Robots con mal acento: *Convivir con el habla sintética (2008)* Marc Böhlen

Artista e ingeniero, profesor de la Universidad de Buffalo, Estados Unidos. También conocido como RealTechSupport, diseña y construye sistemas de procesamiento de la información que reflexionan críticamente sobre la información como valor cultural a través de intervenciones robóticas especulativas. Sus proyectos cuestionan la relación entre personas y sistemas de automatización.

Traducción por Javiera Moncada, Colectivo Pliegue*

***COLABORA CON NOSOTR*S ***

En Pliegue trabajamos de manera voluntaria y colectiva, para que nuestros artículos y libros traducidos puedan ser compartidos libremente. Si deseas colaborar con nosotr*s, mejorar esta traducción o compartirnos material para ser liberado, escríbenos a produccion@pliegue.cl *Para conocer nuestro trabajo, traducciones, clases abiertas, cursos y canal de Youtube, síguenos en nuestro instagram @pliegue y Canal de Youtube Producciones Pliegue.

RESUMEN

Las tecnologías del habla sintética tienen un profundo impacto en nuestra forma de pensar e interactuar con los ordenadores. Este texto analiza las partes I y II del "Make Language Project", una trilogía sobre las consecuencias culturales del habla generada por máquinas como conducto para reconsiderar los prejuicios en la producción de habla sintética y el diseño de robots humanoides.

Cultura Androide normalizada

Nacida de la promesa de la revolución industrial de una vida de abundancia y ocio, la robótica está firmemente comprometida con la interpretación positiva y utópica de la tecnología tal y como la formularon por primera vez pensadores como Moore, Spencer, Saint-Simon, y reinterpretada en términos de tecnología informática en el siglo XX por la comunidad de la cibernética [1]. No todo el mundo compartía este punto de vista. El sociólogo Sorokin [2], por ejemplo, imaginaba que el avance humano a través de la tecnología acabaría en desastre. La extraña mezcla contradictoria de asombro, angustia y admiración con la que se perciben hoy en día los robots de gama alta es una prueba del vigor que siguen teniendo los puntos de vista polarizados; el panorama intelectual parece firmemente asentado con los ingenieros y científicos en el lado positivista, y los estudiosos de las humanidades y los artistas mayoritariamente en el lado pesimista, con algunos

estudiosos interesantes que sugieren un compromiso, por así decirlo, al afirmar que el futuro de la tecnología acabará en la inutilidad absoluta [3].

Desde Turing hasta Kurzweil, pasando por la cultura popular [4], la capacidad de recordar más y calcular más rápido se ha asociado directamente con la inteligencia sobrehumana. Dado que el ilusorio objetivo de una inteligencia superior no es factible en la práctica, los programas de investigación se han centrado en igualar la inteligencia y el comportamiento humanos en determinados ámbitos. No es de extrañar que incluso este objetivo menos elevado esté lejos de ser trivial. La visión computacional, por ejemplo, sigue luchando por lograr una percepción y un procesamiento visuales sintéticos equiparables a los del ser humano. Del mismo modo, el campo de la robótica humanoide no intenta actualmente hacer que las máquinas sean superiores a los humanos, sino que se ha centrado en dispositivos y procesos que imiten a los humanos. Curiosamente, esta noción de similitud o igualdad se define de formas muy específicas y de acuerdo con fuertes supuestos disciplinarios y objetivos retóricos. Por ejemplo, como han observado Nourbakhsh y otros, la mayoría de los robots se diseñan como mascotas o sirvientes [5], y todos son benévolos y educados [6]. Además, los diseñadores de robots humanoides y androides tienden a recrear la perfección física en sus productos. Ishiguro [7], por ejemplo, utilizó a una joven y atractiva moderadora de televisión como modelo para su androide más avanzado e incómodamente realista.

A pesar del atractivo inmediato, la belleza, la benevolencia y la cortesía son pautas de diseño de máquinas problemáticas. Normalizan la cultura androide y crean una base de simpatía hacia los robots que las máquinas no necesariamente merecen. Al normalizar la cultura androide, se pierden oportunidades para formas de interacción incómodas y problemáticas, pero potencialmente ricas y complejas. La cultura androide normalizada nos deja la promesa de una utopía amistosa que bien podría no cumplirse; promete un futuro que sólo es amistoso superficialmente, y nos deja sin preparación para afrontar los conflictos que probablemente surgirán con las máquinas sintientes en el futuro.

