Modelo de evolución de la Tecnología del Habla, y tendencias futuras

Luis Hernández Gómez
ETSI Telecomunicación - Universidad Politécnica de Madrid
Cdad Universitaria s/n 28040 Madrid
luis@gaps.ssr.upm.es

Resumen: Esta ponencia propone un modelo común de evolución para las diferentes tecnologías de base que integran la Tecnología del Habla. En ese modelo se parte de un conocimiento inicial de aspectos básicos de la teoría general de la lengua, sobre el que se incorporan potentes algoritmos de cálculo gobernados por los datos. Tras exponer el notable nivel de desarrollo tecnológico que ese modelo de evolución ha supuesto para la codificación, la síntesis y el reconocimiento de habla, se plantean los principales retos futuros para la consolidación de las aplicaciones de Tecnología del Habla. Postulándose un mayor grado de relación con la Tecnología de Texto, y nuevos retos para el diseño de interfaces, especialmente en aspectos relacionados con la robustez y con el diseño de interfaces multimodales que permitan modos de interacción natural.
Palabras clave: Modelo de evolución de la Tecnología del Habla. Algoritmos gobernados por los datos. Interfaces Telefónicos. Tecnología de Habla y Tecnología de Texto. Robustez. Multilingualidad. Multimodalidad. Interacción natural.

Abstract: This talk proposes a common model for describing the evolution of different technologies involved in Speech Technology. The proposed model describes the evolution from a general knowledge of basic language theory concepts that is combined with powerful data-driven processing algorithms. After presenting the overall technological success of speech coding, synthesis and recognition, we discuss the main future research and development actions to promote new Speech Technology applications. We foresee an increasing degree of interaction with Text Technology and new challenges in the design of more robust and more natural multimodal interfaces.
Keywords: Speech Technology evolution. Data-driven algorithms. Telephone Interfaces. Speech and Text Technologies. Robustness. Multilinguality. Multimodality. Natural Interaction.

1 Introducción

En esta ponencia nos situaremos en el análisis del pasado, presente y futuro de la Tecnología del Habla desde la perspectiva de un grupo de investigación que lleva más de una década apoyando el desarrollo de esta tecnología en la empresa Telefónica Investigación y Desarrollo. Durante este periodo de tiempo podemos afirmar que el avance científico-tecnológico en Tecnología del Habla ha sido espectacular. Este avance queda patente al contemplar el nacimiento y la difusión de productos y aplicaciones como programas de dictado, o servicios de telefonía, hasta hace poco impensables, como el servicio 1003 (11818) automático de Telefónica. Pero a pesar de lo anterior son todavía muchos e importantes los retos a afrontar para alcanzar la ansiada consolidación de los productos y aplicaciones de la Tecnología del Habla.

La exposición de esta ponencia se estructurará alrededor de tres aspectos principales:

- El análisis de la evolución de las principales tecnologías involucradas. Destacando su problemática inicial, proponiendo, como punto de debate y discusión, un modelo común de evolución, y planteando la problemática y conexiones en relación con la tecnología de texto.
- El estado actual del desarrollo de servicios y aplicaciones de acceso a información a través del teléfono.

- Las perspectivas futuras de evolución incidiendo, especialmente, sobre la combinación del habla con otros modos de comunicación orientados a lo que actualmente se viene denominando multimodalidad o, de una manera más amplia, "interacción natural".

2 El modelo de evolución tecnológica

Revisando el pasado *próximo* o *cercano* de la Tecnología del Habla (desde los años 70), podemos situar los primeros trabajos en codificación de voz, conversión texto a habla y reconocimiento del habla como actividades experimentales de laboratorios y centros de investigación. Si bien la codificación de voz fue la primera tecnología que presentó un acceso extendido a aplicaciones comerciales a través de su incorporación a sistemas de comunicación digitales. Contemplando retrospectivamente el avance de esta "primera" tecnología ya encontramos un modelo de evolución tecnológica que, con las lógicas diferencias respecto a las restantes tecnologías, podemos postular como común al que posteriormente sufrieron el reconocimiento y la síntesis de habla.

