¿Qué queremos que sea Tecnología del Habla?

Javier Ferreiros López
Grupo de Tecnología del Habla, Dpto. Ingeniería Electrónica, U.P.M.
E.T.S.I. TELECOMUNICACIÓN, Ciudad Universitaria s/n, 28040-MADRID
jfl@die.upm.es

Resumen: En el artículo se relata la evolución del término Tecnología del Habla tanto con visiones históricas como intentando imaginarse una definición futura, provocando la reflexión de qué nos queda por hacer.
Palabras clave: Tecnología del habla. Historia. Líneas de futuro.

Abstract: The evolution of the term Speech Technology is presented in the paper both through historic visions and trying to imagine a future definition, inducing the reflection about what is left to be done.
Keywords: Speech technology. History, Future work.

1 Introducción

Los implicados en el proceso de investigación en Tecnología del Habla (TH en adelante) podemos describir claramente qué es TH a la vista del trabajo ya realizado. Esta es una manera incompleta de definir este área de investigación ya que también debemos plantearnos su significado con una perspectiva de futuro.

El repaso histórico de la temática que hemos encuadrado en el término TH nos permitirá definir qué parte hemos conseguido realizar de las ideas e ideales que todos podríamos sumar para definir la TH en general y concretar de alguna manera las líneas futuras de trabajo en este campo que permitan acercarnos a dicho ideal.

Si no realizamos este esfuerzo de vez en cuando, corremos el riesgo de que los que definen la financiación para las diversas temáticas, a veces alejados de las necesidades reales de investigación, puedan pensar en algún momento que la TH es algo resuelto y que lo único que hay que financiar son los desarrollos para aplicaciones concretas. Creo que muchos de nosotros somos conscientes de que esta sospecha ha planeado recientemente, y quizás siga planeando, sobre nuestro campo.

La defensa sólo puede nacer de una reflexión como la que fomenta este interesante taller sobre TH para que todos seamos conscientes tanto de los déficit de tecnología que aún detectamos como de las ramificaciones e intensificaciones que aún nos quedan por abordar (y muchas veces por definir).

Pretendo, por tanto, provocar esta reflexión exponiendo algunas ideas sobre qué hemos conseguido, qué estamos en camino de conseguir y qué queremos que se consiga. Una reflexión que espero interesante para los investigadores de las distintas disciplinas, que según iremos relatando, han aportado y deberán aportar a la TH.

2 Visión histórica

La TH nace, tras el estudio y consecuente conocimiento parcial del habla en humanos (explicaré más adelante por qué digo parcial y su relevancia en este discurso), con un interés en reproducir de manera artificial algunos de los procesos implicados.

Este interés se traduce, por ejemplo, en los primeros modelos mecánicos de síntesis de habla (Kempelen, 1791): un fuelle como generador de sobre-presión que atraviesa unas lengüetas a modo de cuerdas vocales seguidas de un conjunto de tubos deformables a voluntad

para reproducir fundamentalmente alófonos sonoros.

De manera más elaborada, y gracias a los conocimientos en tratamiento de la señal y a la aparición de las arquitecturas digitales de procesamiento, surge el modelo de Klatt hacia 1972 (Klatt, 1980), (Klatt, 1987) en que unas señales de excitación atraviesan un esquema de filtros para dar lugar a la primera voz sintética con un nivel mínimamente aceptable de inteligibilidad. Este modelo incentivó interesantes estudios sobre cómo hacer evolucionar en el tiempo los parámetros del sistema para conseguir una señal de voz de calidad razonable. De hecho, a raíz del trabajo con estas primeras voces artificiales se siente la necesidad de que la investigación en este campo sea multidisciplinar. Necesitan de expertos en lengua que les expliquen y les ayuden a sintetizar las unidades básicas, los alófonos, con los que construir palabras y frases. Necesitan conocer los detalles de la entonación y generar los correspondientes modelos adecuados. Todo ello llevará a la necesidad de analizar texto de entrada sin restricciones para manejar la información adecuada y así alimentar la evolución temporal de todos los parámetros del modelo de síntesis. De este modo apareció como uno de los grandes temas dentro de TH la conversión texto-voz.

Otro gran tema bajo el mismo título de TH, ya identificado también en los inicios, era el reconocimiento del habla. Los primeros esfuerzos se relacionaban muy cercanamente con el trabajo realizado en síntesis de habla. De los mismos modelos de producción utilizados en síntesis, se extrajeron las simplificaciones oportunas para obtener un conjunto de características con las que trabajar en análisis, en sistemas de reconocimiento (Itakura, 1975). Los primeros reconocedores de habla, allá por los años 50, fueron los denominados cuasi-fonéticos ya que intentaban reconocer diccionarios muy reducidos (típicamente, los diez dígitos) realizando un conjunto de medidas (Energía, cruces por cero, modelos LPC de bajo orden) con los cuales se pudieran escribir reglas simples que describieran el comportamiento de esas medidas en los alófonos que constituían las palabras objetivo.