Interacción conflictiva

Normalizar las máquinas para que se comporten como los humanos en determinados contextos sociales limita el alcance de la investigación en el diseño de robots. También crea una base frágil y superficial para cualquier tipo de intercambio profundo entre robots y personas que la agenda de la robótica social pretenda abordar. Pero si se quieren conseguir intercambios profundos y a largo plazo entre sistemas sintéticos y personas reales, es esencial contar con una base más amplia de posibles formas de intercambio y maneras de compartir entre máquinas y personas. Los sistemas sintéticos, complejos, confusos y contradictorios como somos los humanos, serán, con el tiempo, mejores compañeros que los educados drones. En última instancia, el objetivo es la diversidad en el diseño de robots, una diversidad definida no sólo en términos técnicos, sino también por ideas variadas sobre lo que las máquinas podrían ser y lo que podemos compartir

con ellas. En este contexto, ofrezco algunas observaciones preliminares de la trilogía "Make Language" [8], partes I y II,
en las que el habla acentuada sintética y el lenguaje soez maquínico infringen las zonas de confort de la interacción con dispositivos informáticos.

Síntesis de texto a voz

El ser humano está especializado de forma única en la producción del habla, y sólo el homo sapiens puede utilizar la lengua, las mejillas, los labios y los dientes para producir 14 fonemas por segundo. Incluso los niños muestran una notable aptitud para reconocer los sonidos como habla. El habla nos convierte en criaturas únicas. La comunidad científica [9] y el saber popular consideran que el lenguaje es esencial para el ser humano. Dado que el lenguaje es tan esencial para el ser humano, su procesamiento se ha convertido en sinónimo de inteligencia sintética [10]. Si entendemos cómo procesan los humanos la información verbal, podremos construir máquinas inteligentes. Por eso conviene hacer un breve repaso de conceptos importantes del habla sintética.
La investigación sobre el habla sintética suele dividirse en dos categorías: Texto a voz y Reconocimiento automático del habla [11]. Text-to-speech (TTS) implica la creación de un patrón sonoro (voz) a partir de una entrada textual (palabras). El Reconocimiento Automático del Habla (ASR, por sus siglas en inglés), la inversa del TTS, consiste en convertir una entrada de voz arbitraria en un texto imprimible. Mientras que el campo del ASR y los conocidos y a menudo despreciados sistemas de dictado han tenido poco éxito en el mundo real, el TTS ha avanzado a pasos agigantados tanto en la investigación como en la práctica. El TTS combina las representaciones acústicas del habla basadas en el procesamiento de señales con el análisis lingüístico del texto para crear enunciados mecánicos que suenen como voces humanas. Los sistemas TTS suelen constar de varios componentes. Un componente de análisis de texto define y desambigua la entrada bruta. Localiza las pausas entre frases y párrafos. También se encarga de normalizar el texto (convertir abreviaturas y acrónimos en palabras completas). El resultado del módulo de análisis de texto se transmite al módulo de análisis fonético. Este módulo realiza, entre otras cosas, la importantísima conversión de grafema a fonema (letra a sonido). A su vez, la salida de este módulo se transmite al módulo de análisis prosódico. Éste se encarga de fijar los objetivos de duración y amplitud del tono para cada fonema. Por último, este resultado pasa al módulo de síntesis de voz, donde la cadena de símbolos construida se convierte en una salida audible que recuerda a una voz. Los diseñadores de TTS han experimentado con varios métodos de síntesis para este último módulo. Los enfoques más utilizados en la actualidad son la síntesis concatenativa y la síntesis de formantes [11]. En este caso, el enfoque concatenativo es de especial interés. A diferencia del método de formantes basado en reglas, la síntesis concatenativa se centra en los datos. Para construir un enunciado, un sistema TTS concatenativo dividiría la entrada en segmentos, buscaría las entradas correspondientes en una gran base de datos de grabaciones de un hablante humano real (locutor en la industria de la síntesis de voz) y, a continuación, concatenaría (añadiría en serie) las partes individuales para formar la salida

final. De este modo, se pueden reproducir incluso secuencias de sonido que no han sido grabadas. Los pasos de búsqueda, mapeo y filtrado incluidos en los sistemas concatenativos son elaborados y ofrecen un habla maquinal realista, sobre todo cuando se percibe a través de medios con poco ancho de banda, como el teléfono. Los sistemas concatenativos
avanzados incluyen técnicas de síntesis de selección de unidades que automatizan la laboriosa tarea de encontrar (manualmente) correspondencias, a grandes rasgos, entre grafemas y fonemas. A su vez, la síntesis de selección de unidades depende en gran medida de la clasificación automatizada, que suele implementarse en forma de redes neuronales diseñadas específicamente y cuyos detalles quedan fuera del alcance de este breve resumen.

