Diseño y creación de un pequeño corpus oral de habla espontánea en español del centro de México para el desarrollo de sistemas de reconocimiento automático de voz

Carlos Daniel Hernández Mena[1]
ca_hernandez@uxmcc2.iimas.unam.mx

José Abel Herrera Camacho[2]
abelh@verona.fi-p.unam.mx

[1]Profesor de Asignatura de la Facultad de Ingeniería de la UNAM y Estudiante del Doctorado en el área de Procesamiento Digital de Señales del Posgrado de Ingeniería de la UNAM; [2] Profesor Titular de la Facultad de Ingeniería de la UNAM.

RESUMEN

En el presente artículo se habla sobre la creación de un pequeño corpus oral de habla espontánea en idioma español, recopilado entre estudiantes universitarios de la ciudad de México. El objetivo es tratar de hacer frente a la escasez de corpus adecuados en español de México para el desarrollo de sistemas de reconocimiento automático de voz. Se describirá la metodología de como fue creado así como sus características generales.

1. INTRODUCCIÓN

Existe un problema recurrente para las personas que se dedican al reconocimiento automático de voz y se trata sin duda de la escasez de bases de datos de calidad suficiente, tamaño importante e idioma adecuado. Aunado a esta escasez se encuentra el hermetismo con la que muchos grupos de investigación mantienen sus corpus, dificultando el desarrollo de este tipo de tecnologías. Algunos ejemplos de importancia histórica de corpus en español aparecen listados en **[LLISTERRI, 2004]**, pero aunque la lista parezca numerosa, la realidad es que para el desarrollo de tecnologías de reconocimiento de voz es muy necesario tomar en cuenta el habla particular de la región en donde se quiere hacer la aplicación o estudio en cuestión, y tomando en cuenta esto, la lista se revela entonces bastante pequeña.

ROC&C'2012 / CP-11 PONENCIA RECOMENDADA POR EL COMITÉ DE COMPUTACIÓN DEL IEEE SECCIÓN MÉXICO Y PRESENTADA EN LA REUNIÓN INTERNACIONAL DE OTOÑO, ROC&C'2012, ACAPULCO, GRO., DEL 11 AL 15 DE NOVIEMBRE DEL 2012.

Sin embargo, hay que decir que existen buenas razones, al menos para que el hermetismo del que hablamos ocurra de esta manera ya que es sumamente arduo el trabajo que se necesita para la realización de una base de datos de este tipo por pequeña que sea y sobre todo si el nivel de etiquetado es muy profundo ya que se requieren especialistas con más conocimientos cada vez. En **[PINEDA Y VILLASEÑOR, 2004]** podemos darnos una idea de lo minucioso que debe ser el trabajo para el diseño y creación de un corpus bastante más profesionales y de mayor tamaño que el que es presentado aquí, además de las herramientas que van creciendo en número y complejidad dependiendo el tipo de objetivos que persiga el corpus.

No obstante todos estos problemas, nosotros nos hemos dado a la tarea de crear un pequeño corpus que resuelva ciertas necesidades de reconocimiento de voz, mismas que incluyen el entrenamiento de nuevos estudiantes que llegan cada semestre a nuestro laboratorio con la intensión de realizar los mas variados proyectos que involucren control por medio de la voz.

En las siguientes secciones se exponen detalladamente las características de nuestro corpus oral, la metodología seguida para su realización, así como también una breve descripción del equipo utilizado para su grabación. Es también preciso mencionar aquí que nuestra metodología estuvo basada en parte en el trabajo presentado por Casacuberta en [**CASACUBERTA y LLISTERRI, 1992**].

2. CARACTERÍSTICAS DEL CORPUS

El objetivo del presente trabajo fue el de crear una base de datos con una cantidad considerable de habla espontánea con condiciones de ruido similares a las de una oficina tranquila y con los sexos de los hablantes balanceados. Esta base de datos permitirá entrenar sistemas de reconocimiento continuo de voz para diferentes aplicaciones y experimentos futuros.

El tamaño total del corpus fue de aproximadamente 4 horas y se recopiló de entre 150 estudiantes universitarios (75 hombres y 75 mujeres) con edades que oscilaban entre los 20 y 30 años de edad y cuya lengua materna era el español que se habla en el Distrito Federal y la zona metropolitana. La idea del número de hablantes la tomamos del trabajo de Barnard en **[BARNARD, 2009]**, en donde se demuestra que un corpus con solo 50 hablantes puede bastar para ser útil en un sistema de reconocimiento de voz, pero son claros al indicar que es aún más importante el número de hablantes que la cantidad de audio recabada.

