Tecnología del Habla: Debate sobre retos pendientes y perspectivas

José Manuel Pardo

Grupo de Tecnología del Habla - Dpto Ingeniería Electrónica
- Universidad Politécnica Madrid
ETSIT Ciudad Universitaria, 28040 Madrid
pardo@die.upm.es

Resumen: En este artículo comentamos las ponencias realizadas en el Taller de Tecnología del Habla realizado en el marco del XIX Congreso de la SEPLN y planteamos una serie de preguntas clave para entender la situación actual y el futuro de dicha tecnología. Además se aventuran una serie de respuestas a dichas preguntas.

Palabras clave: Tecnología del habla, Análisis situación actual, Perspectivas, Progresos

Abstract: In this paper we comment the presentations done in the Speech Technology Workshop held in the framework of the XIX Congress of the SPLN and make a series of key questions important to understand the current state and the future of the mentioned technology. Furthermore a set of possible answers are given to each of the question.

Keywords: Speech Technology, Current state, Perspectives, Progress

1 *Introducción*

En este artículo pretendemos por una parte hacer un resumen comentado de las ponencias presentadas al Taller de Tecnología del Habla del XIX Congreso de la SEPLN. Por otro lado pretendemos introducir al lector en el debate : ¿Cuáles son los retos pendientes? ¿Cuáles son las perspectivas de la Tecnología del Habla?

2 *Resumen comentado de las ponencias presentadas*

Luis Hernández en su ponencia titulada "Modelo de evolución de la Tecnología del Habla y tendencias futuras" (Hernández 2003) analiza un modelo común de evolución de las tecnologías del habla desde un modelo cognitivo hasta un modelo gobernado por los datos y con una alta potencia de cálculo. Ese modelo se aplica mas o menos igualmente a la codificación de voz, al reconocimiento de habla y a la conversión texto-habla. Queda pendiente el uso de este modelo para el desarrollo de sistemas de diálogo en el que se comprende la frase y a la vez se genera una

pregunta o respuesta con sentido al usuario. El modelo es adecuado y hasta ahora se ha cumplido. Sin embargo en tareas de diálogo, será más difícil disponer de datos etiquetados, pues el motor del progreso en los sistemas ha estado ligado permanentemente a la disponibilidad de herramientas que etiqueten de manera automática los datos.

Precisamente las herramientas de etiquetado automático es uno de los campos de desarrollo futuro identificados por Luis Hernández, junto con la adaptación al habla espontánea y la traducción voz-voz. Curiosamente, en un artículo mas o menos simultáneo, Church (Church 2003) defiende la evolución de las tecnologías de la lengua con un mismo modelo primero cognitivo, después (actual) dirigido por datos pero añade que antes del modelo cognitivo hubo uno previo también dirigido por datos y prevé en el futuro de nuevo un modelo cognitivo.

Carmen García Mateo, (García Mateo, 2003), por el contrario, incide más en el problema del desarrollo de las lenguas minoritarias. Aunque algunos métodos solo necesitan de datos etiquetados para poder obtener el sistema, los datos etiquetados son un

bien muy escaso, especialmente en las Lenguas minoritarias. En ese artículo presentado por ella, se incide en algunos desarrollos de Tecnología del Habla para el gallego. Mi comentario es el siguiente: El desarrollo de Tecnología del Habla para lenguas minoritarias será siempre difícil en términos de recursos, pues si considerando un modelo basado en datos, se necesitan datos estructurados, cuando se pase a un modelo cognitivo es presumible que se necesitarán también expertos en el lenguaje minoritario.

Javier Ferreiros (Ferreiros, 2003), después de una breve introducción histórica, en su ponencia hace un repaso de la situación actual de la Tecnología del Habla y sus problemas:

- En conversión texto-voz, los problemas de naturalidad y el uso de estos sistemas en la vida real.
- En reconocimiento de habla, la robustez frente a diferentes condiciones de variabilidad de entorno (temas coincidentes con L. Hernández)
- En comprensión de habla, la cobertura del lenguaje espontáneo.

Prosigue Ferreiros su ponencia con la expresión de las líneas de futuro tangibles y deseables. Las líneas de futuro tangibles se relacionan directamente con los problemas apuntados, mejora de naturalidad, flexibilidad en la generación de voces e inclusión de emociones en la voz. En reconocimiento de habla, trabajo sobre robustez. En comprensión de habla aprendizaje automático y organización de fuentes de información para su reutilización. En control de diálogo, mejora de medidas de confianza para tener mejor control del mismo y de nuevo aprendizaje de modelos.