El conocimiento sobre la señal de habla llevó a enfrentarse con uno de los problemas de fondo que todos los reconocedores de habla deben contemplar: las distintas repeticiones de las mismas elocuciones presentan temporizaciones de cada uno de los eventos acústicos muy diversas: aparece la necesidad del alineamiento dinámico temporal.

Más tarde, desde mediados de los años 60 y fundamentalmente desde los años 70, se empleó de manera más intensa el bagaje existente en reconocimiento de patrones para resolver el reconocimiento del habla (Levinson, Rabiner y Sondhi, 1983), (Rabiner, 1989). Los principiantes eran sistemas de alineamiento dinámico temporal que simplemente comparaban los vectores de características con patrones constituidos por otras secuencias de características almacenadas utilizando métricas sencillas. Los ganadores han sido finalmente las redes neuronales como clasificadores de eventos acústicos básicos y los modelos de Markov, aprovechando tanto sus potentes procedimientos de modelado acústico como de alineamiento dinámico temporal. El sistema insignia que reunía mucho de lo conocido de la aplicación de los modelos de Markov a reconocimiento a finales de los 80, era SPHINX de CMU (Lee, Hon y Reddy, 1990).

El trabajo en conversión texto-voz y en reconocimiento se acompañó de diversas temáticas muy relacionadas para formar el cuerpo de la TH durante muchos años. Me refiero a temáticas como percepción de habla, codificación de habla, adquisición y aprendizaje de habla y reconocimiento del locutor.

El planteamiento de esta primera época que finalizará hacia mediados de los años 80, incluía un porcentaje altísimo de esfuerzo puramente científico, ya que las aplicaciones reales de lo que se iba realizando eran muy limitadas cuando no prácticamente inexistentes. Sólo colectivos con una paciencia y necesidad fuera de lo normal como son las personas con discapacidad eran posibles objetivos para la realización de sistemas con algún sentido y esto provocó desde muy temprano el valioso acercamiento de la TH al mundo de la discapacidad.

El esfuerzo de algunos por convertir la TH en negocio en la época de los 80 solo fue parejo, en general, con el fracaso obtenido. Claro ejemplo de la desmesura de promesas frente a realidades fue el dictado automático. Diversas empresas, nacidas del esfuerzo de grupos investigadores reconocidos, intentaron vender la idea. ¿Conocéis mucha gente que utilice dictado automático en su oficina o en su casa? El error, en mi opinión, no sólo estuvo en

la poca madurez tecnológica del momento en que se realizaron estos esfuerzos, sino en que, además, se eligió una de las tareas más complicadas que se les podía haber ocurrido. Una alta complejidad debido tanto a la tarea en sí, que necesita de sistemas con vocabularios, gramáticas e información en general de dimensiones inabordables en aquel momento y de un fuerte carácter dinámico, como a que necesitan una interfaz con el usuario suficientemente evolucionada que disponga de la capacidad de diálogo suficiente para negociar las interpretaciones correctas y correcciones explícitas necesarias sobre el texto generado.

De todas formas, el esfuerzo no fue baldío. Poco a poco, todos aprendimos cuál es la realidad al llevar un sistema a pruebas de campo con usuarios reales y desde entonces hemos sentido la necesidad de realizar pruebas realistas de ergonomía a nuestros sistemas.

Posteriormente, desde mediados de los 80, y sumando la aportación de otros muchos investigadores, donde debemos mencionar especialmente a los relacionados con el procesamiento del lenguaje natural, conseguimos que bajo el manto de TH se sumaran temáticas que añadían sentido a todo el trabajo realizado. Aparecen las temáticas de comprensión de habla, generación de habla y, finalmente, control de diálogo.

Es en este punto donde han aparecido los primeros sistemas que producen rendimiento económico y, probablemente, los que producen en algunos la confusión de que todo parece estar hecho.

3 El presente

De esta manera, llegamos a la situación actual de lo que entendemos como TH. Relataré los puntos relevantes de la situación de las temáticas centrales en TH.

Disponemos de conversores texto-voz de inteligibilidad excelente, pero con problemas aún de naturalidad. La prueba de fuego, según hemos comentado antes, es enfrentar los productos con los usuarios y que éstos acepten un sistema para un uso continuado es aún muy difícil.

