JOSE LUCIANO MALDONADO
Universidad De Los Andes
Mérida – Venezuela
maldonaj@faces.ula.ve
luzmalvy@telcel.net.ve

Construcción de modelos del habla venezolana para pruebas de reconocimiento automático de oraciones.

En este trabajo se estiman Modelos Ocultos de Markov para un conjunto de fonos del habla de hombres y mujeres de Venezuela, dichos fonos fueron extraídos de un corpus de entrenamiento constituido por 260 pronunciaciones de fechas obtenidas por vía telefónica fija. Los modelos estimados se utilizaron posteriormente para realizar pruebas de reconocimiento automático de un corpus de test de oraciones constituido por 72 fechas.

Las pruebas de construcción de los modelos y de reconocimiento se realizaron a través del HTK.

Para hacer el reconocimiento de las fechas se elaboró una gramática que comprende una red de palabras que admite la mayoría de las formas en que los venezolanos pronuncian las fechas.

Los resultados del reconocimiento fueron los siguientes: para las 72 fechas del corpus de test, el porcentaje de las palabras reconocidas fue 93.69%, mientras que el porcentaje de las oraciones fue 58.33%.

Posteriormente a esta prueba, se revisaron cada uno de los archivos de este corpus y sus etiquetas respectivas. Se determinó que 8 presentaban un alto nivel de ruido por lo que fueron descartados para una segunda prueba, mientras que otros presentaban algunos errores en el etiquetado que fueron corregidos.

Los resultados de esta segunda prueba mostraron un reconocimiento a nivel de palabras del 96.82% y a nivel de oraciones del 70.31%.

Estos resultados nos indican que los modelos de los fonos del habla venezolana fueron estimados de una manera adecuada y que es posible realizar tareas más generales si se sigue el procedimiento de este trabajo.

Justificación

En el campo de la tecnología del habla se vienen realizando esfuerzos en distintos centros de investigación de muchos paises con el objetivo de lograr una comunicación natural entre las máquinas y sus usuarios humanos. En ese sentido se han creado modelos de la voz de hombres, mujeres y niños, modelos de la voz pronunciada en distintos idiomas, modelos de la voz producida en distintos ambientes, modelos de la voz obtenida por distintos medios como micrófonos, teléfonos, etc.

Esos esfuerzos orientados a la construcción de los modelos de la voz han producido resultados exitosos, y es conocido que en la actualidad hay sistemas comerciales que cumplen actividades útiles donde se dan órdenes a las máquinas, se hace transcripción de texto, se hace consultas a algunos sistemas de información y hasta se establece algún tipo de dialogo con máquinas, todo por medio de la voz.

En la búsqueda de sistemas de reconocimiento robustos que reconozcan la voz de cualquier persona, en cualquier ambiente y en cualquier idioma se han utilizado diversos tipos de modelos de voz como son: modelos de palabras, modelos de sílabas, modelos de fonemas, modelos de trifonemas, modelos de semifonemas, etc.

En el caso de los esfuerzos que se están realizando en la Universidad de Los Andes en Venezuela, se han construido modelos de palabras y de fonemas para reconocer en forma automática elementos del español hablado por los venezolanos. Estos modelos fueron construidos en base a la Teoría de los Modelos Ocultos de Markov y las pruebas de reconocimiento se realizaron por medio del sistema de Desarrollo HTK [4].

Base de Datos

Se trabajó con archivos de voz tomados de la base de datos SpeechDat venezolana [2]. La SpeechDat Venezolana es una base de datos obtenida vía telefónica, específicamente a través de teléfonos fijos y construida bajo el formato SPEECHDAT [3]; es propiedad de la Universidad Politécnica de Cataluña, España, ya que fue obtenida como parte del Proyecto SALA de dicha Universidad, cuyo objetivo es construir una gran Base de Datos del Español hablado en el mundo y ponerla al servicio de los desarrolladores de sistemas de reconocimiento de voz para que incorporen ese lenguaje a tales sistemas.