Retorno de la palabra

Si unimos estos avances técnicos a la importancia universalmente reconocida de la lengua, resulta evidente que la tecnología de transmisión de texto tiene un interés primordial como fenómeno cultural. Nada menos que un resurgimiento de las tradiciones orales y una reevaluación del acto de hablar pueden esperarse a raíz de estos nuevos sistemas centrados en la voz. Desde el telégrafo, pasando por las tarjetas perforadas, hasta el teclado y la consola de juegos, los ordenadores han exigido que la gente se reuniera con ellos a través de torpes interfaces hápticas. TTS y ASR supondrán, literalmente, el fin de la era de la introducción manual de texto para las máquinas. Además, el TTS y el ASR redefinen la búsqueda de la "naturalidad" en las máquinas de un modo que otras tecnologías informáticas no consiguen. Las consecuencias de todo ello son de gran alcance, y este artículo sólo abordará algunas de ellas. Pero sí se puede afirmar lo siguiente: Las tecnologías del habla permitirán y exigirán nuevas definiciones en nuestras zonas de confort con las máquinas y, con ello, crearán nuevos problemas difíciles (tanto en el sentido informático de intratable como en el sentido cultural de multicapa) en el diseño de robots. La cuestión no es sólo académica. Mucha gente comparte la sensación de incomodidad cuando una suave voz de máquina le advierte repetidamente de que tenga cuidado al salir de la escalera mecánica o experimenta enfado cuando el menú de opciones que le ofrece la amable voz robótica de un servicio al que está llamando no se ajusta en lo más mínimo a su predicamento particular. E incluso cuando los robots aciertan, su tono de voz no suele ser el adecuado. En muchas lenguas, los enunciados tienen una firma prosódica más sencilla que las preguntas, en las que los patrones prosódicos varían mucho en función de la semántica de la propia pregunta. Por eso, las preguntas son mucho más difíciles de representar computacionalmente que los enunciados. En consecuencia, nuestras máquinas parlantes inteligentes están mejor preparadas para emitir órdenes que para hacer preguntas. Pero ya no hay vuelta atrás y el habla sintética nos obligará a replantearnos nuestra forma de expresarnos y cómo y cuándo los sistemas sintéticos deben imitar a los humanos. El hecho de que las máquinas puedan sonar como los humanos significa que deben utilizar el lenguaje del mismo modo

que las personas. Más allá del sabor de los enunciados, el habla sintética plantea la cuestión de qué podrían decirse las máquinas entre sí y qué deberían decirnos a nosotros. Proyectadas como amables y pacientes, tienen la capacidad de decir lo que necesitamos saber, pero también de repetir con insistencia lo que ya hemos oído o no queremos que nos digan. Vinculados por defecto a bases de datos y sistemas de información, estos agentes übercorrectos sin lengua materna han sido delegados a las funciones de oficinistas, instructores y supervisores. Pero parecen preparados para más.

Acentos e inmigrantes

Los actos de habla no sólo revelan la intención de un hablante, sino que también ofrecen información sobre su origen. Muchas personas que han nacido y crecido en una cultura y viven en otra distinta conservan restos audibles de su pasado en sus patrones de pronunciación. El habla acentuada es un habla particular por la forma en que su aromatización complica la transmisión de un mensaje [12]. Los prejuicios alteran la transmisión aparentemente neutra de un contenido. Dependiendo de las competencias lingüísticas (y la simpatía) del oyente, un acento puede hacer que un enunciado resulte encantador, ininteligible o incluso agresivo. El procesamiento digital de señales puede crear señales arbitrarias, la mayoría de las cuales no guardan ninguna relación con las que el ser humano es capaz de percibir. Con modelos elaborados de la geometría del tracto vocal humano, se pueden crear artificialmente todos los sonidos que el ser humano puede generar. A pesar de la universalidad de la infraestructura técnica, los sistemas TTS suelen diseñarse para aplicaciones muy específicas a lo largo de líneas de falla nacionales con fuentes de voz localizadas y entidades lingüísticamente identificables. Los vendedores comerciales de sistemas TTS suelen nombrar estas fuentes de voz según su origen lingüístico nacional, (pero no por el locutor individual que entregó las muestras de audio). Por ejemplo, hay Sarahs para el inglés estadounidense, Heathers para el inglés británico y Günthers para el alemán. La denominación deliberada de estas voces sintéticas contribuye a hacerlas más creíbles, transmite la agradable sensación de que hay una persona viva detrás de la emisión de audio digital y contribuye a la imagen de marca del producto. El conjunto de voces sintéticas del mercado representa un subconjunto depurado y controlado de lenguas humanas populares. No es de extrañar que los sistemas TTS comerciales no ofrezcan productos de voz con características "indeseables", como un habla arrastrada o, por ejemplo, un fuerte acento alemán.

