CP 10
PON 19

Creación de un diccionario de pronunciación de nombres propios para uso en tecnologías del habla

Carlos Daniel Hernández Mena
ca_hernandez@uxmcc2.iimas.unam.mx

José Abel Herrera Camacho
abelhc@hotmail.com

Laboratorio de Procesamiento de Voz, Facultad de Ingeniería de la UNAM.

RESUMEN

En este artículo se presenta la creación de un diccionario de pronunciación de nombres propios que servirá como herramienta en sistemas orientados a tecnologías del habla, principalmente a sistemas de reconocimiento automático de voz y sistemas de síntesis de voz. Este diccionario cuenta con múltiples pronunciaciones de nombres de pila y apellidos de México, España y Costa Rica, así como nombres Geográficos de toda la República Mexicana con un total de 180,536 nombres propios sin repeticiones.

1. INTRODUCCIÓN

En la actualidad se han alcanzado grandes progresos en el campo del reconocimiento automático de voz, así como en todas las tecnologías del habla en general. Cada vez más empresas están dispuestas a invertir por sistemas que tengan que ver con estas tecnologías; ofreciéndonos ahora, más que nunca, control por voz en automóviles, en teléfonos celulares, en casas inteligentes, en sistemas de operadora automática, en televisiones que cambian de canal o sistemas GPS que nos guían por medio de comandos hablados entre muchos otros.

Pero a pesar de estos significativos avances sigue siendo un problema real enfrentarse tanto al reconocimiento como a la síntesis de palabras con múltiples pronunciaciones, baja frecuencia de aparición ó palabras doblemente confusas, es decir, aquellas que fonéticamente son similares y además tienen la misma probabilidad de aparecer en una oración determinada **[GOLDWATER ET AL, 2010]**.

ROC&C'2013 / CP 10 PONENCIA RECOMENDADA POR EL COMITE DE COMPUTACION DEL IEEE SECCION MEXICO Y PRESENTADA EN LA REUNION INTERNACIONAL DE OTOÑO, ROC&C'2013. ACAPULCO, GRO., DEL 10 AL 14 DE NOVIEMBRE DEL 2013.

Un ejemplo son precisamente los nombres propios, ya que a veces, dependiendo de diversos factores como la ubicación geográfica, el nivel socioeconómico o incluso ciertas modas, pueden tener una o varias pronunciaciones distintas. Por ejemplo, el nombre ***"michel"*** puede pronunciarse como *"michel"* con la *"ch"*, como */maikol/* o incluso como *"mishel"*. Además de esto, es muy probable que dos nombres propios que son fonéticamente similares, tengan también la misma probabilidad de aparición, como en las oraciones:

1. *"Corre y llama a tu hermana **Ailyn**"*
2. *"Corre y llama a tu hermana **Eilyn**"*

Aquí es claro que los nombres ***Ailyn*** y ***Eilyn*** son fonéticamente similares y podrían tener la misma probabilidad de aparecer en una oración como esta.

Es por estos problemas que nos dimos a la tarea de crear un diccionario de pronunciación a gran escala de nombres propios. Este documento contiene un total de 180,536 nombres y apellidos diferentes (es decir, sin que se repita ninguno) de México, España y Costa Rica, así como los nombres geográficos de México.

Cada uno de los nombres esta trascrito en un alfabeto fonético llamado MEXBET en su nivel T-22 como el utilizado en **[PINEDA ET AL, 2004 Y 2009],** con la única modificación de que para representar la grafía ***"ñ"*** se usa el símbolo ***/N/*** en vez del símbolo ***/n~/*** del MEXBET original y con la adición de ciertos símbolos para los diferentes contextos de la letra *"x"* que se explicarán mas adelante en este mismo documento. Otro aspecto importante es que en este diccionario las pronunciaciones son en español del centro de México aunque algunos nombres presentan también múltiples pronunciaciones cuando la manera en que están escritos se presta para ello, como es el caso del

nombre: ***"Jazmín"*** que puede leerse como *"Yasmín"*, o incluso como *"Jasmín"* con la *"j"*.

