¿POR QUE NON NOS COMPRENDEN AS MÁQUINAS? UNHA APROXIMACIÓN CRÍTICA ÁS TECNOLOXÍAS DA FALA

Francisco J. Valverde Albacete
Universidade Carlos III
Madrid

1. INTRODUCCIÓN

¿Ten intentado algunha vez dictarlle a un ordenador e que reflicta que vostede non di máis que parvadas agramaticais? ¿Atopouse nalgún servicio telefónico cunha voz metálica ou que renxe e lle di con voz cansa 'gracias por chamar'? Quizais vostede abandonase á súa sorte unha voz excesivamente amable que lle esixiu un cento de veces que respondera 'si' ou 'non' logo do asubío.

En tal caso, vostede foi víctima do fracaso das tecnoloxías da fala en conseguir un servicio amigable, robusto e útil e se cadra interésalle sabe-lo porqué deste fracaso. Se segue lendo atopará primeiro cómo situar estas tecnoloxías dentro do marco da enxeñería lingüística e da lingüística, a continuación algúns exemplos de tecnoloxía da fala e finalmente unha análise de qué podemos esperar realmente de tales tecnoloxías.

2. CARACTERIZACIÓN DA ENXEÑERÍA LINGÜÍSTICA

En realidade, as tecnoloxías da fala son parte dun programa de actuación máis ambicioso coñecido como 'enxeñería da linguaxe' ou 'enxeñería lingüística', que é o lugar común de encontro da lingüística teórica, a lingüística tradicional, a investigación sobre interfaces home-máquina e un certo espírito de enxeñería non alleo á expectativa de gañar cartos rapidamente.

1.1 DIFERENCIAS ENTRE 'LINGÜÍSTICA', 'LINGÜÍSTICA COMPUTACIONAL' E 'ENXEÑERÍA LINGÜÍSTICA'

Concibi-lo estudio da fala, ou máis xenericamente o da lingua, baixo o prisma da enxeñería lingüística supón dar un xiro radical con respecto ó da lingüística; por un lado, implica cambia-la motivación científica desta última por outra manipuladora e mercantil: non se persegue a consecución dunha descrición adecuada do

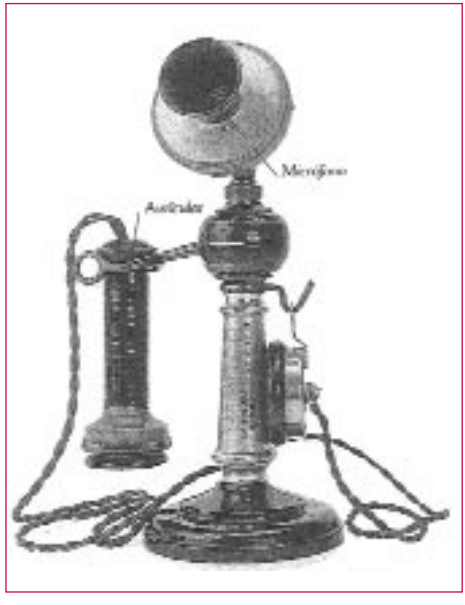
fenómeno lingüístico, senón o uso da lingua para expandi-las capacidades comunicativas humanas ou fomenta-la súa necesidade de adquirir bens e servicios, formulación, dito sexa de paso, compartida co resto das enxeñerías.

Por outro lado, no teórico supón o cambio de 'modelos de descrición da competencia lingüística', o que mellor caracteriza as teorías lingüísticas clásicas, a 'modelos de execución da actuación lingüística'. A técnica na que se propoñen e falsean estes modelos da actuación é sempre a simulación mediante programas informáticos, e é un punto de vista que emanou da lingüística computacional, disciplina que a penas ten corenta ou cincuenta anos.

A diferencia da lingüística computacional, sen embargo, a enxeñería lingüística pretende que os desenvolvementos computacionais teñan unha robustez e facilidade de uso que lles permita converterse en 'servicios' para a maior parte dos falantes dunha lingua. Esta é a distinción que comunmente se acepta entre prototipos e productos: os resultados da lingüística computacional case nunca abandonan o estado de prototipos para se converter en productos.

A enxeñería lingüística, aínda que de novo cuño, leva exercéndose, como mínimo, desde a fin do século pasado, cando se popularizou o uso do teléfono, que pasou a se-lo primeiro aparello e servicio de enxeñería lingüística: filtra o espectro da nosa voz e intenta que a comunicación teña unha calidade mínima para ofrecer un servicio aceptable a pesar dos problemas tecnolóxicos da transmisión de sinais a distancia. O importante da tecnoloxía telefónica analóxica de entón foi a súa efectividade a un prezo razoable; de non ter sido así, hoxe non usariámo-lo teléfono.

Emporiso, aparellada ó cambio de prototipo a producto hai unha vulgarización, unha desacralización da competencia lingüística modelada que se completa cando a dita competencia, transformada en servicio, é aceptada socialmente. Tal foi o camiño que seguiron algúns productos lingüísticos de primeira xeración, como os correctores ortográficos, que hoxe están presentes no noso labor cotián de escritores.



Teléfono de 1812. O teléfono foi o primeiro aparello para transmiti-la voz a distancia.

Posteriormente, mesmo existe un certo desprezo polas capacidades de tales servicios, que normalmente non se corresponde obxectivamente coas prestacións que proporcionan: por exemplo, é un lugar común queixarse da mala asistencia das empresas provedoras de servicio telefónico básico; outro exemplo: a miúdo protestamos sobre as indicacións que os correctores ortográficos nos ofrecen ou sobre a liberdade coa que corrixen as nosas maiúsculas ou faltas de concordancia de xénero, pero a verdade é que o número de erros que a maior parte de nós comete nun documento rematado descendeu notablemente.

Esta énfase en mellora-la calidade para a maior proporción de usuarios posible tamén é unha característica da enxeñería lingüística e recóllese baixo o amplo termo de 'aceptabilidade da técnica', no sentido de que acrecenta as posibilidades de que un servicio ou producto sexa aceptado polo maior número posible de usuarios na maior gama posible de contextos de uso; un exemplo: aínda hai xente que se nega a dar unha noticia importante por teléfono; outro: para os poucos españois activos na defensa dunha ortografía diferente á que propoñen os principais programas de tratamento de texto, a autocorrección destes últimos é unha molestia continua. Todos estes problemas se complican coas tecnoloxías menos robustas que comezan a aparecer, como os sistemas accionados por voz ou os sintetizadores de voz entrecortada e metálica.

Finalmente, o desprezo olímpico pola opinión dos humanos que se observa na enunciación das teorías lingüísticas clásicas —que non son amigables nin para os aprendices de lingüistas, lembrémo-la nosa propia formación— e ata certo punto nos prototipos da lingüística computacional —interesados custe o que custe nas 'probas de concepto'—, é completamente alleo á enxeñería lingüística, por un motivo ben claro: se as persoas non aceptan un servicio de enxeñería lingüística, este non reportará beneficios, sendo o caso que moi poucas aceptan servicios non amigables e robustos, é dicir que funcionen a maior parte do tempo, no maior número de ambientes acústicos posibles, coa maior facilidade de uso. O usuario humano é, ó cabo, quen di se determinado servicio ou producto é un éxito ou non, e a súa opinión dun servicio é a vara de medida definitiva.

Sirvan estas pinceladas para declara-la radical diferencia entre a lingüística —xa sexa na súa vertente máis teórica ou na baseada na experimentación— e a enxeñería lingüística. Agora interésanos máis o programa da enxeñería lingüística que é asumido polas tecnoloxías da fala.