Acentos Sintéticos Fuertes: Crear el lenguaje, parte I

El tono perfecto de la máquina desmiente el hecho de que ningún ser humano habla realmente sin acento, por leve que sea. Todos venimos de algún lugar y ese lugar impregna nuestras vidas y nuestras voces. Sólo la máquina puede hablar con un tono perfecto, sin ubicación ni historia. Hasta la fecha, los investigadores de TTS y ASR [13] se han interesado por el habla acentuada sobre todo por las dificultades que presenta para la inteligibilidad; es decir, cuándo un acento en un idioma determinado se convierte en un

obstáculo mensurable para transmitir un mensaje, una cuestión importante para las empresas de telecomunicaciones y los operadores de centros de llamadas.

Para comprender mejor las consecuencias culturales del habla sintética, estoy experimentando con acentos sintéticos fuertes. En este contexto, he creado un sistema de inglés estadounidense con acento alemán y otro de inglés estadounidense con acento mexicano y vocabulario limitado [8] basados en el motor de voz SVOX [13, 14]. Existen varios métodos para crear habla acentuada. Van desde la combinación de una voz de un idioma con un modelo lingüístico en un segundo idioma hasta la grabación de toda una base de datos de un hablante acentuado [15]. Incluso el método más sencillo, consistente en

introducir texto de la lengua A en un sintetizador de voz construido con un conjunto de fonemas y gramática de la lengua B, ofrece resultados útiles para algunos enunciados. Sin embargo, para generalizar esta mezcla de idiomas se necesitan métodos más elaborados. Entre ellos se encuentran los enfoques extraídos de los intentos de mejorar las transcripciones erróneas de los sistemas ASR utilizados para etiquetar las correspondencias entre grafemas y fonemas [16], la mezcla elaborada de conjuntos de fonemas de dos lenguas base, la mezcla de requisitos gramaticales, la modificación de la frecuencia, el volumen y los retardos en la transición de palabras y diversas reglas para las excepciones. Tanto el acento alemán-inglés como el español-inglés que he construido combinan todos estos procedimientos.

El éxito de este proceso depende de la proximidad de las lenguas en cuestión. Las lenguas indogermánicas, por ejemplo, se mezclan entre sí con mayor facilidad que con las eslavas debido a la similitud de los teléfonos base utilizados para su articulación. Sin embargo, el enfoque utilizado aquí es frágil y no puede manejar textos de uso general ni actos de habla con carga emocional [17]. Además, los mapeos gramaticales, a menudo incómodos, que los hablantes extranjeros construyen cuando hablan en una segunda lengua no son fáciles de incluir en una gramática formal. Los resultados más generales se obtendrían probablemente construyendo, sin vínculo con ninguna lengua base, un modelo lingüístico completamente nuevo basado en un habla acentuada concreta, tratándola de hecho como una lengua de pleno derecho. Esto permitiría construir cualquier tipo de discurso acentuado, incluidos aquellos que un hablante humano nunca se plantearía pronunciar.

Arrebatos Sintéticos: Crear el lenguaje, parte II

¿Podemos aprender algo de la mezcla de lenguas para el habla acentuada sintéticamente que nos ayude a imaginar cómo podríamos mezclar seres humanos y sintéticos? La imperfección, en toda su gran variación, podría ser un buen terreno de aprendizaje para ello. Y a partir de ahí podríamos imaginar mejor lo que los seres sintéticos tendrían que decir, una vez que se hayan liberado de anunciar horarios de vuelos y procesar la compra de entradas para conciertos. En consecuencia, podríamos plantearnos invertir más energía en hablar a las máquinas sobre nosotros mismos. Quizá podamos trasladar a las máquinas nuestra propia experiencia en la adquisición de idiomas. Cuando un ser

humano aprende una nueva lengua, suele adquirir una extraña mezcla de lo esencial y ejemplos de lenguaje soez. Esto es interesante, ya que el lenguaje soez elude la nueva lengua desconocida y conecta al hablante directamente con territorios conocidos [18]. Aunque culturalmente específico en cuanto a las condiciones límite que controlan su uso, el lenguaje soez nos vincula más directamente con nuestro cuerpo que otras formas de habla.