La SpeechDat venezolana contiene grabaciones de voz de 1040 personas de las distintas regiones del País. De cada persona se obtuvieron 44 archivos, por lo que la base de datos completa consta de 45760 archivos, material suficiente para realizar estudios a nivel de ingeniería y a nivel lingüístico con el dialecto venezolano.

El contenido de la SpeechDat Venezolana comprende diversas categorías de pronunciaciones, entre las que encontramos: pronunciaciones de palabras aisladas, como dígitos, nombres, pronunciaciones de palabras conectadas como números telefónicos, números de tarjetas de crédito, pronunciaciones de frases y oraciones mas complejas gramaticalmente como fechas, horas, etc.,[2].

Para las pruebas de reconocimiento realizadas en este trabajo se seleccionaron pronunciaciones de fechas como las oraciones a ser reconocidas. Se trabajó con 332 archivos de pronunciaciones de fechas de hombres y mujeres de toda Venezuela. De los 332 archivos que se utilizaron en las pruebas, 260 fueron tomados como Corpus de Entrenamiento y 72 como Corpus de Reconocimiento.

Las pronunciaciones del Corpus de Entrenamiento estaban distribuidas de la siguiente manera: 140 pronunciaciones de mujeres y 120 de hombres.

Las pronunciaciones del Corpus de Reconocimiento estaban distribuidas de la siguiente manera: 38 pronunciaciones de mujeres y 34 pronunciaciones de hombres.

Las edades de las personas cuyas voces intervienen en las pruebas oscilan entre los 18 y los 60 años y con un grado de instrucción que alcanza la educación media.

Ubicación y Preparación de los Archivos de Voz

La forma como se utilizaron los archivos de voz fue la siguiente: se localizaron dentro de la base de datos SpeechDat venezolana, un conjunto de pronunciaciones de fechas. Se realizó la transcripción fonética de cada pronunciación siguiendo el formato SAMPA. Esto es lo que se ha llamado etiquetado de la voz y consistió en asociar cada sonido distinguible por el oído, con un símbolo fonético del lenguaje español venezolano.

Posterior a ese tratamiento, cada archivo fue convertido del formato propio de la tarjeta de telefonía con que se realizaron las grabaciones a formato WAVE, con el fin de que el HTK [4] los pudiera procesar.

El último paso necesario para iniciar el modelado y la construcción del reconocedor consistió en la parametrización de las señales, para las que se obtuvieron los coeficientes cepstrales en escala de frecuencia Mel (Mel Cepstrum ó MFCC), la energía, la primera y segunda derivada, sobre segmentos de 25 milisegundos, enventanados a través de una ventana Hamming y desplazados 10 milisegundos; por lo que cada segmento se solapaba con el anterior en 15 milisegundos, además cada señal se pasaba por un filtro de pre-énfasis cuyo coeficiente tenía el valor 0.97. Se trabajó con 38 parámetros por segmento.

Sistema de reconocimiento desarrollado

La construcción de los modelos y todas las pruebas de reconocimiento fueron realizadas con el HTK [4]. HTK es un paquete de programas que permite realizar todas las fases de procesamiento de señales para crear sistemas de reconocimiento de voz con Modelos Ocultos de Markov [1][6].

Modelos del habla venezolana utilizados en las pruebas

Los modelos del habla que se construyeron para el reconocimiento de las pronunciaciones de fechas están constituidos por fonos.

En los archivos utilizados en las pruebas se encontró el siguiente conjunto de fonos Venezolanos: a, b, B, c, d, D, e, f, g, G, h, i, j, k, l, m, n, N, M, o, p, r, R, s, t, u, w, y y sil. Es decir, no se trabajó con todos los fonos del habla venezolana, sino que se trabajó con los fonos que se encontraron en los archivos de los corpus de entrenamiento y de reconocimiento.