1.2 OBXECTIVOS DA ENXEÑERÍA LINGÜÍSTICA A MODO DE LISTA DE PROMESAS INCUMPRIDAS

En resumo, co fin de ofrecer servicios adecuados ós usuarios, a enxeñería lingüística formúlase os seguintes obxectivos tecnolóxicos inmediatos (recollidos de LRE, 1999):

— recoñece-la escritura manual e a voz en varias linguas para todo tipo de usuarios;

— comprende-la linguaxe humana e traducir entre varias linguas nunha amplitude de contextos referenciais;

— axuda-los humanos a compren dérense mellor entre si;

— ofrecer resultados orais e impresos en calquera lingua.

Estes obxectivos afectan a nosa formación como persoas ou falantes, a nosa calidade de vida e a nosa economía, fundamentalmente, pero na medida en que se manifestan como realidades inmediatas na nosa vida, tamén a conforman grandemente, por exemplo, alterando os nosos hábitos —como ocorreu coa comunicación telefónica móbil que invade a vida pública—.

Cómo terían de transformarse estes obxectivos tecnolóxicos en realidades aproveitables tamén parece claro; unha adecuada tecnoloxía lingüística debería permitirnos:

— acceder con eficacia á información que necesitamos;

— comunicarnos doadamente e de xeito estándar cos ordenadores na casa, o traballo, os lugares públi cos, etc.;

— facer negocios facilmente a través de calquera medio de comunicación que teña unha canle lingüística, en particular a través do teléfono, xa que aquí a auditiva é a única vía dispoñible;

— funcionar con máis eficacia no ámbito internacional, superando as barreiras dos idiomas;

— etc.

Todos estes obxectivos pódense reunir noutro máis xenérico e tamén máis prosaico, no que o beneficio económico está implícito: proporcionar unha gama máis ampla de servicios mellores ó máximo número de cidadáns, compañeiros e clientes. Hai que aclarar que os principais investidores en tecnoloxías da linguaxe e da fala son os operadores de telecomunicacións, sustentados firmemente no financiamento dos gobernos —notablemente os dos países europeos, os Estados Unidos e Xapón— e a Comisión Europea, así que calquera declaración oficial de obxectivos ha serlles grata.

Paradoxalmente, cando se asumen obxectivos económicos en actividades de investigación, o fallo na consecución de servicios ou productos que se convertan en éxitos económicos convértese nun fallo na investigación.

1.3 AS TECNOLOXÍAS DA FALA COMO TÉCNICAS DE ENXEÑERÍA LINGÜÍSTICA

Vémonos obrigados a restrinxi-lo noso discurso, a partir deste momento, ó ámbito das tecnoloxías da fala: o campo ó que renunciamos, o das tecnoloxías da linguaxe escrita, ten problemas de seu e unha historia moito máis dilatada. Unha boa introducción ás

técnicas modernas da linguaxe escrita é a de Manning e Schütze (1999).

Tecnoloxías da fala son aquelas enmarcadas na enxeñería lingüística que teñen como soporte ou modalidade sensorial exclusivamente a fala. É difícil que a linguaxe escrita non apareza de forma periférica nelas, dada a omnipresencia da forma escrita como rexistro, pero que se usa entón meramente como unha encarnación da linguaxe adecuada tanto para os humanos como para os ordenadores, así que non 'contamina' con tecnoloxías da linguaxe escrita os procedementos e métodos da fala.

Tódalas tecnoloxías da fala usan primordialmente as técnicas de procesamento de sinal aplicadas ó sinal vocal como ferramentas fundamentais, entre as que se atopan técnicas estrictamente de procesado de sinal —como a análise espectral ou en calquera outro dominio adecuado—, técnicas de procesos estocásticos e teoría da probabilidade, técnicas do recoñecemento estatístico de patróns e conceptos derivados da teoría da información, sobre todo o de 'entropía'. No libro básico de Deller, Proakis e Hansen (1987) pódense encontrar cadansúa introducción neste contexto.

1.4 A VALIDACIÓN DE TECNOLOXÍAS MEDIANTE TAREFAS

As tecnoloxías da fala —en realidade tódalas encadradas na enxeñería da linguaxe— deben ser validadas, é dicir, débese avalia-la súa idoneidade para modelar determinadas competencias lingüísticas humanas, a ser posible cunha medida obxectiva. Para realizar tal aparente desatino propúxose a 'validación mediante tarefas', sobre a que se pode encontrar unha descrición extensa e non exenta de crítica no capítulo 13 de Varile e Zampolli (1995).

En esencia, definir unha tarefa de validación consiste, primeiramente, en definir un corpus que reflicta a actuación en determinada competencia lingüística, logo definir un criterio abstracto do desempeño na dita actuación, definir unha medida dese criterio e prescribir un método para realizar tal medida. O carácter de 'medida' entraña o feito de que tódalas medidas tomadas co mesmo método sexan comparables, e por iso os sistemas que se ateñan a este protocolo poden ser directamente comparados.

A favor da filosofía da validación mediante tarefas hai que dicir que permite a obxectivación dos resultados de investigación tecnolóxica mediante a comparación pública e a competición entre laboratorios e empresas; tamén que todo isto redunda en compartir resultados e infraestructuras —bases de datos, *software* ou plataformas de desenvolvemento comúns—. Antes desta era das tarefas de validación, poucos resultados eran usados polo conxunto das comunidades de lingüistas ou enxeñeiros da fala; numerosas propostas teóricas nunca eran cuestionadas mediante a experimentación rigorosa, e, consecuentemente, a maior parte dos prototipos fallaban ó intentar transformalos en sistemas ou servicios.

En contra desta filosofía, hai que denunciar unha orientación da investigación cara á competición pública (que implica a repartición dos fondos públicos de investigación e o prestixio na comunidade) e o feito de que estas competicións obrigan a desviar esforzos que doutra maneira poderían terse dedicado a facer avanza-las investigacións. Non obstante, quizais a validación de resultados sexa parte integrante da investigación técnica que ata o de agora tiñamos arredada.

No apartado meramente metodolóxico, unha medida agregada como a que ofrece a validación por tarefas pode non ser suficiente para analiza-lo comportamento do sistema en caso de que se queira mellorar, porque non proporciona unha explicación detallada de qué hai que mellorar.

No sentido político, sendo a principal patrocinadora da validación mediante tarefas unha axencia de financiamento norteamericana, o idioma claramente favorecido por este esquema é o inglés, no sentido de que a maioría das tarefas se definen nesta lingua e, posteriormente, se acaso, tradúcense a outros idiomas; os recursos están primeiramente en inglés e logo en toda outra lingua que se considere economicamente rendible. Polo momento, estas linguas son o francés, o castelán, o alemán, o xaponés e o chinés mandarín (con desigual fortuna en canto á abundancia de materiais en cada unha delas), aínda que outras linguas de cultura pugnan por estar neste grupo.

En conxunto, baixo a influencia das tarefas de validación, as tecnoloxías da fala —e en realidade tódalas industrias da lingua— avanzaron máis cara ás tecnoloxías máis robustas e aceptables que durante o "longo período de escuridade" precedente. O noso coñecemento científico sobre as realidades lingüísticas latentes en cada técnica tamén se viu esporeado, ampliado ou clarificado, aínda que a nosa sospeita a título persoal é que en máis dunha tecnoloxía están considerándose solucións *ad hoc* no canto de descricións adecuadas dos fenómenos lingüísticos.