2. OBTENCIÓN DE LOS NOMBRES PROPIOS.

Acudimos a diversas instituciones gubernamentales y a nuestra propia universidad para obtener la mayor cantidad de nombres reales posible. Estas instituciones fueron:

- Instituto Federal Electoral Mexicano (IFE)[1]
- Instituto Nacional de Estadística de España (INE)[2]
- Tribunal Supremo de Elecciones (TSE) de la República de Costa Rica[3]
- Facultad de Ingeniería de la Universidad Nacional Autónoma de México (FI-UNAM)[4]
- Posgrado de Ingeniería de la Universidad Nacional Autónoma de México (PI-UNAM)[5]
- Secretaría de la Reforma Agraria (SRA) de México.[6]
- Instituto Nacional de Estadística, Geografía e Informática Mexicano (INEGI)[7]

Lo que obtuvimos del IFE y del INE fueron listas de frecuencias de nombres y apellidos con una frecuencia igual o mayor a 20, del TSE obtuvimos el padrón electoral de Costa Rica con los nombres completos (es decir, nombres de pila y apellidos) de los empadronados, de la FI-UNAM, del PI-UNAM y de la SRA obtuvimos listas de nombres y apellidos mezclados de forma aleatoria para evitar cualquier identificación de alguna persona que figurara en esas listas y finalmente del INEGI obtuvimos una lista de los nombres geográficos de México (más información sobre los nombres geográficos del INEGI en **[MUÑOZ, 2005]**).

La tabla 1 muestra la cantidad de nombres propios obtenida de cada institución por separado.

[1] http://www.ife.org.mx/portal/site/ifev2

[2] http://www.ine.es/

[3] http://www.tse.go.cr/index.html

[4] http://www.ingenieria.unam.mx/

[5] http://ingenieria.posgrado.unam.mx/sitv3/

[6] http://www.sra.gob.mx/sraweb/

[7] http://www.inegi.org.mx/

3. METODOLOGÍA DE CREACIÓN

Una vez obtenidas las listas de nombres propios de cada una de las instituciones, se utilizaron scripts de Python creados por nosotros mismos, para eliminar signos de puntuación como guiones, puntos, comillas, etc. Luego mezclamos todos los nombres propios en una sola lista, eliminamos las repeticiones y los ordenamos alfabéticamente. El resultado de este proceso produjo una sola lista ordenada de nombres propios sin repeticiones.

Institución	Cantidad de Nombres
IFE	48,050
INE	31,111
TSE	100,557
FI-UNAM	4,868
PI-UNAM	4,118
SRA	1,118
INEGI	44,288
Suma Total	**234,110**
Nombres Repetidos	**53,574**
Nombres Sin Repetir	**180,536**

Tabla 1. Cantidad de Nombres por Institución

Ya con estas listas, se les pidió a nuestros transcriptores que aplicaran las siguientes reglas sencillas de trascripción "semi-fonológicas" a cada nombre propio:

- Marcar la vocal tónica con una mayúscula. Ejemplo: ***margarIta***, ***renE***. En palabras donde solo hay una vocal se pone esa vocal en mayúscula como *"A"*, *"E"* u *"O"*. Si la vocal tónica va acentuada hay que ignorar el acento y poner esa vocal en mayúscula como en ***marIa***, ***suAres***, etc.
- La *"ñ"* o *"Ñ"* debe ser sustituida por *"N"* como en: ***NONo***, en vez de ***ñOño***, o ***sUNiga*** en vez de ***sUñiga***.
- No usar la *"y"* y en su lugar usar solo la *"ll"* o la *"i"*. (ejem: ***rei*** en vez de ***rey***, ***nallEli*** en vez de ***nayEli***, etc).
- No usar la *"c"* ni *"q"* y en su lugar usar la *"k"*. (Ejem. ***kArlos*** en vez de ***cArlos***, ***barAk*** en vez de ***barAq***)
- No usar la *"z"* sino solo la *"s"*. (Ejem. ***sOrro*** en vez de ***zOrro***, ***sOna*** en vez de ***zOna***)
- No usar la *"v"* sino solo la *"b"* (Ejem. ***bIno*** en vez de ***vIno***, ***bAca*** en vez de ***vAca***)
- No usar jamás la *"w"* (Ejem. ***gualter*** y no ***walter***). Y si se permite el uso de la diéresis como en *güe*, *güi*.

- La doble *"z"* como en ***pizza*** se transcribe con *"ts"* (***pitsa***).
- Transcribir los nombres de la forma más simple que se pueda, es decir que no se deben usar dobles consonantes por ejemplo ***wAtt***, sería ***guAt*** o ***clarIssa*** sería ***clarIsa***, etc.