La tecnología dominante es la de concatenación de unidades de una base de datos que ofrece un gran repertorio y con sistemas de búsqueda de la mejor secuencia de unidades. Existen también esfuerzos, aún demasiado puntuales, por reproducir emociones en la voz. Dentro de esta temática también hay esfuerzos en cambio o generación de nuevas voces que no precisen de grabaciones extensas y generación de voces corporativas para entornos restringidos de muy alta calidad (naturalidad).

En reconocimiento de habla se han alcanzado tasas de reconocimiento razonablemente altas para vocabularios medios (algunos miles de palabras), pero nos queda mucho que decir en cuanto a la robustez de los sistemas ante las diversas condiciones de trabajo (independencia de locutor en sentido amplio, por ejemplo, con no nativos, canales limitados y posiblemente ruidosos, como la telefonía móvil, reconocimiento con micrófonos distantes que añaden reverberación y ruido, como en el coche o en entornos domésticos o profesionales, etc.). Los eventos de habla espontánea constituyen otro de los efectos que mayores pérdidas de precisión causan.

En comprensión de habla, los sistemas existentes trabajan siempre en entornos muy restringidos y, aún así, tienen una pobre cobertura del lenguaje espontáneo con el que se enfrentan en las tareas realistas.

En generación de habla aún no comprendemos bien la relación entre la relajación de expresiones necesaria para conseguir naturalidad y un cierto orden y estructura más fijos a la hora de presentar un conjunto de datos de manera que éstos sean fácilmente asimilables.

En control de diálogo existen algunas soluciones para dinamizar y adaptar la secuencia de intervenciones en función de medidas como la confianza sobre el reconocimiento o rudimentarios modelos del usuario que necesitan aún más trabajo para ser realmente eficientes.

Con la suma de estas tecnologías se han desarrollado sistemas que resuelven de manera razonable servicios automáticos en dominios restringidos, básicamente de recuperación de información. De todas formas, uno de los grandes problemas detectados es el elevadísimo coste de producción de estos sistemas, al implicar el trabajo manual en demasiados aspectos, trabajo que difícilmente es reutilizable en una aplicación distinta.

Estos sistemas necesitan de varias fases de acercamiento a la solución final: un primer diseño intuitivo del diálogo, procedente típicamente del análisis de diálogos reales de los servicios preexistentes que se desean automatizar, es seguido por la captura de

interacciones con sistemas tipo mago de Oz, prototipos en que algunos de los módulos del sistema aún no existen (por ejemplo, el reconocedor de habla y los generadores mensajes, al menos con la calidad final).

De las relaciones con los proveedores de estos servicios, a través de proyectos de automatización gracias a TH, surgió (espero que no de la población investigadora...) la idea de utilizar los magos de Oz no como herramienta de investigación, sino como producto real de alto rendimiento: un control de diálogo y una cadena de generación de mensajes interaccionan con el cliente, mientras que el reconocedor de habla es un humano (muchos humanos) transcribiendo ventanas de audio que les pasan sin que, de hecho, el humano conozca el sentido de reconocer ese ítem en ese instante. De esta manera, un único sistema informático, procesa varios diálogos en paralelo solicitando ayuda a los humanos en la parte difícil: el reconocimiento. La rentabilidad es aún superior dado que se procesan muchas peticiones en la misma unidad de tiempo ya que, al hablar con una máquina, los clientes no añaden mensajes irrelevantes a la consecución de la información. Estos sub-productos, a mi juicio no deseables, también consiguen su éxito y, lo que es peor aún, redundan de nuevo en que la población general piense que la TH ha evolucionado mucho más de lo que lo ha hecho en realidad.

4 Líneas de futuro tangibles

En el punto anterior de revisión del estado actual hemos incluido tácitamente las líneas de futuro a corto plazo. Mencionaré a continuación, sin ánimo de ser exhaustivo, algunas líneas de las temáticas centrales:

Conversión texto-voz

- Trabajo sobre la naturalidad. Hay que abordar el problema de que los mejores sintetizadores actuales aburren al oyente en breve tiempo.
- Diseño automatizado de variantes de voz con pocas muestras de la voz deseada.
- Expresión de estados de ánimo.

Reconocimiento de habla

- Trabajo sobre habla espontánea. Necesita de diccionarios y modelos de lenguaje dinámicos, incluso aprenderlos sobre la marcha.
- Robustez. Incluye todo tipo de trabajo para contender contra las diversas fuentes de variabilidad.

Comprensión de habla

- Construcción jerárquica de las fuentes de información para aumentar la posibilidad de reutilización de algunas partes comunes a todas las aplicaciones.
- Aprendizaje automático que minimice el coste de desarrollo.

Generación de habla

- Estudio sobre la adaptación entre los modelos y las expectativas de los usuarios que maximice la ergonomía que perciben.