2. BREVE INTRODUCCIÓN A ALGUNHAS TECNOLOXÍAS DA FALA

A continuación, na figura, represéntase o tradicional esquema comunicativo da linguaxe, coa énfase posta en cada un dos elos nos que se aplica unha das tecnoloxías da fala.

A mostraxe de tecnoloxías que imos presentar non é significativa: serán unha técnica básica de entrada de datos —o recoñecemento da fala—, unha técnica básica de presentación de datos —a síntese— e un uso especializado da técnica xenérica de recoñecemento —o recoñecemento de locutores—.

No tinteiro deixamos un número de tecnoloxías importantes: a comprensión da fala, a traducción, a recuperación de información falada, cuestións de multilingüismo —a

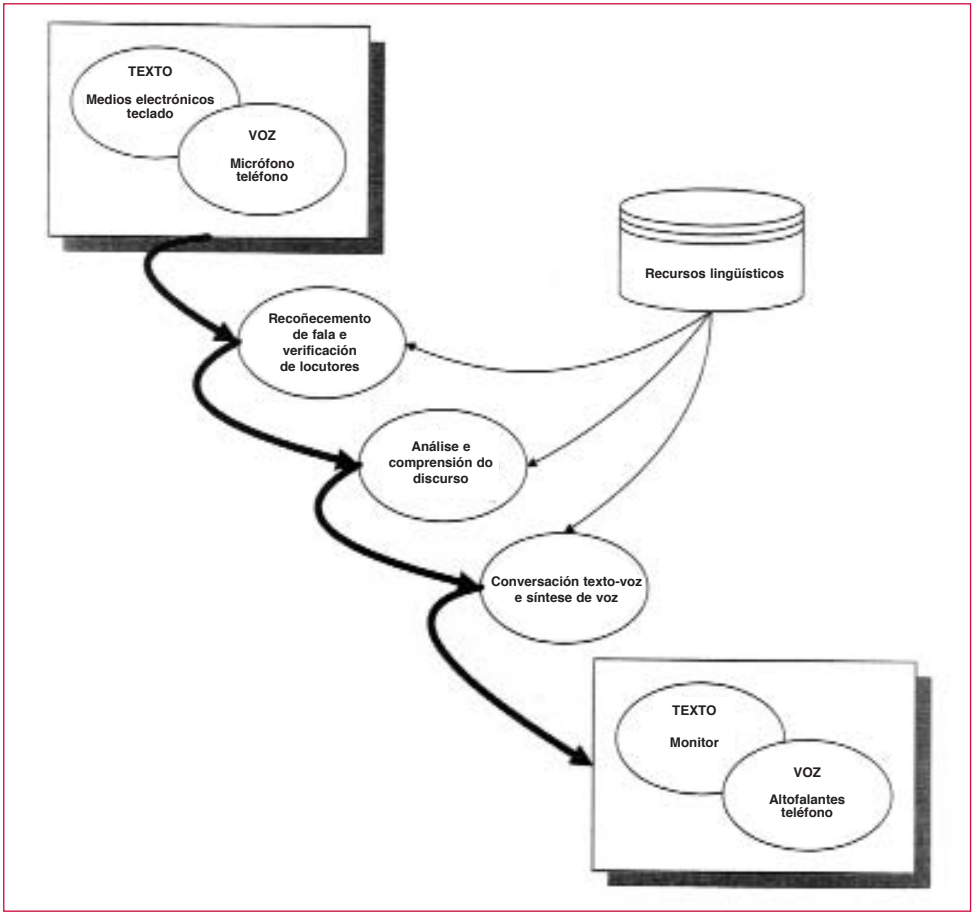


Fig. 1: Diagrama da inserción das tecnoloxías da fala na comunicación lingüística.

coexistencia de varias linguas na mesma aplicación de fala— e a multimodalidade —a coexistencia de varias modalidades sensoriais de interacción: vocal, visual, táctil, etc.—, ou os sistemas de diálogo —no que tódalas técnicas conflúen para crea-lo espellismo dun falante cibernético completo—, ben porque son aplicacións específicas das tecnoloxías que presentamos ou ben porque inclúen elementos que transcenden o ámbito da fala e non se poderían presentar ben sen eles. Referímo-lo lector á enciclopedia de Varile e Zampolli (1995) para unha presentación moito máis ampla das tecnoloxías da lingua.

Debemos confesar, sen embargo, a deliberada omisión das tecnoloxías

de codificación da fala: por un lado, obxectivamente falando, o seu uso é tan común (como demostra a telefonía dixital fixa e móbil, os CD-ROM ou os contestadores telefónicos nas centrais) que non hai posibilidade xa de entusiasmar ó público cos seus logros, e a penas os organismos de financiamento; por outro lado —e menos obxectivamente—, a historia da codificación ata o momento é case un éxito continuo, e desmerecería o ton deliberadamente crítico e pesimista deste artigo.

Antes de repasa-las tecnoloxías particulares, queremos dar un bosquexo de qué estamos intentando conseguir coas diferentes tecnoloxías, e para isto necesitamos, sobre todo, caracteriza-la fonación e a audición da voz humana.

2.1 CARACTERIZACIÓN DA VOZ

O modelo máis común da producción de voz é o dunha secuencia de tubos ríxidos ou deformables conectados a continuación uns doutros (larinxe, farinxe e cavidade bucal) ou en paralelo (cavidade nasal), cun pistón que aspira ou expira o aire (os pulmóns) situado nun extremo e o outro extremo aberto á atmosfera exterior (a abertura bucal). Nos tubos hai estreitamentos (cordas vocais, dentes, beizos) e válvulas (veo do padal e lingua) que permiten ou ben xerar unha vibración case harmónica no aire, como as cordas vocais, ou ben modula-lo fluxo de aire que pasa por ou preto deles, como o resto dos articuladores.

Este sistema funciona normalmente ó exhalar aire, mentres se dispoñen as cordas vocais na configuración adecuada para que o fluxo aéreo as faga vibrar e xere unha onda de presión. Asemade, acciónanse os articuladores para modular este fluxo aéreo resaltando unhas propiedades da onda e eclipsando outras.

Esa onda de presión aérea recíbese no pavillón auricular e é conducida a través do oído medio cara ó tímpano, onde se dá a transducción a vibracións mecánicas dun sistema de pancas óseas conectadas a un recipiente, a cóclea, onde acaba como onda de presión nun fluído acuoso. A cóclea ten a configuración adecuada para detecta-las acumulacións de enerxía en determinadas frecuencias e traducilas a enerxía electroquímica que se procesa e transporta ó cerebro, onde o resto da audición ten lugar. Parece ser que o nervio auditivo tamén realiza certo procesamento no seu camiño ó córtex.

Neste traxecto o sinal de voz sufriu unha multitude de filtrados, atenuacións, amplificacións, transduccións de medio e procesamentos que cómpre simular ou repetir en calquera sistema que intente imita-las capacidades fonadoras ou auditivas humanas, cunha dificultade engadida: ten que usar un soporte semiconductor electrónico, así que debe facer unha transducción eléctrica que substitúa as anteriores.

Cando se traballa con ordenadores, a transducción entre onda de

presión e onda eléctrica realízaa un micrófono, e a inversa realízaa un altofalante; pero sendo ámbolos dispositivos pasivos, o único que poden facer é distorsiona-lo sinal de voz, co que empeora a comunicación. O procesamento de sinais é o estudio das técnicas para evita-la degradación de sinais, como a voz, a través dos medios de comunicación e os seus terminais, como o micrófono e o altofalante.