2.1 Codificación de Voz

La evolución de la tecnología de codificación de voz, parte del conocimiento base de la teoría general de la lengua, en este caso del modelo fuente (excitación) – tracto vocal (filtro) propuesto por la fonética acústica, y, en un momento de su desarrollo tecnológico, encuentra una algorítmica o técnica de procesado dirigida o gobernada por los datos. Surge de este modo un importante avance tecnológico de "altas prestaciones" (alta calidad) a expensas de, también, una elevada exigencia en operaciones de cálculo (carga computacional). Los algoritmos de codificación de voz que hoy encontramos en la práctica totalidad de los sistemas de comunicaciones exigían, en los años 80, una potencia de cálculo que requería 125 segundos de proceso del ordenador más potente de la época para procesar (codificar / descodificar) un segundo de voz. Pero el espectacular avance en la potencia de cálculo de los procesadores, al que seguimos asistiendo, ha permitido que hoy encontremos este tipo de algoritmos de codificación en nuestros teléfonos móviles. Bien, pues un punto de debate sería que ese modelo de evolución, desde el conocimiento de la teoría general de la lengua hasta modelos matemáticos gobernados por los datos, y con elevadas exigencias de cálculo, es el que también encontramos para las siguientes tecnologías.

2.2 Reconocimiento de Habla

La tecnología de reconocimiento de habla se ha ido decantando hacia el uso de técnicas estadísticas de clasificación de patrones, que precisan un aprendizaje o entrenamiento a partir de grandes bases de datos de voz. Son sistemas que han ido progresando desde el reconocimiento de conjuntos reducidos de unos pocos comandos, hasta sistemas con capacidad para reconocer vocabularios próximos al millón de palabras. Sin menospreciar un grandísimo esfuerzo en la mejora y evolución de los algoritmos de reconocimiento, gran parte del avance tecnológico logrado ha sido posible por la disponibilidad de medios y recursos de cómputo y memoria cada vez mayores. También en este caso encontramos, obviamente, unas bases fundadas en el conocimiento general del lenguaje: desde la fonética, para la definición y estructuración de unidades de reconocimiento, el diseño de corpus de entrenamiento o la generación de diccionarios de reconocimiento, hasta modelos para la construcción de redes o gramáticas de reconocimiento. Si bien, en aspectos tales como el correspondiente a este último punto, el contacto con la teoría general de la lengua o con la Tecnología de Texto ha quedado, nuevamente, marcada por el uso de modelos que, en el mejor de los casos, son análisis estadísticos del lenguaje. Es decir, son modelos que adquieren mayor potencia en función de la cantidad de datos de "entrenamiento" disponibles. Cabe aquí destacar que para el desarrollo de sistemas de voz interactivos, el texto proporcionado por el reconocedor de habla generalmente precisa ser procesado por un módulo de análisis semántico encargado de extraer los conceptos y contenidos necesarios para el control del diálogo. Es en este un ámbito donde ha sido deseable, posible, y, en muchos casos, fructífera la colaboración con la Tecnología de Texto. Tomemos como ejemplo los trabajos realizados por el grupo de investigación Julietta de la Universidad de Sevilla. Estos trabajos han estado orientados a adaptar esquemas de análisis semántico a las