Se puede apreciar que no aparece la v; ésto se justifica debido a que en general, en la pronunciación venezolana no se distingue la ocurrencia de una v de una b. Lo mismo sucede con la s y la z. Tampoco aparece la x, puesto que es poco frecuente el uso de tal fonema en su conversación y menos aun, en pronunciaciones de fechas. El símbolo c se seleccionó para representar el sonido de la ch y M para representar el sonido de la ñ. B, D y G representan la realización oclusiva (inicio de frase después de una nasal) de los fonemas /b/, /d/ y /g/, mientras que b, d y g representan la realización fricativa de los mismos fonemas. N representa la realización velar del fonema /n/ en distensión, es decir, al final de sílaba seguida por consonante.

Se utilizó el símbolo sil para representar las zonas de silencio presentes en las pronunciaciones. El resto de símbolos representan sonidos bien definidos e identificados en los archivos.

Para cada uno de estos sonidos se crearon Modelos Ocultos de Markov. Los modelos de estos fonos fueron creados a partir de las 260 pronunciaciones del corpus de entrenamiento y de las realizaciones de cada sonido en dicho corpus.

Los Modelos Ocultos de Markov que representan los fonos son del tipo Bakis [1] de 5 estados y se estimaron por medio del algoritmo Baum-Welch [1].

Inicialmente los modelos de los fonos son clones de un modelo prototipo que se creó con los vectores de medias y covarianzas globales obtenidos a partir de todos los parámetros extraídos de las señales del corpus de entrenamiento. Estos modelos inicialmente contenían una gausiana por estado.

Una vez que se obtuvieron esos modelos iniciales, el modelo de cada fono fue re-estimando progresivamente con sus propias realizaciones hasta encontrar los mejores parámetros para cada modelo.

El proceso de búsqueda de los mejores parámetros de los modelos consistió en hacer re-estimaciones y en posteriormente observar la capacidad de reconocimiento con el corpus de reconocimiento.

A continuación se describe el proceso de re-estimación de los modelos con los cuales se obtuvieron los mejores resultados:

A partir de los modelos iniciales descritos, se realizaron 3 re-estimaciones sucesivas a través de la versión “embedded training” del algoritmo de Baum-Welch [4]. Luego, se fue aumentando el número de gausianas por estado hasta llegar a seis gausianas. Cada vez que se aumentaba el número de gausianas se efectuaban dos re-estimaciones del tipo indicado antes, a excepción del último caso donde con seis gausianas por estado se realizaron entre 1 a 27 re-estimaciones más.

Pruebas de reconocimiento de las oraciones

Las pruebas de reconocimiento consistieron en presentarle, al sistema de reconocimiento, como entradas pronunciaciones de fechas y en observar la capacidad que tenía dicho sistema para identificar esas fechas.

Se realizaron en forma separada pruebas donde se utilizaron archivos que formaban parte del corpus de entrenamiento, y pruebas con archivos que formaban parte del corpus de reconocimiento. Los archivos para estas últimas pruebas fueron tomados de personas de lugares de origen distinto al origen de las personas cuyas voces intervinieron en el entrenamiento.

Todas las pruebas de reconocimiento se realizaron con el algorítmo Viterbi del HTK [1].

Diccionario utilizado en las pruebas

Como se mostró en la sección anterior, el sistema de reconocimiento está dotado de 29 modelos de sonidos del habla venezolana, lo que significa que dada una pronunciación como señal de entrada, debe ser capaz de identificar la secuencia de esos sonidos presentes en dicha pronunciación y hacer la transcripción ortográfica de dicha secuencia.

Debido a que en este trabajo nos planteamos como objetivo hacer reconocimiento de palabras y oraciones, además del reconocimiento de fonos, dotamos al reconocedor de un diccionario cuyas entradas consistían de secuencias de los símbolos asociados a los sonidos (secuencia de fonos) y las salidas eran las palabras ortográficas asociadas a esas secuencias de sonidos.