3. METODOLOGÍA

El corpus fue grabado dentro de nuestro laboratorio con condiciones de ruido similares a las de una oficina pequeña y tranquila. Para el proceso de grabación se utilizaron una computadora de escritorio que era utilizada por el experimentador para almacenar la grabación y una lap-top conectada a un monitor externo para desplegar imágenes que el usuario pudiera describir. De esta manera, el usuario se quedaba sentado frente al monitor a una distancia aproximada de 50cm del micrófono dinámico cardioide utilizado y se le pedía que describiera con sus palabras una de 5 imágenes posibles (ver figura 1). Cuatro de estas imágenes eran cuadros del muralista Diego Rivera que contenían escenas de la conquista de México por parte de España, y la imagen restante era un autorretrato del pintor Juan O'Gorman.

Las imágenes se seleccionaron debido a la cantidad de objetos que contenían (utensilios, herramientas, armas, etc.) así como por la cantidad de emociones que expresaban, de esta manera les era muy sencillo a los hablantes describir detalladamente cada objeto en la pintura y las emociones que les provocaban.

Al principio de las grabaciones hicimos experimentos para que los hablantes contestaran preguntas a temas de opinión como: "¿Qué opinas de la política y de los políticos?" ó "¿Para ti que es la libertad?" basándonos un poco en el trabajo de Maekawa en **[Maekawa, 2003]** en donde se describe como los voluntarios hablaban sobre temas como "¿Cuál es el recuerdo más alegre/doloroso que recuerdes?" y eran llamados a hablar entre 10 y 12 minutos en un estudio de grabación profesional. En nuestro caso hubiera sido imposible retener a los hablantes tanto tiempo por que se trataba de estudiantes que nos donaban su voz entre clases, además de que optamos mejor por utilizar los cuadros de pinturas para que los hablantes los describieran por que con las preguntas a temas de opinión simplemente casi no hablaban nada.

En nuestro caso cada voluntario hablaba entre 2 y 3 minutos hasta que consideraba que había descrito todo lo que veía en la imagen. Un registro detallado del hablante fue recabado junto con su firma dando autorización a utilizar su voz para fines de investigación. Entre los datos recabados se tiene la edad y sexo del hablante, así como su lugar de nacimiento y de residencia actual, nacionalidad de los padres y se les pedía especificar los idiomas extranjeros que hablaran aunque fuera de manera muy básica.

Las edades de los hablantes oscilaban entre los 20 y 30 años de edad siendo la mayoría de ellos estudiantes de nivel licenciatura, aunque también había estudiantes de posgrado. Otra característica es que la mayoría de ellos estudiaba en alguna de las ramas de la ingeniería, es decir, que podemos considerar que todos tenían un "habla educada".

4. PROCESOS HECHOS AL AUDIO

Para seleccionar el audio, limpiarlo y convertirlo al formato deseado se utilizo la herramienta de software libre llamada Audacity[1]. Fue necesario escuchar las grabaciones para determinar que solo la voz del hablante era la que se escuchaba y todo fragmento de audio con sonidos de fondo, o más de una persona hablando al mismo tiempo fue descartado del corpus.

Después de seleccionar los fragmentos de audio que cumplían con los requerimientos fueron sometidos a una herramienta de eliminación de ruido del propio Audacity que tenia por parámetros un tiempo de ataque de 0.15 segundos y una reducción del ruido de 24dB.

A los fragmentos de audio resultantes finalmente se les asigno un nombre de archivo con el que serían reconocidos dentro del corpus y fueron exportados a un formato tipo WAV del Nacional Institute of Standards and Technology (NIST[2]) de los Estados Unidos a una frecuencia de muestreo de 16 khz con muestras de 16 bits. Este formato es ideal para trabajar con el sistema de reconocimiento de voz SPHINX

[1] http://audacity.sourceforge.net/?lang=es

[2] http://www.nist.gov/index.html

(versión 3[3]) de la universidad Norteamericana Carnegie Mellon (CMU[4]).

También es preciso mencionar que en este punto se seleccionaron los archivos para conformar el corpus de entrenamiento. El criterio de selección consistió en elegir 2 enunciados al azar por hablante, lo que daba un total de 300 enunciados diferentes, es decir, un 11.26% del corpus total.