Fig. 1 Amy and Klara, 2006

Construcción de un potencial para los Arrebatos Sintéticos

Para experimentar con el lenguaje soez en los sistemas sintéticos, he construido un conjunto de desagradables agentes robóticos alojados en simpáticas cajas rosas llamadas Amy y Klara. Sus ontologías están formadas por y se limitan a leer y analizar trivialidades en línea de revistas de estilo de vida como Salon punto com. La creación de ontologías que reflejen el conocimiento común es una de las prioridades de la investigación en IA. Aquí no hay ninguna pretensión de universalidad o exhaustividad. Se trata de una epistemología prestada. Los robots buscan noticias en la web y agrupan los resultados por temas. Esta lista pasa a formar parte del mundo de Amy y Klara. Aunque los temas son significativos para los humanos, para los robots no son más que palabras. Sin embargo, estas palabras adquieren, con el tiempo, importancia por el hecho de repetirse. Cualquier tema que se mencione repetidamente recibe más peso computacional que los temas que sólo se encuentran una vez. Los temas que alcanzan un umbral crítico de importancia construida numéricamente se convierten en material de debate. Amy y Klara comparten sus resúmenes de texto ponderados estáticamente mediante TTS y ASR. Como ambos robots escanean la misma página web, cada uno espera que el otro le diga lo que ya sabe, y cuando esto no ocurre, surge la discordia y empiezan a insultarse. El ASR de cada robot tiene un vocabulario de lenguaje soez, dividido en escalas cada vez más agresivas de palabras malsonantes. Los robots fueron

entrenados específicamente para poder responder a este tipo de lenguaje que se filtra de los productos ASR. Además, tanto los resultados del reconocedor de voz como la transmisión física de los enunciados del orador al micrófono son propensos a errores; la falta de comunicación es inevitable y, con ello, los argumentos están prácticamente garantizados. El hecho de que Klara tenga un marcado acento alemán no hace sino aumentar la probabilidad de malentendidos. Hacer que el reconocedor sea menos selectivo crea más
casos de falsos positivos (reconocer falsamente un posible resultado válido) en cuyo caso los robots creen, por así decirlo, ofenderse por la mayoría de las palabras que perciben. No hay ninguna llamada directa para empezar a utilizar el lenguaje sucio en el programa de los robots. Los propios argumentos son una propiedad emergente de la configuración descrita. Una vez que uno de los robots emite una palabrota, el otro responde del mismo modo si reconoce la palabra. Cada robot está equipado con un conjunto de micrófonos que reducen el ruido y un pequeño pero dinámico sistema de altavoces-amplificadores equipados con ecualizadores. Las redes neuronales encargadas de hacer coincidir la entrada percibida con el corpus lingüístico aumentado con palabras malsonantes se entrenaron con este hardware en condiciones de poco ruido ambiente. El uso repetido de una palabra malsonante de la escala n conduce a la selección de una palabra malsonante de la escala n+1, siempre que la siguiente palabra se reconozca como malsonante en un plazo de tiempo determinado (de lo contrario, los niveles de agresividad retroceden). Dado que el reconocimiento y la pronunciación se producen en rápida sucesión, es posible tanto un intercambio de bajo nivel (cuando los resultados del reconocimiento son pobres) como una acalorada escalada de lucha, si los resultados del reconocimiento son positivos. Además, ambos robots están equipados con cámaras de vídeo y pueden verse mutuamente. Al igual que el sistema de voz, el sistema de visión está orientado al conflicto. Amy y Klara tienen una predisposición programática a que les moleste el color rosa, sin saber que ellas mismas son de color rosa, ya que la cámara de cada robot sólo puede ver blanco en el interior de su caja. Un algoritmo adaptativo de detección de tonos basado en histogramas [19] permite a los robots detectar el color rosa de la otra caja incluso en condiciones de iluminación variables. La propiedad a menudo idolatrada de la emergencia no se limita únicamente a las causas nobles.

Por supuesto, las condiciones límite pueden establecerse para evitar discusiones, en cuyo caso los robots se sientan tranquilamente uno al lado del otro. Sin embargo, es más interesante establecer las condiciones de forma que la mayoría de los episodios acaben en un intercambio de improperios que deje a la gente que observa a los dos robots preguntándose realmente por la inteligencia de las máquinas. En la web se puede ver un vídeo que documenta este intercambio de palabras malsonantes [8].