Finalmente, propone una ampliación de la cobertura del término Tecnología del Habla para incluir sistemas que además de comprendernos, entiendan nuestras “intenciones” y además también nuestros “gestos”. Las “intenciones” se entiende en un término mucho más amplio que la comprensión de un mensaje, se trata de poner ese mensaje en contexto para comprenderlo pero en relación con el conocimiento que el sistema tiene del mundo y las relaciones de este mundo con el interlocutor humano. Es una tarea ciertamente ambiciosa. Sin embargo la visión coincide con visiones de dos expertos en Tecnología del Habla presentadas en la sesión especial “Towards a roadmap for Speech Technology”

de Eurospeech 2003, la de Ron Cole (Cole 2003) y Bjorn Granstrom (Granstrom 2003).

3 ¿Cuáles son los retos pendientes? ¿Cuáles son las perspectivas de la Tecnología del Habla?

En estas líneas pretendo introducir el debate central del Taller celebrado en el Congreso desde mi punto de vista y para ello voy a plantear algunas preguntas y aventuro una respuesta personal a las mismas.

3.1 ¿Cómo ha evolucionado la Tecnología del Habla desde sus comienzos hasta nuestros días?

Pues bien, en los comienzos, no existía ni el término de Tecnología del Habla. Todo lo relacionado con Habla estaba incluido en el entorno de procesamiento de señal, y más que de procesamiento de habla se hablaba de procesamiento de la señal de voz (sin significado). Los primeros codificadores de voz utilizaban métodos de procesado de señal. Los reconocedores de voz utilizaban métodos de reconocimiento de patrones relativamente simples (Dynamic Time Warping) y los sintetizadores eran meros filtros en serie o en cascada aplicados a una función excitación, típicamente una serie de pulsos un otra función simple que intentaban simular rudimentariamente la generación de fonemas por el método fuente-filtro. Poco a poco se fueron introduciendo a la vez una serie de conocimientos lingüísticos, más computación y más datos para llegar a la tecnología actual que es prácticamente estadística. La tecnología actual está bastante probada y sedimentada y mejora automáticamente con la disponibilidad de nuevos datos. De ahí la importancia capital de disponer de datos, pero datos que puedan procesarse por métodos automáticos. En todos los entornos e idiomas no existen datos suficientes por lo que aún queda espacio de mejora solo añadiendo más datos. Sin embargo en unos años esta metodología se saturará, y habrá que volver a considerar los modelos de base, más complejos.

3.2 ¿Se consiguió todo lo que se esperaba o ha habido muchos retrasos o decepciones?

Como en otras tecnologías, la evolución ha sido con avances y estancamientos y casi siempre las predicciones han sido muy optimistas. Llevamos más de 15 años oyendo predicciones de gurús que pronosticaban grandes negocios en tecnología del habla que no se han cumplido. Los errores más importantes de predicción que se han cometido han sido los siguientes:

- Pensar prematuramente que el problema de síntesis de voz era más simple que el de reconocimiento y que estaba resuelto. Se ha invertido comparativamente mucho menos en síntesis de voz que en reconocimiento, y por ello existen actualmente muy pocos conversores texto-habla ampliamente utilizados. Sobre todo, nadie pensó que un habla inteligible no era suficiente para la aceptación por el usuario y que el usuario demandaba naturalidad, es decir no grandes diferencias con una grabación.
- Pensar prematuramente que un sistema de reconocimiento de laboratorio, probado con los desarrolladores iba a funcionar en el mundo real. No se pensó en dos factores muy importantes, la reacción de un usuario ante una máquina que es de por sí un problema muy complejo de modelo de interacción del usuario, y la variedad de entornos en los que es necesario utilizar los sistemas.
- Pensar prematuramente que un sistema de diálogo guiado por la máquina era un sistema útil. En muchos casos, los innumerables menús hacían que el sistema fuera difícil de usar y tedioso y el usuario acababa por no utilizarlo. Por otro lado, y según Heisterkamp propone, no estaría de más forzar en algunos casos convenciones. La historia nos demuestra que ha habido modos de uso no naturales que se han propagado simplemente por convención, es el usuario el que se aproxima al funcionamiento de la máquina y no al revés (Heisterkamp 2003).
- Pensar prematuramente que la medida de calidad más importante de un sistema de diálogo es la tasa de reconocimiento y sin embargo la medida más importante es la

tasa de aceptación del usuario o las ventajas que proporciona frente a otro método alternativo, coste, rapidez de servicio etc...

Aún con todos estos errores, la industria ha ido evolucionando positivamente, corrigiendo los errores iniciales y haciendo de la Tecnología del Habla una tecnología cada vez más necesaria (por útil).

3.3 ¿Cuáles han sido hasta ahora los problemas más importantes que ha habido que resolver? ¿Cuáles se han resuelto y cuáles no?