Ahora bien, también se les pidió que especificaran los contextos de la letra *"x"* mediante las siguientes reglas:

- ***Xochimilco*** se transcribe ***$ochimIlco*** por que aquí la *"x"* suena como *"s"*.
- ***exámen*** se transcribe ***eKSAmen*** por que la *"x"* suena como *"KS"*.
- ***Xolos*** se transcribe ***SOlos*** por que la *"x"* suena como el fonema *"esh"* que se escribe como *"S"* y que aquí sonaría como ***"SHOLOS"***.
- ***México*** se transcribe ***mEJico*** por que la *"x"* suena como *"J"*.
- Sustituir la *"sh"* como en ***Sharon*** por la *esh "S"* que es la misma *"S"* que en ***SOlos*** (***xolos***) o ***shicotEncatl*** (***xicotencatl*** que debe transcribirse como ***Sicotencatl***).

Teniendo estas reglas en mente, una trascripción correcta sería:

zussie sUsi
zussy sUsi
zusy sUsi
zutacota sutacOta
zutlay sutlAi

Es decir, el nombre propio, seguido de un espacio en blanco, seguido de la misma palabra transcrita con las reglas de trascripción descritas.

Finalmente también se les pidió a los transcriptores que si reconocían más de una pronunciación del mismo nombre debería transcribirse de la siguiente manera:

michel michEl
michel(1) maIkol
david dabId
david(1) delbid
daniel daniEl
daniel(1) dAniel
daniel(2) delniel

Donde se aprecia que cada pronunciación adicional es añadida seguida de un número entre paréntesis que indica el número de pronunciación extra de la palabra.

Cuando todo este procedimiento de aplicar las reglas de transcripción ha sido concluido aplicamos nuestra herramienta de software FONETIZADOR para convertir las transcripciones de las palabras en transcripciones en alfabeto MEXBET nivel T22.

En la figura 1 se muestra el alfabeto MEXBET T22 utilizado y su equivalente en el alfabeto fonético internacional (IPA).

IPA	Mexbet	IPA	Mexbet
a	a	s	s
e	e	ʃ	S
o	o	x	x
i	i	j	Z
u	u	m	m
p	p	n	n
t	t	ɾ	r(
k	k	r	r
b	b	ɲ	N
d̪	d	l	l
g	g	f	f
t͡ʃ	tS		

Figura 1. Alfabeto MEXBET utilizado y su equivalente en IPA

4. APLICACIONES

Existen varios usos que se le podrían a un diccionario de nombres propios como el nuestro, por ejemplo en el campo de la síntesis de voz se puede contribuir a mejorar la pronunciación de nombres de lugares y esto puede ser bastante deseable en sistemas GPS que guían por medio de la voz.

Es indudable también que en el campo del reconocimiento automático de voz se puede contribuir a disminuir el "word error rate" (WER) ya que se podrá verificar una palabra entrante con múltiples pronunciaciones de un mismo nombre o incluso verificar varias formas de escribir un nombre que de hecho suenen igual como: ***Christian***, ***Cristian*** y ***Kristian***.

En la literatura también hay ejemplos de personas que han investigado con modelos de lenguaje para nombres propios como en **[TAPIA ET AL, 2010]** o que han tenido la necesidad real de contender con nombres en un sistema de operadora automática como en **[VARELA ET AL, 2003].**

Ciertamente se espera que con nuestra herramienta puedan profundizarse investigaciones como las mencionadas en el párrafo anterior o incluso generarse otras totalmente nuevas.

5. CONCLUSIONES

Un diccionario de pronunciación de nombres propios es una herramienta que puede ayudar a mejorar sistemas de reconocimiento y síntesis de voz de manera directa y sencilla, ya que su aplicación no implicaría grandes cambios en la arquitectura de sistemas como estos.

Por otro lado la idea de crear diccionarios de pronunciación no es nueva en lo absoluto y ha demostrado ser útil en muchos casos (por ejemplo, cuando queremos saber la pronunciación de palabras en un idioma extranjero).

Algunos ejemplos de diccionarios de pronunciación son "The CMU Pronouncing Dictionary"[8] de la Universidad Carnegie Mellon, El proyecto MOBY[9] creado por Grady Ward, El "British English Example Pronciation dictionary"[10] (BEEP) creado por la universidad de Cambridge.

Sin embargo, todos esos diccionarios de pronunciación están en el idioma inglés y contienen pronunciaciones de palabras en general y no solo de nombres propios, por lo tanto, consideramos que este trabajo es en cierto modo pionero, al menos entre los países de habla hispana.