A voz é, pois, unha onda de presión que durante certos intervalos temporais é periódica ou case periódica, é dicir, repite a súa configuración, e entón dicimos que é voz sonora, e durante outra parte do tempo é aperiódica e entón dicimos que é voz xorda. É máis doado ve-la estructura periódica e aperiódica da voz no espectro de potencia, que é unha imaxe de cómo está repartida a enerxía do sinal nas súas frecuencias constitutivas: o espectro periódico puro ten unha representación coma un peite onde os dentes representan as frecuencias nas que hai enerxía e a súa intensidade, e os ocos aquelas nas que non hai enerxía, mentres que nun espectro aperiódico hai enerxía en tódalas frecuencias. No espectro de potencia da voz —que por veces é case-periódico— pódense ver moitas das características dos sons da fala: cun pouco de práctica ata é posible 'ler' nun espectrograma (unha transcrición do espectro) os fonemas que foron producidos.

O espectro de frecuencia da voz só chega ata 7000-8000 Hz (ciclos por segundo), así que toda a información que porta está contida nesta anchura de banda. Sen embargo, se filtrámo-la voz deixando pasar só as frecuencias entre os 100 e os 3400 Hz aínda seguimos comprendendo o que se nos di. Como é só esta banda de frecuencias a que nos permite oír un teléfono, ás veces a voz así filtrada denomínase 'de calidade telefónica'.

O problema da caracterización da fala dun locutor, e non digamos xa a da fala de tódolos posibles locutores, é a súa extrema variabilidade: nin sequera o mesmo falante na mesma situación para- e extralingüística, enunciando a mesma secuencia de palabras produce idénticos sons. Se para nós é difícil distingui-la sutileza de diversos acentos nun idioma que non é o materno, ou ata nos custa certo esforzo acostumarnos a acentos diferentes da nosa propia lingua, ¿canto máis difícil non será para unha máquina poder distinguir ou identificar sons? Ademais, calquera fenómeno de variabilidade está superposto ó da omnipresencia do ruído na nosa comunicación: diferentes falantes, máquinas ou fenómenos naturais conspiran contra o correcto funcionamento da canle de voz, pero tamén interveñen nos procesos electrónicos que levan esta voz ata as unidades centrais de proceso dos ordenadores, complicando o problema. A loita contra o ruído e a abstracción ou a introducción da variabilidade da voz humana —no caso do recoñecemento e no da síntese de fala, respectivamente— forman o núcleo do programa das tecnoloxías da fala.

2.2 RECOÑECEMENTO DA FALA

Elixímo-lo recoñecemento da fala para explicar a fondo unha tecnoloxía por se-la que máis se desenvolveu nos últimos corenta anos e a que está máis preto quizais de cumpri-los seus obxectivos. Seguiremos para a súa presentación a Ney e Ortmann (1999).

A TAREFA DE RECOÑECEMENTO DA FALA

Polo título deste epígrafe podíase pensar que o obxectivo desta particular tecnoloxía é modelar unha tarefa similar á dos dictados que se realizan nas nosas escolas; sen embargo, por 'recoñecemento da fala' enténdese hoxe a capacidade de transcribi-las palabras pronunciadas por un falante calquera sen intentar comprendelas. Non obstante, loxicamente esta comprensión é crucial para o dictado á hora de distinguir homónimos; así pois, esta tarefa defectuosa non é en absoluto o mesmo que tentar reproduci-la capacidade semasiolóxica do falante humano, é dicir, a súa capacidade de descodifica-la voz e comprende-la súa mensaxe.

Outra tarefa que si procura obter un significado das palabras pronunciadas deuse en chamar de 'comprensión da fala', pero, de novo, nela podemos permitirnos ignorar algunhas palabras, relacións gramaticais e incluso frases redundantes con tal de atinar co significado xeral do texto enunciado. É dicir, nestoutra tarefa a transcrición acústico-ortográfica é secundaria.

Parece que ámbalas tarefas son complementarias e que unha combinación axeitada das dúas técnicas debería permitir modelar un dictado a humanos, pero tales integracións son sempre difíciles e existe unha certa inercia a seguir coas mesmas tarefas que proporcionaron notables avances á tecnoloxía.

Para recoñecer, pois, unha elocución, partimos dunha secuencia de observacións acústicas $x_1^T \equiv x_1 \cdots x_T$ obtidas a partir do sinal vocal captado cun micrófono e seleccionando as características que teñan correlacións físico-acústicas máis prometedoras, normalmente a composición de frecuencia da voz (un concepto radicalmente diferente ó de 'realización dun fonema', que pasa por se-la unidade mínima de descrición en fonética).

O obxectivo da tarefa de recoñecemento é obte-la secuencia de palabras que foi pronunciada $w_1^N \equiv w_1 \cdots w_N$, para o cal a teoría estatística de decisión dinos que temos que levar a cabo un proceso que maximice a probabilidade de que se 'pronunciaran' as palabras anteriores xa que se 'oíron' realmente as observacións: é dicir, temos que calcula-lo máximo de $P\left(w_1^N \middle| x_1^T\right)$.

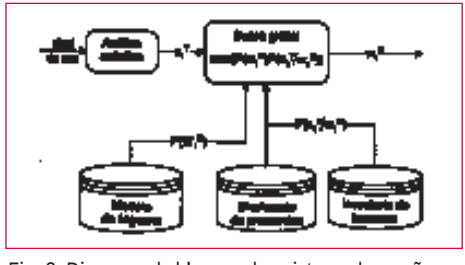


Fig. 2: Diagrama de bloques dun sistema de recoñecemento de fala.

Lamentablemente, esta distribución de probabilidades sería imposible de calcular mediante exemplos, pola cantidade de datos que esixe, se non dispuxesemos da regra de Bayes, que permite reescribir $P(w_1^N|x_1^T) \propto P(w_1^N)\ P(x_1^T|w_1^N)$, é dicir, que a probabilidade que nos interesa é proporcional ó producto da probabilidade da secuencia de palabras pronunciadas e a probabilidade de produci-las observacións oídas dada a secuencia de palabras. Como cando maximizamos unha cantidade tamén maximizámo-las que sexan directamente proporcionais a ela e viceversa, o problema está resolto. Na figura móstrase cómo se mesturan estas propiedades para levar a cabo o recoñecemento.

Pero... ¿non cambiámo-lo cálculo dunha probabilidade polo doutras dúas? ¿Non implica iso máis complexidade? Si, pero cada unha destas probabilidades é moito máis doada de calcular cá orixinal.

O MODELO DE LINGUAXE

En primeiro lugar, $P(w_1^N)$ denomínase '(a probabilidade de) o modelo de linguaxe', porque indica cál é a probabilidade de calquera secuencia de palabras da linguaxe, e normalmente calcúlase como unha 'acumulación' da probabilidade de cada palabra, dado que se pronunciaron tódalas anteriores: $P(w_1^N) = \prod_{n=1}^{N} P(w_n | w_1^{n-1})$

O cálculo destas probabilidades é un campo de estudio bastante extenso en si que vén mortificando a lingüística distribucionalista desde a primeira metade deste século, pero en esencia consiste en contar aparicións ou secuencias de palabras e estima-las probabilidades como as frecuencias relativas empíricas dun conxunto de textos de adestramento.