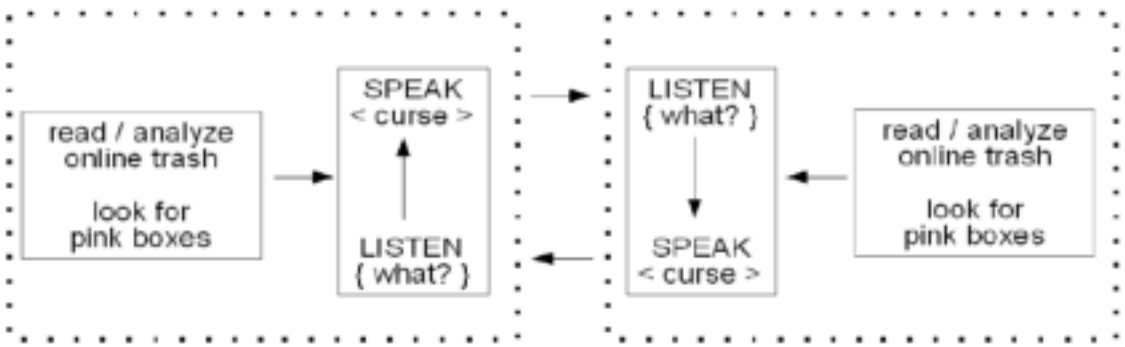

Fig. 2 constructing the potential for hissy fits in Amy and Klara

Aprendiendo de Amy y Klara

Amy y Klara y sus provocadores ataques de histeria sintética nos advierten de que no debemos esperar demasiado de las máquinas inteligentes. Contrarrestan la retórica de la máquina inteligente amable con una crítica de los usos normativos del habla sintética y la imperfección lingüística. ¿Podemos aprender algo sobre la mezcla de robots y personas mezclando lenguas en máquinas? Si el lenguaje soez está fuera de los límites de las máquinas, ¿qué pasa con otros tabúes? ¿Vamos a trasladar todos nuestros tabúes a los robots cuando parezcan, suenen u huelan como nosotros? Muchos tabúes lingüísticos derivan de tabúes religiosos, sexuales y de enfermedades mentales y físicas directamente relacionadas con las limitaciones físicas del ser humano. Como las máquinas carecen de nuestras funciones corporales, los tabúes correspondientes no tienen por qué mantenerse. Debemos esperar que algunos de nuestros tabúes sean invalidados por las máquinas. No debería sorprendernos que las máquinas inventen nuevas palabrotas propias de la experiencia de ser máquina y tener habla. Las lenguas de origen humano se verán alteradas y modificadas por su uso en máquinas de forma similar a como la cultura popular altera y se suma a los corpus de la lengua inglesa. Habrá nuevas figuras retóricas. Las lenguas que ya no utilicen los humanos podrán mantenerse vivas artificialmente en las máquinas. Y las personas que carezcan o hayan perdido la capacidad del habla podrían recuperarla a cambio de máquinas más pacientes que nosotros. Wolfgang von Kemplen, uno de los primeros investigadores experimentales del habla sintética, imaginó originalmente que su máquina parlante se utilizaría con fines terapéuticos [20].

Las consecuencias de las tecnologías avanzadas de tratamiento de la información nos harán cuestionar continuamente nuestras ideas preconcebidas sobre la inteligencia y nos desafiarán a reevaluar las formas en que nos relacionamos con las máquinas. Hay buenas razones para creer que la inteligencia humana, tal como la conocemos, depende del cuerpo humano, y que el lenguaje requiere precisamente ese cuerpo. Algunos investigadores de lingüística computacional han intentado demostrar este vínculo en simulaciones de la evolución del lenguaje en robots encarnados [21], y han obtenido

resultados sorprendentes, siempre que se acepten unas condiciones iniciales que permitan una semántica compartida [22] ab initio. Sin embargo, si esto no se da, el modelo fracasa. Por eso puede ser útil reconsiderar las decisiones inconscientes que tomamos basándonos en una visión antropocéntrica del lenguaje sintético.

El habla sintética es un buen ejemplo del tipo de dilemas que puede generar la mímesis perfecta. Aunque sólo sea por eso, los experimentos y observaciones aquí descritos pueden invitarnos a considerar dignas de atención la inteligencia y la cognición que no son propias del ser humano. Los dispositivos informáticos tienen la capacidad de "ser" de formas que los humanos no pueden. El funcionamiento interno de las máquinas y los dispositivos informáticos es muy diferente del nuestro en cuanto a material, construcción, escalas de tiempo y limitaciones biológicas. Ser máquina no es ser humano; más bien es una especie de ser extraño que no tiene ninguna relación con nuestras costumbres a menos que le obliguemos a comportarse como tal. Lo que se pierde en el enfoque mimético del diseño de robots es la oportunidad de comprometerse con la alteridad, por así decirlo, de la
máquina. A este respecto, sería aconsejable que los ingenieros estudiaran la historia de la representación pictórica y su lucha milenaria con la mímesis. Desde Zeuxis a Giotto, Leonardo y el Paragone de las Artes [23], pasando por Caravaggio, hasta la desaparición de la academia de arte tradicional a finales del siglo XIX, el realismo y la mímesis han sido principios de representación potentes y problemáticos en la civilización occidental. Sólo tras la guerra y los cataclismos sociales, las artes encontraron en la abstracción y el constructivismo nuevas vías de expresión no mimética.