El diccionario contiene entonces, la lista de las palabras que intervienen en las pronunciaciones de las fechas y los sonidos asociados a esas palabras.

En el diccionario se puede encontrar una misma palabra asociada con secuencias diferentes de sonidos, esto se debe a que se trató de cubrir para el reconocimiento, las diferentes variantes de pronunciación de las palabras que se detectaron en los archivos bajo estudio.

Para ilustrar esta situación, presentamos como ejemplo la palabra diciembre, que puede estar constituida por la secuencias de sonidos siguientes: disjembre, Disjembre, disjemBre ó DisjemBre.

Esto significa que como las personas pronuncian las mismas palabras de manera diferente, entonces hay que buscar una forma adecuada para asociar diferentes secuencias de sonidos a una misma palabra. Esto se logra construyendo el diccionario en la manera descrita.

Por razones de espacio, no se muestra el diccionario completo, en su lugar se presentan a continuación cuatro ejemplos de su salidas:

abril a b r i l
agosto a g o s t o
agosto a g o h t o
año a M o
catorce k a t o r s e

Gramática utilizada para el reconocimientos de las oraciones

Así como a través del diccionario se puede identificar la secuencia de palabras presente en una pronunciación, para identificar oraciones hay que dotar a los reconocedores con las reglas gramaticales del lenguaje al cual están asociadas las pronunciaciones que se pretenda que éstos manejen. En este caso, se incorporó al reconocedor una gramática que consiste en la combinación adecuada de todas las palabras que usan los venezolanos para pronunciar o escribir fechas.

El programa que modela la gramática permite aceptar fechas de las formas que mostramos a continuación como ejemplo:

“El martes cuatro de diciembre de mil novecientos treinta y cuatro”, “martes cuatro de diciembre de mil novecientos treinta y cuatro”, “cuatro de diciembre de mil novecientos treinta y cuatro”, “diciembre de mil novecientos treinta y cuatro”, “cuatro doce de mil novecientos treinta y cuatro”, “cuatro doce treinta y cuatro”, “el cuatro doce de mil novecientos treinta y cuatro”, “en el mes de diciembre de mil novecientos treinta y cuatro”, “en diciembre del treinta y cuatro” y algunas otras.

A continuación se muestra la programación de esa gramática, la cual está escrita en el formato HTK.

$num1 = cero| uno | dos | tres | cuatro | cinco | seis | siete | ocho | nueve;
$num2 = diez | once | doce | trece | catorce | quince | dieciseis | diecisiete | dieciocho | diecinueve | veinte | veintiuno | veintidos | veintitres | veinticuatro | veinticinco | veintiseis | veintisiete | veintiocho | veintinueve | treinta;

$num3 = cuarenta | cincuenta | sesenta | setenta | ochenta | noventa;

$fin2 = y $num1;
$fin1 = $num1 | $num2 | $num3 $fin2 | $num3;
$nmes = $num1 | diez | once | doce;
$dia = lunes | martes | miércoles| jueves | viernes | sábado | domingo;
$mes = enero | febrero | marzo | abril | mayo | junio | julio | agosto | septiembre | octubre | noviembre | diciembre;
$delanodosmil = del año dos mil;
$deldosmil = del dos mil;
$delanomil = del año mil;
$demil = de mil | del mil;
$dosmil = $delanodosmil | $deldosmil;
$ano = $delanomil | $demil;
$demes = de $mes| $mes;
$eldia = el $dia | $dia;
$enmes = en $mes;
$num = $num1| $num2;
$elnum = el $num | $num;
$parte4 = $eldia $num | el $num | $num;
$parte1 = $parte4 $demes | $enmes | $elnum $nmes;
$parte2 = $ano | $dosmil;
$parte3 = $fin1;
$parte5 = $parte2 novecientos | $parte2;
$parte6 = $parte5 $parte3 | del $parte3;
$parte7 = $parte4 $demes;
(SIL ($parte7 | $parte1 [SIL] $parte6) SIL)

Resultados obtenidos

Resultados más favorables obtenidos en las pruebas de reconocimiento de fechas donde se trabaja con archivos del corpus de entrenamiento.