Por ultimo, será importante mencionar el equipo de grabación utilizado que era de la marca Behringer:

- Interfase USB para digitalizar la señal de audio modelo UCA200[5]
- Mezcladora analógica de 5 canales modelo XENYX502[6]
- Micrófono electrodinámico cardioide modelo XM8500[7]

5. TRANSCRIPCIÓN Y MODELO DEL LENGUAJE

Una vez tenido los fragmentos de audio en el formato correcto se hicieron las transcripciones de cada archivo a mano en un editor de texto ASCII. Después de obtenido el archivo de transcripción completo se utilizó el paquete de scripts "Trascriber" del Dr. Luis Villaseñor del INAOE (presentado en **[PINEDA Y CASTELLANOS, 2009]** y en **[PINEDA Y VILLASEÑOR, 2004]**) para realizar la transcripción fonética con los 22 fonemas del español del México utilizando el alfabeto ASCII llamado Mexbet del Maestro Javier Cuétara Priede (presentado en las mismas referencia anteriores).

La transcripción ortográfica y fonética del corpus son requisito para la creación del modelo de lenguaje, para lo cual se utilizó la herramienta en linea Sphinx Knowledge Base Tool (Mejor conocida como LMTool[8]) de la universidad Carnegie Mellon. Esto debido a que nuestro corpus no excedía las 10 horas de duración (de haber excedido las 10 horas hubiera sido necesario utilizar otras herramientas dentro de nuestras computadoras, y no como la LMTOOL que corre en los servidores de la CMU).

6. CONCLUSIONES

Nuestro objetivo de creación de un pequeño corpus se logró con éxito y representará una herramienta importante en nuestras investigaciones, sin embargo, es preciso trabajar en un modelo de lenguaje mucho mejor del que se elaboro aquí, ya que este modelo se hizo muy especifico para este corpus y por lo tanto carece de generalidad. Es muy probable que la elaboración de un modelo de lenguaje adecuado de pie a un nuevo trabajo para presentar en el futuro.

También es preciso mencionar aquí que como trabajo futuro contemplamos el etiquetado fonético de este corpus que esperemos de inicio para principios del siguiente año.

Y como es nuestra intensión apoyar el desarrollo de estas tecnologías podemos poner nuestro corpus a disposición de otros investigadores, profesores o estudiantes con solo enviarnos un correo electrónico comentándonos brevemente cual será su experimento o aplicación y realizar la cita correspondiente si es que llegan a utilizarlo.

---

[3] Para descargar SPHINX 3 y obtener mas información sobre su utilización consultar la página: http://www.speech.cs.cmu.edu/sphinx/tutorial.html

[4] http://www.cmu.edu/index.shtml

[5] Manual de usuario del UCA200 en: http://www.produktinfo.conrad.com/datenblaetter/300000-324999/303350-an-01-en-U_Control_UCA200.pdf

[6] Características de la mezcladora XENYX502 en: http://www.behringer.com/EN/Products/502.aspx

[7] Características del micrófono XM8500 en: http://www.behringer.com/EN/Products/XM8500.aspx

[8] Disponible en: http://www.speech.cs.cmu.edu/tools/lmtool-adv.html

Figura 1. Disposición del equipo de grabación del corpus

AGRADECIMIENTOS

Queremos agradecer al Dr. Ivan Vladimir Meza del Instituto de Investigación en Matemáticas Aplicadas y Sistemas (IIMAS[9]) de la UNAM por su valiosísima asesoria para la creación de este corpus y al Dr. Luís Villaseñor del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE[10]) por permitirnos utilizar su "Transcriber" para la trascripción fonética de nuestro corpus.

REFERENCIAS

[BARNARD, 2009] Barnard, E.; Davel, M.; and van Heerden, C. 2009. "ASR corpus design for resource-scarce languages", In Proceedings of Interspeech, 2847–2850.

[CASACUBERTA y LLISTERRI, 1992] Casacuberta, F.- García, R.- Llisterri, J.- Nadeu, C.- Pardo, J.M.- Rubio, A. (1992) "Desarrollo de corpus para investigación en tecnologías del habla (Albayzín)", Procesamiento del Lenguaje Natural, Boletín no 12: 35-42. http://www.sepln.org/revistaSEPLN/revista/12/12-Pag35.pdf

[LLISTERRI, 2004] Llisterri, J. (2004) "Las tecnologías del habla para el español", in SEQUERA, R. (Ed.) Ciencia, tecnología y lengua española: la terminología científica en español. Madrid: Fundación Española para la Ciencia y la Tecnología. pp. 123-141. http://liceu.uab.es/~joaquim/publicacions/TecnolHablaEsp_FECyT03.pdf

[MAEKAWA, 2003] K.Maekawa. "Corpus of Spontaneous Japanese: Its design and evaluation", In Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (this volume), 2003

[PINEDA Y CASTELLANOS, 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 Y VILLASEÑOR, 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.

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 anual, en 2001, en Carnegie Mellon University;, y una estancia de investigación anual en University of Southern California. Autor de más de 50 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.

.

[9] http://www.iimas.unam.mx/

[10] http://www.inaoep.mx/