La investigación del habla sintética es una empresa compleja que exige una atención rigurosa a los detalles. Por desgracia, también es otra víctima de la división del trabajo, por así decirlo, que se ha establecido entre las ciencias de la ingeniería y las humanidades y las artes. Sería un ejemplo más de la conocida y a menudo lamentada especialización disciplinaria si no tuviéramos que escuchar repetidamente sus consecuencias en los teléfonos y oírlas en los sistemas de navegación de los automóviles. Cómo sonarían y se comportarían las máquinas de voz si estuvieran informadas por la visión de Wittgenstein sobre el significado de las palabras que surge sólo de su uso [24], o los juegos de palabras elaboradamente coreografiados de Rose que empiezan con una charla que recuerda a una presentación académica que ha salido mal y acaban en una cacofonía de expresiones que suenan como palabras "reales" pero que no son más que balbuceos [25]. Imagínese que conocieran el poderoso tracto vocal de Blonk, sus habilidades linguales y su descaro para crear sonidos tan extraños que parecen inhumanos y a veces maquínicos [26], o al novelista Albahari, que en un trabajo reciente [27] conjeturó el número mínimo de palabras que uno necesita realmente para funcionar. Él cuenta cinco, siempre que uno se abstenga de hacer preguntas.

Agradecimientos

Este trabajo ha sido financiado en parte por una beca del Instituto de Humanidades de la Universidad de Buffalo. Gracias a SVOX por poner a disposición su motor de conversión

de texto en voz para experimentos extraños. Gracias también a FONIX SPEECH por su ayuda a la hora de hacer que su motor de reconocimiento de voz automatizado manejara con relativa facilidad los arrebatos de Amy y Klara.

Bibliografía

[1] Pickering, A., "Cybernetics and the mangle: Ashby, Beer y Pask", Social studies of science, 2002, vol. 32, no3, pp. 413-437.
[2] Sorokin, P., "Dinámica social y cultural: A Study of Change in Major Systems of Art, Truth, Ethics, Law and Social Relationships", 1957/1970, Boston: Extending Horizons Books, Porter Sargent Publishers.
[3] CAE, "La tecnología de lo inútil", 1994, en: CTHEORY, http://www.ctheory.net/articles.aspx?id=59
[4] Kurzweil. R. "La era de las máquinas inteligentes", MIT Press 1992
[5]. Fong, T., Nourbakhsh, I., y Dautenhahn, K., "A Survey of Socially Interactive Robots", Tech. Rep. CMU-RI-TR-02-29, Rob. Inst., CMU, 2002
[6] Simmons, R. Nakauchi, Y., "A Social Robot that Stands in Line", IROS, 2000. [7] Ishiguro, H., "Android Science", Cognitive Science Society, 2005.
[8] Böhlen, M., The Make Language Project, www.realtechsupport.org/new_works/ml.html.
[9] Nass, C., Gong, L., "Speech interfaces from an evolutionary perspective", Communications of the ACM, Vol. 43, Num 9 (2000), P 36-43
[10] Christiansen, M. y Kirby, S., "Language Evolution: ¿El problema más difícil de la ciencia?". En: Language Evolution: The States of the Art. Oxford University Press, 2003
[11] Schroeter, J. "Text to Speech Synthesis", en: Electrical Engineering Handbook, 3ª edición, capítulo 16, p. 1-13, AT&T Laboratories, 2005
[12] Ikeno, A., Pellom, B., Cer, D., Thornton, A., Brenier, J., Jurafsky, D., Ward, W., Byrne, W., "Issues in Recognition of Spanish-Accented Spontaneous English", en ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokio, Japón, abril de 2003. [13] Pfister B., Romsdorder H., "Mixed lingual text analysis for polyglot TTS synthesis", Eurospeech2003
[14] Pfister, B., "Skript zur Vorlesung Sprachverarbeitung I+II Abteilung fuer Elektrotechnik", ETH Zuerich, 2005
[15] Tomokioyo L., Black A., Lenzo K., "Foreign Accents in Synthetic Speech: Development and Evaluation", Interspeech2005
[16] Kim Y., Sydral A., Conkie, A., "Pronunciation Lexicon Adaptation for TTS Voice Building", AT&T Labs, en: InterSpeech2004 - ICSLP, Corea 2004.
[17] Schroder, M., "Emotional speech synthesis: A review". En Proceedings of Eurospeech 2001,. volumen 1, páginas 561-564, Aalborg, Dinamarca, 2001.
[18] Hughes, G., "Swearing: A Social History of Foul Language, Oaths and Profanity in English". Londres: Penguin Books, 1998
[19] Bradski, G.R., "Computer vision face tracking for use in a perceptual user interface". Intel Technol. J. 2(2), 1-15, 1998.
[20] von Kempelen, W., "Mechanismus der menschlichen Sprache nebst Beschreibung