Se realizaron varias pruebas con este corpus: inicialmente se hizo reconocimiento de los 260 archivos, en ese caso el sistema lograba reconocer completamente el 40% de las fechas y el 80.57% de las palabras que formaban esas fechas. Posterior a esta prueba inicial, se descartaron los archivos que contenían palabras que normalmente no se usan en las pronunciaciones de fechas, los archivos que presentaban altos niveles de ruido ambiental y se corrigieron aquellos que estaban acompañados de errores en las transcripciones.

Los resultados después de la revisión, se muestran de manera resumida en la siguiente salida propia del HTK.

---------- Overall Results ----------
SENT: %Correct=69.44 [H=150, S=66, N=216]
WORD: %Corr=96.94, Acc=94.96 [H=1963, D=15, S=47, I=40, N=2025]
==========

Resultados más favorables obtenidos en las pruebas de reconocimiento de fechas donde se trabaja con archivos del corpus de reconocimiento.

Igual que en las pruebas anteriores, se seleccionaron archivos que presentaban menos niveles de ruido ambiental y con palabras propias de las fechas. A continuación se muestran los resultados que constituyen los rangos de reconocimiento, a nivel de fechas completas y a nivel de las palabras.

---------- Overall Results ----------
SENT: %Correct=58.33 [H=42, S=30, N=72]
WORD: %Corr=93.69, Acc=88.77 [H=609, D=13, S=28, I=32, N=650]
==========

---------- Overall Results ----------
SENT: %Correct=70.31 [H=45, S=19, N=64]
WORD: %Corr=96.82, Acc=94.14 [H=578, D=7, S=12, I=16, N=597]
==========

Discusión de Resultados

Los resultados alcanzados son satisfactorios, especialmente si se toma en cuenta que no sólo se estaba haciendo reconocimiento a nivel de fonos y de palabras, sino también de oraciones, y que para lograr este último tipo de reconocimiento se requieren buenos resultados a nivel de las unidades previas, y eso sólo se logra cuando los modelos están bien entrenados. Otro aspecto a considerar a la hora de evaluar los resultados, es el hecho de que se trabajó con voces de múltiples hablantes es decir se hizo reconocimiento de oraciones de manera independiente del locutor lo que le da amplitud al sistema reconocedor, puesto que no es esclavo de la voz de una sola persona.

Estos resultados indican que es posible crear modelos del habla de los venezolanos, a nivel de fonos para hacer reconocimiento automático de oraciones y que en el futuro se podrán construir reconocedores del español hablado en Venezuela, siempre que se trabaje con una gran cantidad de pronunciaciones de entrenamiento por modelo para mejora la capacidad de reconocimiento obtenida en este trabajo.

Referencias Bibliográficas

[1] Deller J., Proakis J. y Jansen J., Discrete-Time Processing of Speech Signals. New York : Macmilllan Publishing Company 1993.

[2] Moreno A., Mora E., Speechdat Spanish Venezuelan database for the fixed Telephone network, Universidad politécnica de Cataluña, España y Universidad de Los Andes, Venezuela. 1999.

[3] Moreno A., Speechdat Spanish Database for the fixed Telephone Network. Universidad Politécnica de Cataluña, España. 1997.

[4] Entropic Speech Technology, HTK Manual.

[5] Savage J., A hybrid System with Symbolic AI and Statistical Methods for Speech Recognition. A dissertation for the degree of Doctor of Philosophy, University of Washington, USA 1995.

[6] Zhao Y., A Speaker-Independent Continuous Speech Recognition System using Continuous mixture Gaussian Density HMM of phoneme-sized units. IEEE Transaction on Speech and Audio Processing, Vol. 1, No. 3, July (1993).

[7] Hoge H., Moreno A., Technical Annex of the SALA Contract. SALA project feb. 1999.