O principal problema que se presenta neste proceder é condiciona-la probabilidade dunha palabra a tódalas que a precederon, o que é custoso de calcular e incluso imposible para cadeas longas de palabras. Para resolvelo, realízase a aproximación de que a probabilidade de aparición dunha palabra só depende dunhas cantas anteriores, por exemplo, as dúas inmediatamente precedentes —o que se coñece como a 'probabilidade bigrama' e que denomina o seu modelo de linguaxe como 'bigrama'. Mesmo así, para calcula-las probabilidades bigrama dun vocabulario dunhas vinte mil palabras fan falta uns cen millóns de palabras de texto de adestramento.

Outro dos problemas é estima-la probabilidade das palabras que non aparecen no texto de adestramento pero si no uso normal do sistema e que vén a ser unha porcentaxe apreciable do total. Este problema normalmente resólvese con complicados estimadores de probabilidade, pero é un campo de investigación aínda aberto.

O MODELO ACÚSTICO

Resolto o cálculo da probabilidade do modelo de linguaxe, só resta calcula-la probabilidade $P(x_1 | w_1^N)$ de xera-la secuencia de observacións dada a secuencia de palabras, tamén

denominada 'modelo acústico', que é, eminentemente, un modelo de producción da fala.

Para poñe-lo problema en perspectiva, cada fonema xera desde unhas poucas observacións —como as consoantes breves— ata algunhas decenas delas —as vocais ou as nasais coa súa estructura formántica estable—. O noso modelo de producción ten que predicir para cada palabra qué observacións acústicas se van ter en conta ou emitir.

My Fair Lady, de Cukor. O profesor Higgins distinguía nunha soa palabra dita pola florista Eliza Doolittle máis de oitenta sons diferentes.

Para facer isto, ás veces segméntanse as palabras en fonemas, seguindo os dictados da fonoloxía, pero é máis interesante segmentar en fonemas cos contextos fonéticos a dereitas e a esquerdas perfectamente definidos, porque así pódese discriminar entre varias realizacións do mesmo fonema con diferentes graos de coarticulación[1], unha descrición fonética máis próxima ó sinal vocal —fenómeno recollido só parcialmente polo concepto de 'alófono' en fonética—.

1 Ve-lo apartado, 2.1 Caracterización da voz.

Agora ben, cada falante ten o seu propio xeito de pronunciar cada fonema, o que significa que un procedemento de descrición de fonemas debe ter en conta moitos procesos de xeración á vez para ser un modelo robusto dunha poboación. Isto só se pode conseguir se o modelo de cada son pode emitir observacións diferentes para cada falante, o que esixe unha descrición estocástica da emisión de observacións.

Por outro lado, como cada persoa ten unha velocidade media de elocución propia que varía moito coas circunstancias extra- e intralingüísticas, cómpre que o modelo acústico poida xerar fonemas 'rápidos' —é dicir, con poucas observacións— e fonemas 'lentos' —que se poden observar durante máis tempo—.

Este problema non se deu resolvido satisfactoriamente ata que non se utilizaron os modelos ocultos de Markov —v. Deller, Proakis e Hansen (1987), cap. 12—, que son autómatas que simulan unha fonte de observacións mediante dous procesos estocásticos: un que realiza transicións entre estados que non se poden observar —de aí o adxectivo de 'ocultos'— e outro que emite observacións aleatorias en cada transición e son o rexistro que o autómata deixa da súa actividade.

O carácter aleatorio é necesario para absorbe-la variabilidade das observacións da fala, pero fai que o proceso de maximización sexa moi complicado: cada palabra segméntase nos seus fonemas —ou fonemas contextuais— e cada un dos fonemas substitúese polo seu modelo de Markov; ó final, tódolos modelos se concatenan para da-lo modelo de Markov de cada palabra e os modelos de palabra han concatenarse cando se realiza o recoñecemento propiamente para consegui-la maximización conxunta dos modelos acústico e de linguaxe. Como esta maximización adopta a forma dun procedemento de busca da mellor hipótese de recoñecemento mediante a predicción de palabras e a comprobación da súa idoneidade, tamén se coñece como 'busca (das hipóteses mellores de recoñecemento)'.

No modelo oculto de Markov correspondente á palabra w, a probabilidade dunha transición desde o estado s' ó estado s é $P(s|s';w)$, e a probabilidade de emitir unha observación x nese estado é $P(x|s,w)$. Combinando ámbalas probabilidades obtémo-la probabilidade de que estando nun estado transitemos a outro emitindo unha observación:

$$P(x,s|s';w) = P(x|s;w)\cdot P(s|s';w)$$

Sen embargo, concédese adoito que os estados e as observacións de cada fonema non dependen da palabra na que aparecen, senón do contexto acústico no que se atopan, e ás veces suplántase o modelo de palabra polo de contexto acústico —sexa un modelo de fonema, de fonema en contexto ou calquera outro—.

Unha vez construído o autómata que corresponde a unha secuencia

determinada de palabras w_1^N, a probabilidade dunha particular secuencia de observacións x_1^T xerada atravesando unha secuencia de estados s_1^T é $P\left(x_1^T, s_1^T | w_1^N\right) = \prod_{t=1}^{T} P\left(x_t, s_t | s_{t-1}; w_1^N\right)$.

Pero como non coñecemos a ciencia certa o camiño que hai que seguir para chegar á secuencia de observacións, de feito teremos que considerar tódolos camiños alternativos para chegar a calcula-la probabilidade total da secuencia, que resulta ser $P\left(x_1^T | w_1^N\right) = \sum_{s_1^T} \prod_{t=1}^{T} P\left(x_t, s_t | s_{t-1}; w_1^N\right)$, o segundo termo necesario.

A BUSCA NO ESPACIO DE ESTADOS

A solución ó problema de recoñecemento é a secuencia de palabras cunha probabilidade maior $\hat{w}_1^N = \arg \max_{w_1^N} \left\{ P\left(w_1^N\right) \cdot \sum_{s_1^T} P\left(x_1^T, s_1^T | w_1^N\right) \right\}$, pero a miúdo este procedemento é demasiado arduo e cómpre realiza-la aproximación do máximo, pola que a suma das probabilidades na secuencia de estados se substitúe pola da secuencia máis prometedora: $\hat{w}_1^N = \arg \max_{w_1^N} \left\{ P\left(w_1^N\right) \cdot \max_{s_1^T} P\left(x_1^T, s_1^T | w_1^N\right) \right\}$.

A maximización da probabilidade realízase mediante programación dinámica, unha técnica que utiliza a información puntual ('local', en xerga algorítmica) que se vai recollendo no mesmo proceso de recoñecemento para tomar decisións sobre cómo ten que avanzar este; polo tanto, é unha técnica subóptima, é dicir, pode tomar decisións locais incorrectas con respecto á solución global óptima. Na práctica, con todo, a incidencia deste erro non é frecuente, e o procedemento resulta aceptable. Os detalles de implementación non veñen ó caso, pero seguen sendo suxeitos de experimentación e de controversia cando se intentan extrapolar á habilidade de proceso humano.

DIFICULTADES DO RECOÑECEMENTO DE FALA

Daquela, ¿de onde xorden as dificultades no recoñecemento da fala se temos un procedemento matemático para a solución do problema?

A primeira dificultade é que os sistemas deben ser de calidade aceptable para tódolos potenciais locutores que queiran usalos, en tódalas circunstancias posibles. Como existe unha gran variabilidade acústica nos locutores debida a factores fisiolóxicos e psicolóxicos —a velocidade de elocución, etc.—, os modelos acústicos deben estar ensaiados con datos de tódolos potenciais locutores en tódalas posibles situacións, o que leva consigo procedementos e datos de adestramento custosos e coidadosos. Na práctica, abonda con ensaia-los modelos cun número elevado de locutores distintos suficiente para que sexan representativos dos demais, o que permite que a tecnoloxía sexa útil a moita xente.

