obtenido se podrá utilizar para otras interfaces, pero en primera instancia servirá para usarlo en el modulo oral de SAEL.

2. Herramientas a utilizar

Existen varias herramientas que ayudan a realizar el entrenamiento de las redes neuronales, así como el etiquetado, sin embargo en este proyecto se usa el CSLU toolkit debido a que con esta herramienta se desarrollo el SAEL, además de que el sistema ya está configurado para trabajar con los archivos específicos del CSLU. Por estas razonas se continuo trabajando con el CSLU toolkit.

2.1. ¿Qué es CSLU?

CSLU toolkit fue desarrollado en 1992, es un conjunto completo de herramientas para permitir la exploración, el aprendizaje y la investigación en la palabra y de la interacción humano-computadora.[Too]

2.2. ¿Cuales son las herramientas de CSLU?

CSLU contiene herramientas para Audio entre estas podemos encontrar: segmentado de audio, sincronización de entrada y salida de audio. Entre las herramientas de visualización ofrece: visualización de datos de ex-

presión, información de la etiqueta de voz, hacer click sobre las imágenes. En el reconocimiento del habla hace uso de los estándares HMM e híbridos HMM/enfoques ANN. CSLU también ofrece herramientas para el reconocimiento de voz; desarrollo rápido de la aplicación(RAD); herramientas PSL, estas herramientas facilitan la percepción de los experimentos humanos dentro de RAD (rapid application development), analizador robusto. Todas estas herramientas se pueden usar para la enseñanza de idiomas, educación, usos corporativos, investigación y desarrollo de un corpus de voz.[Too]

3. Definición del Corpus

Para definir el corpus se debe saber qué frases debe grabar cada persona, el número promedio de personas a grabar, el número de frases que cada persona grabará y bajo qué condiciones se deben realizar las grabaciones. El corpus que se eligió es un vocabulario fácil de leer y que contiene dígitos, letras, palabras comunes en la vida de los infantes; se eligieron los números y el abecedario debido a que es lo primero que los niños aprenden en la escuela. El dominio del corpus de voz a utilizar se encuentran desglosados en la tabla siguiente:

3.1. Población de locutores

El corpus debe ser grabado por niños de cuatro hasta los diez años, debido a que deben existir ejemplos de pronunciaciones correctas e incorrectas, de esta forma cuando se entrene la red neuronal con estos datos, se podrá tener una mejor clasificación de las frases correctas; al momento de usar el corpus en la interface no se restringirá al locutor a un sonido "perfecto"si no que se dará un margen de error, esto se debe a que se buscará igualar una voz que haya sido validada en la red neuronal o cuando menos la más cercana. El número de locutores que se plantea grabar es de 50 puesto que se necesitan muchos datos para entrenar un sistema de buena calidad. [Gar99]

3.2. Condiciones de Grabación

Las grabaciones deben realizarse bajo condiciones similares en el ambiente, el tipo de micrófono y las características de los locutores.[Gar99] Las grabaciones se llevan a cabo en un lugar cerrado (en este caso será un salón

	No. Frase	Frase a leer
Dígitos	1	1 2 3 4 5
	2	6 7 8 9 0
Letras	3	A B C D E F
	4	G H I J K L
	5	M N Ñ O P Q
	6	R S T U V W
	7	X Y Z
Palabras	8	Bicicleta, Balón, Patines
	9	Cometa, Columpio, Árbol
	10	Mamá, Papá, Tío,
	11	Hermano, Abuela, Bebé
	12	Comer, Tomar, Jugar
	13	Correr, Dormir, Repetir
	14	Gordo, Feo, Fuerte
Frases	15	Mi mamá se llama
	16	Tengo una casa grande
	17	Vamos a Jugar
	18	Llévame al parque
	19	¿Cuántos hermanos tienes?
	20	Me gusta mucho el helado

Tabla 1: Frases

de la institución donde se realizan las grabaciones), mediante el uso de una computadora personal y un micrófono de diadema con filtro de ruido.