Control de Diálogo

- Aumento del número de parámetros medidos y de las correspondientes influencias en el flujo del diálogo.
- Control de diálogo por objetivos y negociación que dinamice las intervenciones.
- Modularización en subdiálogos resueltos de manera satisfactoria y reutilizables.
- Aprendizaje automático de los modelos.

5 Líneas de futuro deseables

Por fin explicaré la frase en que decía "...tras el estudio y consecuente conocimiento *parcial* del habla en humanos..." y su implicación en mi discurso.

Para definir las líneas de futuro deseables, entendidas como a largo plazo, debemos admitir que necesitamos que siga ocurriendo lo que hemos relatado en la historia, que la definición de TH se amplíe cubriendo más disciplinas que doten de más posibilidades a este campo de investigación.

TH, en mi opinión, debe abarcar toda tecnología relacionada con el habla (hasta aquí, evidente). ¿Y qué puede ser esto que intentamos definir?.

El habla implica la existencia de dos seres conscientes que intercambian mensajes con un cierto propósito... Bueno, al menos eso fue hasta que comenzó la TH. Sin embargo, hoy en día, disponemos de sistemas automáticos (podemos convenir que no conscientes) que utilizan el habla para comunicarse con nosotros y darnos información o efectuar las operaciones

que deseamos. Por lo tanto, el mismo término "habla" ha evolucionado.

Y digo que tenemos un conocimiento parcial del habla humana no sólo por su evolución, sino porque desconocemos aún muchísimo de todo el proceso implicado. Pensemos en nuestro desconocimiento sobre los procesos mentales en nuestros cerebros y su relación con la consciencia (término dificilísimo simplemente de definir, aunque lo haya utilizado ya más arriba).

Creo que a la luz de estos pensamientos, el ideal de la temática que debe recoger la TH, se amplia de manera interesante. Por describir un objetivo que nos sirva de sumario de intenciones, deberíamos confluir en sistemas con los que pudiéramos dialogar para conseguir fines muy diversos. Estos sistemas deben saber extraer de nuestra comunicación con ellos la intención de la comunicación y respondernos adecuadamente de manera que se cubran nuestras necesidades.

Por lo tanto, estamos hablando de interfaces que utilizan el habla con potencia muy superior a la que podemos observar hoy en día, con relaciones con las aplicaciones que nos ofrezca nuestro ordenador de trabajo, con nuestros equipos domóticos, con internet, etc. Una característica que considero esencial es su capacidad continua de aprendizaje. Si utilizamos una expresión no conocida por el sistema, éste debe tener la capacidad de dialogar con nosotros para que le expliquemos su significado para que la admita la próxima vez.

Por supuesto, debemos añadir todas las características que esperamos de semejante ayudante: robustez frente al canal por el que nos dirijamos a él (por ejemplo, un micrófono distante o un teléfono), independencia del locutor para que admita que varias personas interaccionen (y colaboren) en la petición, capacidad de comprensión de los eventos de habla espontánea, etc.

Otros factores interesantes pueden ser la interacción, más íntima de lo que hoy podemos encontrar, con los mensajes gestuales. El gesto es también parte del lenguaje y debe ser considerado para detectar el locutor fuente del mensaje, para comprender su estado de ánimo, para observar sus indicaciones hacia algo concreto, etc. También el interfaz debería ser capaz de devolvernos información gestual, ya sea representada en un monitor o porque el interfaz sea un robot con aspecto semejante al nuestro y múltiples grados de libertad.

Si pensamos en todo lo que nos falta para este objetivo, tenemos definido estudio y trabajo para muchas personas y para mucho tiempo.

6 Referencias

W.V. Kempelen, 1791, Le mechanisme de la parole suivi de la description de un machine parlante, J.V. Degen, Vienna.

D.H. Klatt, 1980, Software for a cascade/parallel formant synthesizer, Journal of the Acoustical Society of America **67**, 971-995.

D.H. Klatt, 1987, Review of text-to-speech conversión for English. Journal of the Acoustiacl Society of America **82**, 737-793.

F. Itakura, 1975, Minimum prediction residual principle applied to speech recognition, IEEE Trans. ASSP **ASSP-23**, 67-72.

S.E. Levinson, L.R. Rabiner and M.M. Sondhi, 1983, An Introduction to the application of the theory of probabilistic functions of a Markov Process to automatic speech recognition. Bell System Technical Journal **62**, 1035-1074.

L.R. Rabiner, 1989, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**, 257-286.

K.-F. Lee, H.-W. Hon and R. Reddy, 1990, An overview of the SPHINX speech recognition system. IEEE Trans. ASSP **38**, 35-45.