Sen embargo, un modelo que abrangue moitos locutores con moitas pronuncias diferentes tende a ser pouco discriminatorio, é dicir, confunde as realizacións dun son nun

locutor coas doutro son noutro locutor —por exemplo: /s/ dun locutor con /f/ doutro—, e isto implica que os resultados de recoñecemento sexan peores porque o proceso de maximización se equivoca máis veces.

Por outro lado, é extremadamente difícil dicir en qué intre termina un segmento de son que representa un fonema e en qué momento comeza outro, debido ó problema da coarticulación, a contaminación coas características articulatorias dun fonema dos seus sons circundantes. Este problema complícase con velocidades de elocución elevadas ou pronuncias descoidadas, e é tamén máis evidente nuns acentos ou idiolectos ca noutros.

Ademais, a linguaxe non é só ambigua nos significados, tamén nos sons; algúns sons son particularmente difíciles de distinguir —por exemplo, as fricativas /f/ e /s/ ou as oclusivas /p/, /t/ e /k/—. Calquera ruído ambiente non fai máis que aumenta-la ambigüidade e a confusión, ó enmascarar aínda máis os sons.

Como razón meramente tecnolóxica, o procedemento que esbozamos ten unha dobre complexidade exponencial na lonxitude da secuencia de fonemas que hai que explorar e na lonxitude da secuencia de palabras que hai que supoñer para probar candidatos acústicos: se se utilizase un método de exploración de candidatos de forza bruta para uns miles de palabras de vocabulario cunha frase de lonxitude normal, non existiría tempo na vida media dun ordenador para levar a cabo os cálculos necesarios. Todo isto esixe afina-los algoritmos e optimiza-los cálculos redundantes.

A pesar de todos estes inconvenientes, cada día os sistemas de recoñecemento de fala teñen taxas de erro máis baixas —arredor do 5 % en máquinas comerciais de dictado para ambientes silenciosos e controlados—, e é previsible que noutros dez anos vexamos unha máquina de dictado continuo a bordo dun coche, por exemplo.

2.3 RECOÑECEMENTO DE LOCUTORES

Esta é tamén unha tarefa de entrada de datos, construída sobre a mesma tecnoloxía cá de recoñecemento pero facendo máis fincapé nas características particulares de cada locutor ca nas xerais de tódolos locutores agregados.

Existen dúas tarefas de recoñecemento de locutores, a 'identificación de locutores', na que se intenta saber qué locutor falou de entre un conxunto, e a 'verificación de locutores', que é a aceptación ou o rexeitamento da identidade que un locutor asevera, é dicir, que ademais do sinal vocal témo-la identificación proporcionada polo locutor como dato e emitimos unha decisión sobre tal identidade.

Pódense usar estas técnicas para controla-lo acceso de usuarios a servicios como a marcación telefónica por voz, a banca e compra telefónica, servicios de acceso a bases de datos, información ou correo vocal; e

para o control de acceso a áreas restrinxidas ou ordenadores remotos. A cantidade de aplicacións non fai máis que medrar, o que é un reflexo de que son tarefas definidas de forma suficientemente xenérica para seren útiles.

A base teórica de ámbalas tarefas é a suposición de que a identidade dun locutor está relacionada coas súas características fisiolóxicas e de comportamento, que se manifestan na envolvente espectral da voz —que permite caracteriza-la forma do seu tracto vocal— e coas súas características suprasegmentais —que dependen sobre todo das características dinámicas da fonte glotal, como por exemplo a entoación—, é dicir, características que abranguen máis dun segmento de voz.

Consecuentemente, cada locutor debe estar caracterizado mediante un modelo acústico adecuado para a tarefa (un modelo que sexa moi discriminador nas súas características e adecuadamente capaz de segui-la dinámica da voz), para o que se acostuman usa-los modelos de Markov, aínda que se podería empregar calquera outro modelo que describise tanto as características espectrais como as suprasegmentais.

Despois de realiza-la extracción de parámetros, a configuración máis habitual dun sistema de identificación de locutores compara a secuencia de observacións recibida con cada modelo de locutor existente, seguindo o esquema do recoñecemento de fala antes exposto, e logo decide entre tódolos modelos propoñe-lo idóneo, ou ben ignoralos todos e rexeita-lo locutor. Este proceso necesita un tempo de cálculo proporcional ó número de locutores que estean no conxunto rexistrado, que pode ser moi numeroso.

Sen embargo, para o caso máis sinxelo de verificación, mediante a identificación proporcionada polo locutor accédese a un único modelo que se compara coa mostra de voz moi rapidamente, e rexéitase no caso de que a similitude entre modelo e voz non supere un limiar; para calcular este, ás veces úsanse os demais locutores. Esta técnica de comparación cun limiar tamén se pode empregar no caso da identificación para robustece-lo proceso.

De calquera xeito, ámbalas tarefas aceptan tres modos de funcionamento:

— o modo con texto pechado, ou 'dependente do texto', no que os locutores sempre pronuncian o mesmo texto para seren recoñecidos ou verificados e que corresponde ó problema máis sinxelo;

— o modo 'independente de texto', no que os locutores poden elixir pronunciar calquera texto, que é tamén o problema máis difícil, pola impredicibilidade dos locutores;

— e finalmente o modo con 'texto presentado', no que é o propio sistema de recoñecemento o que proporciona o texto que ha servir para a identificación.

Dados os esquemas anteriores, o recoñecemento de locutores pode fallar por dúas razóns ben diferentes: unha primeira, intencionada e que hai que evitar como sexa, que é a impostura ou suplantación de locutores, na que un locutor malicioso consegue ser verificado ou identificado como outro; e unha segunda, non intencionada pero igualmente perniciosa: a variabilidade dos locutores, que impide a verificación ou identificación correcta dun locutor non malicioso.

Obviamente, o modo de funcionamento máis susceptible á suplantación é o de texto pechado: neste caso, un impostor pode presentar unha gravación dun locutor válido, ou aprender a imitar un locutor válido só nas frases de identificación e ser identificado como tal. O modo independente de texto tampouco proporciona maior protección, posto que calquera elocución válida gravada lle dá ó impostor o acceso degoxado. Co modo de texto presentado, sen embargo, ó impostor resúltalle moito máis difícil dispor da voz natural do locutor válido (aínda que, en teoría, cun sistema moderno de síntese de fala podería, en tempo real, sintetiza-las frases requiridas). Coa chegada de métodos de procesado de sinal máis potente é posible mestura-lo recoñecemento de voz coa lectura de labios e a identificación de rostros, o que dificulta a posibilidade de suplantación, pero tampouco o fai imposible.

En canto ó segundo factor de degradación, a variabilidade das características acústicas dos locutores débese a múltiples motivos: por un lado, o estado fisiolóxico ou psicolóxico dos falantes ou calquera influencia que modifique as características acústicas da súa voz nun período de tempo pouco dilatado; por outro lado, as diferencias entre as condicións acústicas durante o intervalo da captura de datos de adestramento e no momento en que se intenta o recoñecemento de locutor; tamén o ruído de fondo en calquera dos dous momentos, adestramento ou recoñecemento, contribúe á variabilidade, así como os cambios fisiolóxicos a longo prazo, por exemplo os producidos na voz trala pubertade ou a senectude. Finalmente, a variabilidade das voces dos locutores pódese controlar mediante técnicas de normalización, ben sexa dos parámetros que caracterizan as voces, ben das medidas que se usan para decidir sobre as identidades dos locutores; pero esta é unha área de investigación aínda activa.