necesidades de robustez que son exigibles para su utilización en sistemas de tecnologías del Habla. Puede considerarse este caso como paradigmático en la relación entre Tecnología de Texto y Tecnología del Habla. Ya que son especialmente las limitaciones en el funcionamiento de los reconocedores de habla lo que, por un lado, restringe las posibilidades de uso de técnicas avanzadas de procesado de texto, y, por otro, exigen su re-planteamiento. Replanteamiento para incorporar las exigencias de robustez necesarias para la operatividad de los sistemas de Tecnología del Habla. Adicionalmente, e intentando avanzar más hacia lo que ya puede vislumbrase en desarrollos futuros, una vez que se produce este acercamiento entre tecnologías, y la Tecnología del Habla comienza a mejorar sus prestaciones (tanto por grado de evolución algorítmica como por potencia de cálculo) es donde pueden plantearse nuevos ámbitos de integración más potentes. Por ejemplo, siguiendo con el caso anterior, hoy es posible plantearse una aplicación real que contemple la inclusión de un potente módulo de análisis semántico, como el desarrollado por la Universidad de Sevilla, en el proceso de decodificación acústica de un reconocedor de habla comercial, como el de Telefónica I+D. Otro caso en cierta medida análogo e ilustrativo de lo que pretendemos explicar puede encontrase analizando la evolución de los sistemas de reconocimiento de locutor. En este campo las técnicas de modelado acústico parecen haber alcanzado un techo tecnológico difícil de superar, y es ahora cuando se replantean los beneficios de la incorporación de otras fuentes de conocimiento del lenguaje: como el modelado prosódico ó el modelado de lenguaje o de cadenas de fonemas utilizados por un locutor determinado.

Pero antes de seguir con lo que serían perspectivas futuras, analizaremos el modelo de evolución de los sistemas de conversión texto a habla.

2.3 Conversión Texto - Habla

Tradicionalmente el desarrollo de los sintetizadores de habla se ha organizado en torno a un núcleo de análisis acústico y otro de análisis lingüístico-prosódico. En el núcleo de análisis lingüístico-prosódico es donde, obviamente, encontramos un mayor contacto con la Tecnología de Texto. Pero volviendo otra vez la vista al modelo de evolución para la síntesis de habla, que como en el caso de reconocimiento ha sido espectacular en la última década, encontramos nuevamente un patrón de comportamiento similar. El mayor reto planteado, una vez superado el inicial de la inteligibilidad del mensaje sintetizado, es el nivel de calidad y naturalidad de la voz sintética; y éste se está alcanzando a partir de la llegada de algoritmos de síntesis a partir de extensas bases de datos de voz pre-grabada: técnica conocida como *síntesis por corpus*, que permite desarrollar sintetizadores de habla de altísima calidad, en muchos casos indistinguible de habla natural. Así, sobre una base de técnicas de análisis de texto, sobre vienen un conjunto de potentes algoritmos de generación, anotación y selección de unidades en grandes bases de datos, para producir como resultado una voz sintética de calidad muy elevada. Pero nuevamente, al igual que sucedía en la evolución de los sistemas de reconocimiento, una vez alcanzado un alto grado de madurez entra en juego la exigencia de incorporar potentes técnicas de procesado de texto orientadas a dar solución a aspectos críticos tan importantes como: la corrección automática del texto a leer, la identificación del idioma, o la definición de modelos prosódicos adaptados a síntesis de emociones, situaciones de diálogo, etc.

3 Tecnología del Habla en Sistemas Telefónicos Interactivos

Comentamos al principio de esta ponencia, que nuestra discusión iba también a considerar la problemática asociada al diseño y desarrollo de servicios y aplicaciones de acceso a información a través del teléfono. A tal fin, junto a las tecnologías que hemos revisado, -y que podríamos denominar “de base”- encontramos las tecnologías que permiten su integración (tecnologías de integración) con vistas a la construcción de sistemas de voz interactivos, donde destaca especialmente la tecnología de gestión de diálogo. En este campo podemos recoger todas las expectativas y frustraciones que imaginaban el desarrollo de sistemas de interacción hombre-máquina con prestaciones que sólo encontramos en las películas de ciencia-ficción. Pudiendo abrirse otro punto interesante de debate sobre si la comunicación hombre-máquina debe plantearse sobre el paradigma de la comunicación entre personas. Pero en cualquier caso la realidad del