AGRADECIMIENTOS

Agradecemos al Instituto Federal Electoral (IFE) de México por brindarnos una enorme cantidad de nombres por medio de su sitio INFOMEX[11], al licenciado Héctor Eduardo Aguayo Muñoz del Instituto Nacional de Estadística, Geografía e Informática (INEGI) por atender nuestra solicitud de los nombres geográficos de México, al Ing. Gonzalo López de Haro (secretario general de la Facultad de Ingeniería de la UNAM) y al Dr. Luís A. Álvarez-Icaza Longoria (Coordinador del Posgrado de Ingeniería de la UNAM) por proporcionarnos nombres y apellidos reales (mezclados aleatoriamente) de estudiantes de la Facultad de Ingeniería y del Posgrado de Ingeniería respectivamente, a la lic. Sara Molina Guzmán por proporcionarnos nombres y apellidos reales (mezclados aleatoriamente) de la Secretaría de la Reforma Agraria (SRA) de México, y finalmente a todos nuestros prestadores de servicio social que hicieron posible este diccionario gracias a su arduo trabajo.

[8] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[9] http://icon.shef.ac.uk/Moby/

[10] http://www.speech.cs.cmu.edu/comp.speech/Section1/Lexical/beep.html

[11] https://ciudadania.ife.org.mx/infomex/ActionInitSAILoginINFOMEX.do

REFERENCIAS

[GOLDWATER ET AL, 2010] Sharon Goldwater, Dan Jurafsky, Christopher D. Manning, Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates, Speech Communication, Volume 52, Issue 3, March 2010, Pages 181-200, ISSN 0167-6393, DOI: 10.1016/j.specom.2009.10.001.

[MUÑOZ, 2005] Aguayo Muñoz, Héctor E.; "SISTEMA DE CONSULTA DEL REGISTRO DE NOMBRES GEOGRÁFICOS"; Reporte Técnico; Instituto Nacional de Estadística, Geografía e Informática; 2005; pp. 5;

[PINEDA ET AL, 2009] L. A. Pineda, H. Castellanos, J. Cutara, L. Galescu, J. Jurez, J. Llisterri, P. Prez-Pavn, L. Villaseor, "The Corpus DIMEx100: Transcription and Evaluation", Language Resources and Evaluation, 2009.

[PINEDA ET AL, 2004] Pineda L., Villaseñor-Pineda L., Cuétara J., Castellanos H. and López I. "DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish", In Proceedings of the IX IberoAmerican Conference on Artificial Intelligence (IBERAMIA 2004), pages 974-983. Puebla, Mexico, 2004.

[TAPIA ET AL, 2010] Tapia, G., Meza, I. and Pineda, L.: "Language Models for Name Recognition in Spanish Spoken Dialogue System". Research in Computer Science, Vol. 49, pp. 145-154, 2010.

[VARELA ET AL, 2003] A. Varela, H. Cuayáhuitl and J.A. Nolazco-Flores, "Creating a Mexican Spanish Version of the CMU Sphinx-III Speech Recognition System", Progress in Pattern Recognition, Speech and Image Analysis, 251-258, Springer 2003.

SOBRE LOS AUTORES

Carlos Daniel Hernández Mena nació en la ciudad de México en 1983. Es Ingeniero en Comunicaciones y Electrónica por parte de la Escuela Superior de Ingeniería Mecánica y Eléctrica unidad Culhuacán (ESIME Culhuacán) del Instituto Politécnico Nacional (IPN). Estudió la maestría en Ingeniería en Computación en el Posgrado de Ciencias e Ingeniería de la Computación con sede en el Instituto de Investigación de Matemáticas Aplicadas y Sistemas (IIMAS) de la Universidad Nacional Autónoma de México (UNAM). En esta maestría trabajó haciendo reconocimiento de comandos de voz con DSP's. Actualmente estudia un Doctorado en Procesamiento Digital de Señales en el Posgrado de Ingeniería de la UNAM y su trabajo consiste en hacer reconocimiento de voz continua. Es profesor del laboratorio de microcomputadoras en la Facultad de Ingeniería de la UNAM. Ha trabajado como consultor en electrónica para diversas empresas.

Se recibió en 1979 de Ingeniero Mecánico Electricista con reconocimiento a su desempeño, en 1985 de Maestro en Ingeniería Electrónica, y en el 2001 de Doctor en Ingeniería, todos por la UNAM y el doctorado con el apoyo de la Universidad de California en Davis. Realizó una estancia posdoctoral en 2001 en Carnegie Mellon University. Autor de más de 30 artículos sobre codificación, reconocimiento, síntesis y ensanchamiento de voz. Es Jefe del Laboratorio de Procesamiento de Voz, en la Facultad de Ingeniería de la UNAM. Profesor en la UNAM desde 1979. Actualmente realiza una estancia de investigación anual en la Southern California University, EU.