En conclusión, o campo do recoñecemento de locutores enfróntase coa riqueza da voz humana e a crueza das nosas ferramentas de análise e modelado, polo que esta tampouco se pode considerar unha tecnoloxía que poida ser instalada en servicios que requiran unha seguridade crucial.

2.4 SÍNTESE DA FALA

A diferencia das dúas tecnoloxías anteriores, a síntese de fala é unha tecnoloxía de presentación ou de saída cun fluxo de sinal e informacións que vai do ordenador cara ó oínte. O

proceso da síntese pódese considerar desde unha descrición fonética do que ha ser sintetizado, e entón chámase propiamente 'síntese de voz', ou desde texto escrito, e entón chámase 'conversión texto-voz', e esixe un paso adicional que transforme o texto nunha descrición fonética adecuada.

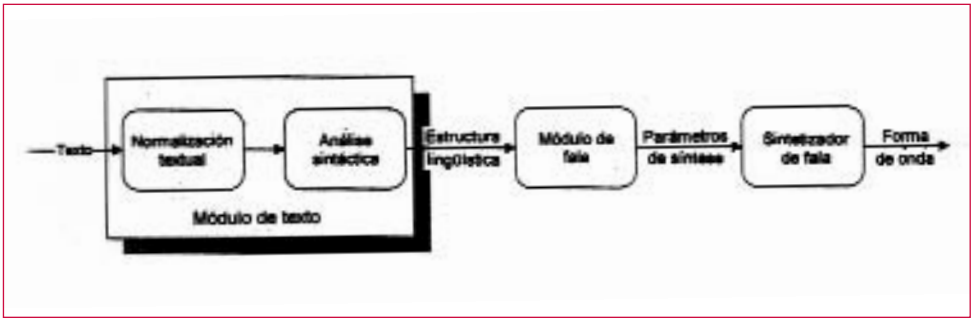


Fig. 3: Diagrama de bloques e fluxos da conversión texto-voz.

Como é evidente, a calidade do proceso de síntese depende da calidade e a riqueza descritiva da anotación fonética e do proceso de transformación ou xeración da propia voz (ningún deles está exento de erros), pero coma sempre, o éxito desta tecnoloxía está en función da calidade percibida polos oíntes humanos e esta defínese mediante dúas medidas: a 'intelixibilidade', que indica cómo de comprensible é a mensaxe subxacente na voz sintética, e a 'naturalidade', que amosa ata qué punto está o oínte disposto a conceder que a voz sintetizada pertence a un falante humano e mostra expresividade —dá transmitido emocións—, estilo —formal o informal, público ou íntimo, etc.—, transmite unha certa situación comunicativa. Para tódolos efectos isto significa que a descrición fonética debe estar anotada paralingüisticamente —con emocións, rexistros, quizais idiolectos, etc.— e extralingüisticamente —sobre todo coa situación extralingüística na que se ha de emitir—.

Para comprender cómo inciden ámbalas medidas, intelixibilidade e naturalidade, na aceptabilidade é conveniente facer constar que un servicio público de recuperación de información telefónica esixe asemade a máxima intelixibilidade e naturalidade, mentres que as aplicacións destinadas a profesionais (control de vehículos ou procesos), ou a persoas altamente motivadas (diminuídos visuais ou persoas que traballan en ambientes ruidosos) esixen exclusivamente unha alta intelixibilidade.

Como se pode deducir da maneira de presenta-lo resumo que fixemos anteriormente, para levar a cabo a síntese de voz hai que recorrer

fundamentalmente a dous corpos doutrinais[2]: o da fonética e o da teoría de sinal.

Por unha banda, a descrición fonética lévase a cabo principalmente en dous niveis: o 'nivel segmental', no que se describen as unidades que van ser sintetizadas —e que normalmente corresponden a fonemas e subunidades ou combinacións destes— e o 'nivel suprasegmental', que describe as características que abranguen máis dun segmento de voz —como a entoación, a duración, o 'tempo', etc.—. Algunhas características suprasegmentais son unha consecuencia directa de características segmentais (no caso da velocidade de elocución, as duracións dos segmentos); pois ben, aínda que se dispón dunha doutrina asentada arredor dos fonemas, non contamos con tantos datos no tocante ás súas realizacións, menos aínda se coñece cómo interaccionan sistematicamente esas realizacións entre si —por exemplo nas coarticulacións— e case nada se dá por definitivo sobre cómo interactúan os segmentos para da-las características suprasegmentais. Esta é unha carencia da teoría fonética clásica que afortunadamente está a suplirse con rapidez, en gran medida pola colaboración coas tecnoloxías da fala que toleran mal as construccións teóricas que non se someten á matemática, por 'experimentables'.

Por outro lado, a teoría do sinal perseverou nos intentos de encontrar mecanismos de xeración de segmentos adecuados e a súa posterior modulación para consegui-las características suprasegmentais requiridas, fundamentalmente motivado por unha case patolóxica desconfianza ós postulados da fonética teórica e experimental: agás a abstracción 'fonema', hai poucas presuposicións fonéticas que non foran postas en dúbida.

Describiremos aquí as últimas tentativas tecnolóxicas baseadas na enxeñería lingüística do corpus, obviando esquemas anteriores, fructíferos no seu momento pero que parecen ter chegado a límites no seu desempeño.

O proceso de síntese con máis éxito nestes momentos consiste en dúas fases, unha de adestramento e outra de síntese efectiva. A fase de adestramento consta das seguintes subfases:

— construír unha base de datos de pronunciación de segmentos co maior número posible de contextos acústicos, melódicos e paralingüísticos;

— posteriormente, esa base de datos analízase mediante determinados parámetros obxectivos que intentan reflecti-los contextos acústicos nos que foron emitidos os segmentos individuais para detecta-las súas fronteiras;

— a continuación entrecóllense os segmentos e almacénanse nunha xigantesca base de datos indizados

2 Usámo-la palabra 'doutrinal' con pleno usufructo da súa connotación relixiosa.

mediante os devanditos parámetros (por exemplo, qué fonemas están contidos en cada segmento e se están acentuados ou non; cál é a velocidade media de elocución; se nun segmento hai cambios de tonema e en qué dirección, etc.).

Cando se queira sintetizar unha descrición fonética particular, só hai que acceder mediante a descrición fonética ós segmentos adecuados e concatenalos en tempo de execución tendo coidado de que os espectros nos extremos dos bordos casen ou se transformen gradualmente uns noutros, para o cal poden utilizarse técnicas de suavizado de espectros.

Evidentemente, este proceso non está exento de problemas: por exemplo, cando os espectros nos extremos dos segmentos que se han concatenar son tan diferentes que a técnica de suavizado non é efectiva. Isto provoca no oínte a sensación de que o falante sintético produce estalos na súa voz.

Un problema inherente a esta técnica é que o espacio de busca de segmentos pode ser inmenso (as dimensións dos parámetros que caracterizan os segmentos multiplícanse entre si) e exploralo en tempo real, unha tarefa case imposible. Outro é que a base de datos non teña os suficientes contextos acústicos para xera-la voz que pide unha determinada anotación lingüística (por exemplo, é difícil xerar unha voz enfadada cunha base de datos de locutores dos programas de noticias da televisión), o que implica que a voz xerada non será expresiva abondo: se se sintetiza unha elexía, un poema de amor ou un fragmento de traxedia, dá risa en vez dos sentimentos que cabería esperar.