progreso tecnológico actual nos presenta un horizonte mucho más limitado que lo imaginado. En este campo todavía no ha llegado a "predominar" los sistemas a partir de los datos (aunque existen propuestas interesantes) y el debate se plantea entre los sistemas "propietarios" y los "estándares". Los sistemas propietarios, desarrollos particulares de empresas o centros de investigación, generalmente presentan mayores prestaciones y competencias lingüísticas. Pero los nuevos estándares, VoiceXML y SALT, propuestos por consorcios de empresas, ofrecen una gran potencialidad para su desarrollo industrial, y puede debatirse si cuestionan o no, e incluso si frenan el desarrollo de sistema más potentes. Una posible vía intermedia podría abrirse recordando que estos estándares no son más que lenguajes de programación específicos de diálogos, y que los resultados de investigación más avanzados podrían orientarse hacia la generación automática de código según esos estándares, o en su integración/combinación con ellos; por ejemplo para la definición de sub-diálogos especializados en el intercambio de datos o informaciones comunes a diferentes dominios de aplicación, como pueden ser horas, fechas, números de teléfono, etc.. Incluso hay propuestas para utilizarlos como fase inicial de recogida de datos para sistemas de diálogo basados en aprendizaje desde ejemplos. Pero esto sería nuevamente una perspectiva de futuro, cuando en el presente la fuerza de la implantación de productos de Tecnología del habla está condicionada en gran medida por el desarrollo tecnológico sobre VoiceXML.

Nos enfrentamos así, en este ámbito concreto, al problema fundamental que frena la implantación de sistemas telefónicos interactivos, que puede resumirse como "falta de robustez". Esa falta de robustez deberá irse rebajando con nuevos desarrollos en reconocimiento y síntesis de habla. Quizá esto se consiga siguiendo alguna de las direcciones previamente apuntadas. Pero, dado que este avance se prevé lento, y los años de inversión y desarrollo ya empiezan a pesar sin grandes retornos económicos, surge la necesidad de realizar un esfuerzo, también costoso, de ajuste módulo a módulo y de mejora de la interacción entre ellos. Algo similar a lo ocurrido en los estándares de codificación de voz, pero ahora en un contexto mucho más complejo.

4 Tecnología del Habla en Español

La problemática del desarrollo de la Tecnología del Habla asociada al Español puede verse tanto desde sus ventajas como desde sus inconvenientes. En las ventajas podríamos situar la dependencia de las peculiaridades del idioma como escudo para frenar el desarrollo desde otros países. Este hecho ha supuesto que muchos de los trabajos y desarrollos de nuestros grupos de investigación hayan sido financiados por, y transferidos a empresas extranjeras.

Pero en el lado negativo tenemos que comenzar apuntando la escasez de tejido industrial capaz de absorber los resultados de grupos de investigación nacionales.

En el caso de Telefónica Investigación y Desarrollo, los excelentes resultados en el desarrollo de productos de Tecnología del Habla en castellano, catalán, gallego, euskera y dialectos del español hablado en Hispanoamérica, han posibilitado la potenciación de la División de Tecnología del Habla, mientras desaparecían grupos similares en otras operadoras tanto europeas como de Estados Unidos. Si bien hay que darse cuenta que, como hemos venido postulando, el éxito de las tecnologías básicas basadas en los datos hace perder fuerza, sin llegar a desvanecerse totalmente, la dependencia del conocimiento de la lengua concreta. Por lo que son empresas específicas las que aglutinan los principales productos multilingües del mercado. Sin embargo, retomando la discusión sobre la problemática de robustez, quizás vuelva a resurgir una alta dependencia con el idioma exigida por el diseño de sistemas de diálogo avanzados.

En relación con lo que acabamos de discutir, otro aspecto importante a desatacar, es la necesidad de afrontar el reto ineludible de la multilingualidad, como característica exigible a cualquier sistema de Tecnología del Habla. Ello supone no sólo disponer de tecnología básica (reconocimiento y síntesis) multilingüe, sino también del desarrollo de sistemas de gestión de diálogo, en la medida de lo posible, "independientes" del idioma.