Los problemas a resolver en Tecnología del Habla surgen del complejo mecanismo que tenemos los humanos para conseguir esta comunicación. Este mecanismo no es aún bien conocido ya que las funciones más importantes las realiza el cerebro y son una cualidad de la inteligencia. No se han descubierto todavía las estructuras fundamentales del procesamiento del habla por el cerebro.

Lo que se ha resuelto hasta ahora ha sido utilizando muchos datos (al menos últimamente) y métodos de entrenamiento robustos. La solución del problema (parcialmente) no implica que la forma de actuar el mecanismo de producción y percepción humana en el hombre esté relacionado con el método utilizado. Extrapolando tasas de sistemas actuales en reconocimiento de habla a los que se les someta a tantos datos como un hombre oye en su vida no se conseguiría la misma tasa de acierto que el hombre (R. Moore 2003).

Se han resuelto aplicaciones que reconocen habla telefónica para un gran número de hablantes y un conjunto de palabras limitada. El reconocimiento de habla continua telefónica es aún muy rudimentario y poco utilizado, a no ser que hablemos de una aplicación muy restringida. Cuando el canal es ruidoso, el hombre funciona mucho mejor que las máquinas.

3.4 ¿Cómo se relaciona la tecnología del Habla con la Tecnología general de la Lengua?

Hasta ahora las tecnologías de la lengua se podían clasificar en dos partes principales, tecnologías del texto y tecnologías de señal. Las tecnologías del texto utilizaban el lenguaje

expresado en texto para obtener información, generarlo, procesarlo, traducirlo etc... Las tecnologías de señal procesaban el lenguaje expresado de forma oral con métodos de procesamiento y reconocimiento de patrones. Estas dos tecnologías, desde hace dos décadas trabajan ya en colaboración para conseguir progresos en Tecnología del Habla. El Habla es el lenguaje más natural por excelencia (mucho mayor que el texto) y por ello es mucho más complicado su proceso. Cada vez existen más expertos de procesamiento de texto interesados por la Tecnología del Habla, véase por ejemplo la red europea ELSNET que es una red que trata de texto y de habla y de los nuevos centros o institutos creados con la unión de ambos expertos (un buen ejemplo es el TALP de la UPC). Sin embargo existen aún mucho camino por recorrer para que la tecnología del texto actual sea plenamente utilizada en Tecnología del Habla, incluso será necesario desarrollar nuevas metodologías de tecnología de texto aplicadas a transcripciones textuales de comunicaciones orales.

3.5 ¿Ha habido mayor evolución en Tecnología del Habla o Tecnología del Texto o la evolución de ambas ha ido paralela?

Aunque fundamentalmente los expertos en ambas tecnologías han trabajado por separado, en ambos campos se han conseguido progresos notables. En tecnología de texto existen herramientas muy potentes para la ayuda a la escritura (correctores de texto y de estilo) y a la traducción de texto. En tecnología de habla existen sistemas telefónicos que aportan un ahorro muy importante a tareas relativamente rutinarias realizadas por operadoras. Este es el campo que con diferencia más éxitos comerciales ha conseguido.

3.6 ¿Cuáles son los pasos previstos en los próximos diez años? ¿Con qué dificultades nos encontraremos?

La Comisión Europea ha pedido a ELSNET¹ que elabore una Hoja de Ruta (en inglés Roadmap) sobre las tecnologías de la Lengua en los próximos diez años <http://www.elsnet.org/> que se ha organizado por

¹ Red de Excelencia Europea de Habla y Lenguaje

talleres, en algunos de los cuales el autor ha participado. En ella se hacen predicciones de qué tecnologías y sistemas estarán disponibles y en qué tiempo. Lógicamente cualquier tipo de predicción es especulativa, pues para que se cumpla, se necesita que confluyan una serie de factores que no controla necesariamente el que las hace. Por ejemplo, es fundamental que para que la tecnología se desarrolle haya suficiente inversión en la misma, y para ello, debe haber también un negocio razonable en la utilización de las tecnologías y productos actuales. Por otro lado, para que la Tecnología se desarrolle se necesita también inversión pública, y eso depende de los gobiernos con sus cambios y contrastes. A nadie se le oculta que gran parte de la evolución tecnológica y la ventaja relativa que tiene Estados Unidos en muchas tecnologías, y en concreto en Tecnología del Habla tiene su base en grandes inversiones en investigación dirigidas a la Universidad y a la Empresa, financiadas por su departamento de defensa. Desde 1970 DARPA ha sido una de las grandes agencias financiadoras de la investigación en Tecnología del Habla y la creadora de sistemas de evaluación competitiva que ha potenciado enormemente la capacidad de las universidades y empresas norteamericanas en el area.