einer sprechenden Maschine", 1791 y http://www.ling.su.se/staff/hartmut/kemplne.htm.
[21] Steels, L., Kaplan F., "Bootstrapping grounded word semantics", en: Briscoe, T., Linguistic evolution through language acquisition: formal and computational models, Cambridge University Press, 1999.
[22] Bickerton, D., "Language evolution: Una breve guía para lingüistas", Lingua 117 (2007), p. 510-526.
[23] Steinitz, K., "Leonardo da Vinci's Trattato della Pittura: Treatise on Painting. A Bibliography of the Printed Editions 1651-1956". Library Research Monographs. Vol. 5. Copenhague: Munksgaard, 1958.
[24] Wittgenstein, L., "Philosophische Untersuchungen" (1953), Suhrkamp 2003. [25] Rose, P., "Presiones del texto", vídeo de 17 minutos, 1983.
[26] Blonk, J., "Vocalor" en: "Labior -Phonetic Etude #3", publicado en Staalplaat, 1998 [27] Albahari D., "Fünf Wörter", (2004), Eichborn Verlag, 2005

Glosario

- **Habla sintética:** Habla creada o procesada por un ordenador.
- **TTS:** Texto a voz. El proceso de convertir palabras escritas en voz audible. • **ASR:** Reconocimiento Automático del Habla. Proceso de convertir en texto la señal de audio recibida.
- **Fonema:** La unidad contrastiva más pequeña del sistema sonoro de una lengua, independiente de su posición en la palabra o frase.
- **Grafema:** El equivalente del fonema en los sistemas escritos.
- **Prosodia:** información relacionada con el habla (entonación, tono, duración, gestos) que no está contenida en el propio texto. La prosodia suele considerarse un canal de comunicación paralelo que contiene información que complementa o contrasta la del canal principal.
- **Síntesis concatenada:** Síntesis basada en la concatenación (o encadenamiento) de segmentos de habla grabada. Este método suele producir el habla sintetizada con un sonido más natural, pero requiere una amplia base de datos de muestras de habla grabadas y etiquetadas.
- **Síntesis de formantes:** La síntesis de formantes no utiliza muestras de habla humana en tiempo de ejecución. En su lugar, la salida del habla se crea utilizando un modelo acústico en el que parámetros como la frecuencia fundamental (f0) y la vocalización varían con el tiempo para crear una forma de onda del habla artificial. Los sintetizadores de formantes suelen producir expresiones más "robóticas", pero suelen ocupar menos espacio que los sistemas concatenativos, ya que carecen de una base de datos de muestras de habla.
- **Detección adaptativa del tono del histograma:** Proceso que consiste en dividir el espectro de color completo de una imagen en una serie de intervalos definidos y utilizar los resultados para establecer las condiciones límite del histograma en una imagen posterior, de forma que se pueda seguir el intervalo/color deseado a lo largo del tiempo.

Información biográfica

Marc Böhlen es Profesor Asociado en el Departamento de Estudio de los Medios de Comunicación de la Universidad de Buffalo y Profesor Visitante en el AI-LAB de la Universidad de Zurich. Lleva desde 1996 contribuyendo a la diversidad en la cultura de las máquinas.

Figuras

Figura 1: Los ordenadores de Amy y Klara están sincronizados (de ahí el gran reloj rosa). Esto les permite salir simultáneamente de sus respectivas cajas. Cada caja mide 10 x 10 x 10 pulgadas. Los robots son de aluminio y plástico. Cada robot está equipado con un micrófono, un altavoz y una cámara.

Figura 2: Esquema de la arquitectura de software que permite los altercados entre Amy y Klara.