5 Otras posibles tendencias futuras para la Tecnología del Habla

A lo largo de la exposición anterior ya hemos apuntando algunas de las posibles perspectivas de futuro para la Tecnología del Habla.

Destacando el común denominador de las propuestas que hemos discutido, habría que volver a resaltar que quizás sea el momento en que, dada la relativa madurez de los sistemas de reconocimiento y síntesis, haya llegado el momento del surgimiento de un mayor grado de interacción con la Tecnología de Texto.

Pero junto al planteamiento general anterior también podríamos apuntar algunas líneas de actividad que se prevé puedan ofrecer importantes beneficios para la potenciación del uso de la Tecnología del Habla, como son:

- La utilización de técnicas de personalización, adaptación y aprendizaje. Estas técnicas ya cuentan con notables desarrollos pero que en el futuro próximo podrán hacerse realidad sobre sistemas y aplicaciones reales, donde la generación, gestión y uso de información personalizada, aunque se trate de sistemas accedidos por grandes poblaciones de usuarios, es ya hoy en día técnicamente factible.
- El establecimiento de nuevos paradigmas de interacción hombre-máquina que den prioridad a los criterios de usabilidad de las aplicaciones. Aspecto que auque es reconocido como de vital importancia en la comunidad científica, es muchas veces dejado fuera de las consideraciones de diseño y planificación de aplicaciones, que no se establecen desde el beneficio de los usuarios, sino desde la óptica de reducción de costes.
- Por último podemos comentar que, paradójicamente, algunas de las aplicaciones de la Tecnología del Habla que tienen o podrían tener una lata demanda industrial se encuentran todavía con importantes retos científicos y tecnológicos, y ofrecen soluciones parciales, por debajo de lo que supondría un uso satisfactorio. Por tanto es en estos campos donde deberá concentrarse un importante esfuerzo de innovación y desarrollo futuro, nos referimos a:
    1. Sistemas de transcripción automática de habla.
    2. Sistemas adaptados al habla espontánea (incluyendo altos niveles de expresividad y emociones).
    3. Y sistemas de traducción automática voz-voz.

Finalmente, y también en la dirección de caminos de evolución ya iniciados y que, muy probablemente, veremos desarrollar en el futuro se encuentra la incorporación de nuevos modos de interacción que complementen la dificultad del uso de la voz. Esta línea de trabajo en interfaces multimodales, o más genéricamente en lo que se denomina interacción natural se prevé como una línea de trabajo que aportará indudables beneficios al diseño de futuras interfaces. Así el empleo de otros modos de comunicación (al menos un modo más de interacción) está demostrando plantear importantes mejoras.

- El primer motivo es la importante merma en prestaciones de una interfaz vocal en:
    1. Situaciones de ruido, donde la tecnología de reconocimiento de habla no permite garantizar un nivel de funcionamiento satisfactorio, y;
    2. En situaciones donde un usuario puede encontrase incómodo para utilizar la voz en su interacción con el terminal, por ejemplo en presencia de otras personas.
- El segundo de motivos hace referencia a la limitación intrínseca que un solo modo de interacción puede suponer a la hora de presentar al usuario determinados tipos de información; pensemos, por ejemplo, en la dificultad que supone la presentación de una larga lista de nombres o la localización en un plano en una interfaz exclusivamente vocal.

Por tanto, la multimodalidad, con la inclusión de técnicas de interfaces gráficos y/o de texto pueden complementar de una forma eficaz aquellas funcionalidades de difícil accesibilidad desde el control por voz.

En esta dirección ya encontramos los recientes estándares X+V y SALT, si bien como ya indicamos anteriormente, son estándares con prestaciones y potencia claramente limitadas, son un punto de referencia inicial que quizás podrán beneficiarse del desarrollo derivado de trabajos de investigación sobre otros sistemas multimodales propietarios.