4. Recolección del Corpus de Voz

Se realizaron las grabaciones con las frases mencionadas en la tabla 1.Todos los archivos se almacenan como XX.WAV. Antes de realizar las grabaciones se debe tener la siguiente organización de directorios:

- El corpus debe residir en **c:/data/corpora.**
- El archivo corpus debe estar en **c:/data/corpora/corpora.**

- 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 sonido **xx.wav.**
- Dentro de **transcriptions** hay sub-directorio para cada locutor,el cual contiene los archivos con las transcripciones a nivel de texto **xx.txt,** p**alabra xx.wrd**y fonema **xx.phn.**
- Los reconocedores reside en **c:/data/recognizers.**
- Los sub-directorios de **c:/data/recognizers** se nombrarán de acuerdo a lo que se reconoce**.**

[Gar99]

4.1. RAD usado para las grabaciones

El proceso de recolección de voces se realizo con ayuda de un RAD(Rapid Application Developer) un RAD es una herramienta que nos ayuda a construir aplicaciones eficientes. Se realizaron dos RAD´s uno para los niños que ya sabían leer y otro para los que aun no sabían leer. A continuación se muestra en la figura 1 el RAD utilizado.

En el siguiente párrafo se explica el funcionamiento del RAD mostrado en la imagen, dentro de los paréntesis encontraremos los nombres de los objetos. El RAD mostrado en la figura 1 nos permite crear los directorios de cada locutor (CREAR_DIR), dar la bienvenida al locutor(BIENVENIDA),realizar la grabación y escucharla(GRABAR), nos pregunta si la deseamos grabar otra vez(OTRA _VEZ), el objeto COMPROBAR comprueba la respuesta recibida del objeto OTRA_VEZ, continua comparando el numero de las frases si el numero es menor a 20 continua con las grabaciones(SIGUIENTE) si el numero es igual a 20 se despide del locutor y le da las gracias(DESPEDIR).

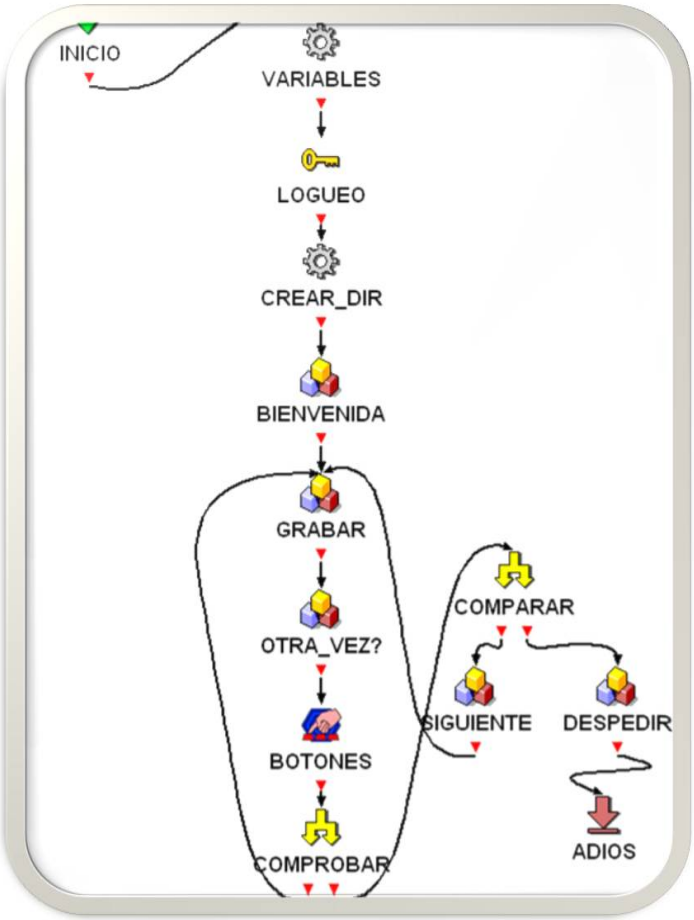

Figura 1: RAD de grabaciones

5. Etiquetado del corpus

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. De acuerdo a las especificaciones del CSLU, se manejan tres niveles de transcripciones:

- Transcripciones a nivel de texto
- Transcripciones a nivel de palabra
- Transcripciones a nivel de fonemas

Para este proyecto se utilizarán transcripciones a nivel de fonemas, ya que de esta forma se puede representar el contenido fonético de una pronunciación. Las unidades utilizadas para etiquetar el corpus de voz comprenden tres conjuntos de símbolos:

- Los usados para definir sonidos que pertenecen al lenguaje
- Los diacríticos que dan información detallada, a continuación se presentan los diacríticos a utilizar.
  - **_h** hace referencia a un aspirado
  - **_x** hace referencia a un fonema muy corto
  - **_fp** hace referencia a un llenado de pausa
  - **_lm** hace referencia a un ruido en la linea
  - **_bn** hace referencia a un ruido de fondo
- Los que definen sonidos que están fuera del lenguaje, entre estos símbolos encontramos los siguientes:
  - **.pau** Representa una pausa entre las palabras
  - **.br** Representa una respiración
  - **.bn** Representa un ruido de fondo en la grabación
  - **.ln** Representa un ruido en la linea
  - **.unk** Representa un elemento que no se ha podido identificar

[Gar99]

6. Entrenamiento de la red neuronal

En este caso se hace uso de los scripts que trae el CSLU toolkit, sin embargo se debe crear los archivos niños.parts y el niños.vocab. El archivo .vocab define dos cosas: la pronunciación de cada palabra en el vocabulario y la gramática. Consiste de dos partes, en la primera parte se definen las posibles pronunciaciones para cada palabra. El formato es el siguiente:

- W {D o bc b l e [u/w]}
- Balón {bc b a l o}

Del lado izquierdo se escribe la palabra del vocabulario a reconocer, y a su derecha se especifica su pronunciación a nivel fonético. La segunda parte del archivo describe la gramática que vamos a permitir durante el proceso de reconocimiento.

El archivo .parts define el número de partes en las que se va a dividir un fonema y las clases de fonemas que constituyen los contextos. Las opciones para dividir un fonema en partes son:

- Una (1) parte: el fonema es independiente del contexto
- Dos (2) partes: la primera mitad del fonema depende del contexto izquierdo y la segunda mitad del contexto derecho
- Tres (3)partes: la primera parte del fonema depende del contexto izquierdo, el centro es independiente del contexto y la ultima parte depende del contexto derecho.
- Right-dependent(r): el fonema no se divide en partes, pero si depende del contexto derecho.

[Gar99]

Una vez configurados los directorios y teniendo el etiquetado de las frases, se prosigue a realizar el entrenamiento de la red neuronal, para ello se hará uso de los siguientes scripts:

- **Gen_catfiles.tcl:** Crea las etiquetas alineadas con el tiempo de las categorías a entrenar. Estas categorías son escritas en archivos con la extensión .cat y se colocan en sub-directorios que reflejan la estructura del directorio del corpus a ser usada.
- **Genvec.tcl:** Se utiliza para crear el vector de características para cada frame que se entrenará. El proceso se lleva a cabo seleccionando los frames para entrenamiento, de los archivos creados con gen_catfiles.tcl, y procesa las características de todos los frames.

- **Make_recognizer.tcl:** Encuentra los archivos para entrenamiento y determina las categorías que serán clasificadas por el reconocedor.
- **Train_and_test_nets.tcl:** Entrena la red neuronal en el archivo vector name.train.vec. Este programa crea un archivo de pesos en cada iteración.

7. Conclusiones

Hasta el día de hoy se cuenta con la estructura de directorios necesaria; con las transcripciones a nivel fonema y a nivel palabra así como con los archivos ninos.parts y ninos.vocabs. Nos encontramos realizando la parte del entrenamiento de la red neuronal. El resultado final de esta investigación sera un reconocedor de voces infantiles.

Etiquetar las voces manualmente es un proceso lento, aunque existe la manera de hacerlo automáticamente ésta no es del todo optima debido a que una vez creadas las etiquetadas se requiere realizar la alineación de cada etiqueta con su respectiva parte de la señal.

CSLU Toolkit es una herramienta completa debido a que con ella se puede realizar el etiquetado de las voces, las grabaciones de las mismas y el entrenamiento de la red neuronal. En algunas cosas deja de ser eficiente por algunos bugs del sistema, CSLU proporciona los scripts necesarios para realizar el entrenamiento de la red sólo se deben tener los archivos necesarios y configurar algunas de los scripts.

Referencias

[Gar] Nancy Aguas Garcia. Proyecto sael.

[Gar99] Nancy Aguas García. Verificación de pronunciación basada en tecnología de reconocimiento de voz para un ambiente de aprendizaje, 1999.

[Too] CSLU Toolkit. http://cslu.cse.ogi.edu/toolkit/.