3.7 ¿Qué temas quedarán resueltos en los próximos diez años?

Como hemos dicho, hacer predicciones es siempre difícil, pero se espera que en los próximos diez años, los sistemas de reconocimiento por teléfono mejorarán tanto el número de palabras manejadas como el diálogo con el usuario, que será más natural y más entendible por el mismo y por tanto será más fácil de utilizar.

En conversión texto-habla, se espera disponer de más naturalidad en la voz, más sentimiento (emoción), más flexibilidad (número de voces y de idiomas) y más aceptación por parte de los usuarios (en algunos casos concretos, el habla será indistinguible del habla grabado).

Para predicciones más especulativas, pero con cierta base metodológica, ver los libros publicados de Ray Kurzweil (1999, 1990)

3.7 ¿Qué acciones deberemos llevar a cabo para que se cumplan los objetivos previstos?

Habrà que trabajar sobre todo en el etiquetado y procesado automàtico de bases de datos. Hoy por hoy, està claro que con màs datos se mejoran los sistemas, pues el marco està dado. Pero el mètodo para conseguir màs datos etiquetados de forma barata aùn no se tiene. Eso no quiere decir que no sea importante tambièn el trabajo en modelado y en ciencia fundamental para mejorar y optimizar los mètodos de entrenamiento de los sistemas y los motores fundamentales de generaci3n y de reconocimiento de habla. Segùn Church (2003) esto se harà necesario inevitablemente dentro de 10 aõos.

3.8 ¿C3mo se relacionarà la Tecnología del Habla con otras tecnologías: Multimedia, Multilinguaje, Reconocimiento y Generaci3n de Imàgenes Animadas?

Es sabido que el mensaje del habla se complementa con el de nuestro interlocutor. Es màs difìcil para un humano reconocer un habla por telèfono que en directo. Esto es no s3lo porque el canal telef3nico tiene un ancho de banda reducido sino porque en la comunicaci3n cara a cara, los gestos y movimiento de la boca llevan informaci3n sobre las palabras.

En el futuro serà cada vez màs importante el que podamos integrar reconocimiento de la seõal de voz junto con reconocimiento de los movimientos de la cara. Con ello podremos reconocer tanto el significado lingüístico como el significado metalingüístico (la emoci3n, la ironía etc.) Esta integraci3n tambièn serà útil en el caso de que queramos generar un habla sintética. Este habla sintética serà màs inteligible si aõadimos una cara sintética que imite los gestos humanos.

En relaci3n con la multimodalidad, el lenguaje deberà integrarse con otros modos de actuaci3n (rat3n, teclado, etc.) para conseguir una interacci3n con las màquinas màs ergon3mica e "inteligente".

Conclusi3n

Es siempre relativamente fàcil analizar el pasado, ya que conocemos su evoluci3n. Predecir el futuro para intentar adelantarnos a

los problemas es mucho màs difìcil, pero el ejercicio puede ser útil para definir posibles areas de trabajo comunes para los grupos que trabajan en este tema.

Bibliografía

- Hernàndez L, 2003. Modelo de evoluci3n de la Tecnología del Habla y tendencias futuras *Actas del XIX Congreso de la Sociedad Espaõola para el Procesamiento del Lenguaje Natural, SEPLN*. Madrid, Septiembre.
- García Mateo C. 2003. Tecnologías del habla y lenguas minoritarias. *Actas del XIX Congreso de la Sociedad Espaõola para el Procesamiento del Lenguaje Natural, SEPLN*. Madrid, Septiembre.
- Ferreiros J., 2003. ¿Qué queremos que sea la tecnología del Habla"? *Actas del XIX Congreso de la Sociedad Espaõola para el Procesamiento del Lenguaje Natural, SEPLN*. Madrid, Septiembre.
- Church K. 2003, Speech and Language Processing: Where Have We Been and Where Are We Going? *Eurospeech 2003 Geneve*.
- Cole R. 2003, Roadmaps, Journeys and Destinations Speculations on the Future of Speech Technology Research *Eurospeech 2003 Geneve*.
- Granstrom B. and House D. (2003) Multimodality and speech technology: Verbal and non-verbal communication in talking agents. *Eurospeech 2003 Geneve*.
- Heisterkamp P. (2003) "Do not attempt to light with match!": Some thoughts on progress and research goals in Spoken Dialog Systems, *Eurospeech 2003 Geneve*.
- Moore R. (2003) A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners, *Eurospeech 2003 Geneve*.
- Kurzweil R. (1999) The age of the Spiritual Machines, *Phonenix*.
- Kurweil R. (1990) The age of the intelligent Machines, *MIT Press*.