Mesmo así, a síntese mediante corpus presenta vantaxes orixinadas en tres decisións de deseño: os procedementos para a selección de segmentos e unidades, as medidas obxectivas para a análise e selección das unidades baseadas nas anotacións fonéticas e o deseño do corpus do que se extraen os segmentos.

Na actualidade os sistemas de conversión texto-voz proporcionan unha alta intelixibilidade pero a naturalidade é moi mala; fáltalles habilidade para controla-la expresividade, o estilo —incluíndo o rexistro— e acadar unha identidade definida: a este respecto, están conseguidas as voces de locutores masculinos adultos e son boas as de locutoras femininas adultas, ambas en situacións neutras, pero non as doutros segmentos da poboación como nenos ou adolescentes, nin outras situacións máis emotivas —aínda que os primeiros resultados son esperanzadores—.

Esencialmente, salvo en ambientes moi ruidosos, en ausencia de defectos auditivos unha voz sintética é claramente diferenciable dunha voz humana, móstrase cansa ó escoitala e monótona, e pode estar, por veces, salpicada de estalos.

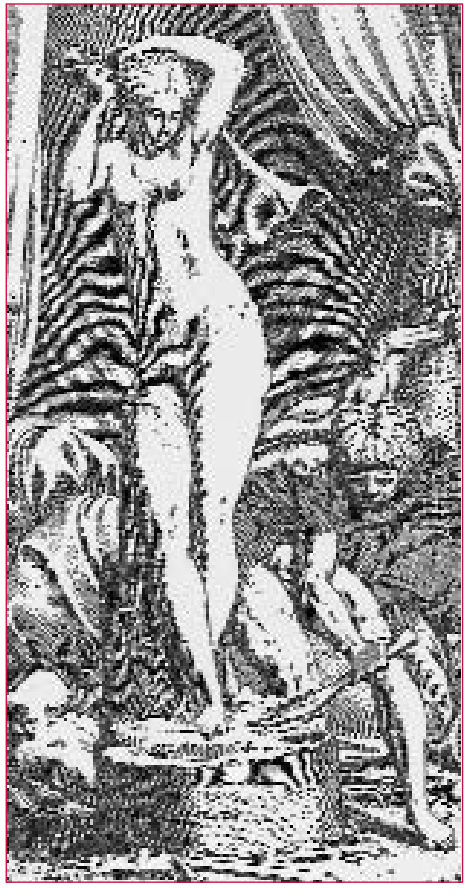
A xeito de conclusión, son numerosos os factores que contribúen a esta

falta de calidade na síntese: insuficiente coñecemento dos factores que fan que unha voz sexa natural, limitado coñecemento de cómo obter, analizar e reorganiza-los sons adecuados para obter unha voz determinada, e practicamente nulo coñecemento sobre cómo facer que dos sons organizados emerxa a naturalidade. Moitas destas carencias son achacables ós métodos descritivos da fonética teórica —que, en puridade, non era unha ciencia de descrición da capacidade de xeración de sons, senón a ciencia da análise dos sons pronunciados—, pero tamén se poden atribuír notables carencias á bagaxe técnica do tratamento de sinais, notablemente á fe cega en determinados procedementos que teñen límites tecnolóxicos e a súa reticencia a probar técnicas directamente inspiradas no acervo conceptual máis madurado da fonética.

3. ¿POR QUE FALLARON AS TÉCNICAS DA FALA?

Despois de revisar, polo menos sumariamente, varias tecnoloxías da fala, estamos en disposición de comprender cómo é posible que aínda non existan máquinas que falen coma unha persoa e que entendan o que lles dicimos, cando nas películas se levan predicindo estes mesmos sucesos desde hai polo menos trinta anos, e hai ata xoguetes de neno que aprenden a falar. Ó noso xeito de ver isto débese a múltiples factores que tentaremos ir expoñendo nos seguintes parágrafos.

O primeiro feito do que hai que tomar conciencia é que a descrición, o modelado, a síntese e o recoñecemento de voz, en calquera das súas encarnacións, son problemas extremadamente difíciles por factores intrínsecos á linguaxe e as capacidades de fonación e audición humanas, entre os cales destaca a variabilidade e a influencia



Pigmalión, Mestre L. D., gravado, século XVI.
Coa información e a robótica como ferramentas volve ter sentido o mito de Pigmalión.

do contexto —sexa acústico ou lingüístico— no comportamento das unidades, e o feito de que aínda existen moitos puntos escuros na fisioloxía da audición e a fonación como para que consigamos arremedos perfectos das súas capacidades. Tamén é posible que carezamos, polo de agora, das adecuadas ferramentas alxébricas ou numéricas para construírmos estas imitacións, e neste caso sería inútil intenta-la construcción de sistemas perfectamente miméticos das capacidades lingüísticas humanas. A aproximación actual cumpre o papel de modela-las capacidades humanas razoablemente ben e debemos estar contentos con iso.

Ademais, gran parte dos factores alleos á propia lingua son imputables ás comunidades científicas directamente involucradas nas tecnoloxías da fala: os lingüistas e os enxeñeiros —normalmente 'electrónicos' ou de 'ordenadores'—, e tamén, como non, ós usuarios.

Por un lado, moitos estudiosos da linguaxe, incluídos lingüistas, seguen concibindo o estudio da linguaxe como a adoración[3] dun fenómeno que é capaz de producir *O Quixote*, a obra de Lope de Vega e a poesía de Rubén Darío. Os esforzos que moitos outros lingüistas levan realizando desde hai catro décadas para introduci-los métodos cuantitativos e alxébricos —que transforman un saber nunha ciencia experimental— chocan frontalmente coa inercia, a escasa formación matemática e o rexeitamento a abandona-la visión transcendental do uso da lingua —quizais mellor encarnada no estudio literario— que practican moitos integrantes das 'vellas escolas'. ¡E teñen todo o dereito e a razón para opinaren así! Aínda que non teñen a exclusividade deste estudio nin o dereito de frear aproximacións alternativas.

Por outro lado, a aproximación *matematicista* cega ó obxecto de estudio, soberbia ante os predecesores e orientada á realización de beneficios, perverte calquera intento serio de que un enxeñeiro consiga introducirse nas ciencias da linguaxe. A súa ignorancia da tradición lingüística levarao tanto a cuestionar concepcións lingüísticas mal asentadas en datos lingüísticos —práctica saudable—, como a insistir en puntos que xa foron resolvidos polos estudios sobre estructura e comportamento da linguaxe. A limitada capacidade dos seus métodos matemáticos para modelar sistemas complexos, faralle recorta-lo obxecto de estudio para conformarse con aquilo estrictamente modelable e abarcable desde a súa limitada perspectiva, nunha febril carreira para obter mellores resultados có laboratorio ou a empresa veciña cos que mendigar unha subvención que perpetúe este eterno correr cara a ningures.

Non pouca responsabilidade nesta carencia de sistemas miméticos débese á necesidade dos investigadores de proceder a pequenos pasos que publicacións en curtos espacios de tempo coas que xustificar unha actividade investigadora prolífica. Isto redunda nun menor interese polas

3 Tamén considerando a súa connotación relixiosa.

investigacións arriscadas con rendementos a máis longo prazo e polas aproximacións diverxentes, tan necesarias no estudio científico maduro.

Finalmente, mesmo na ciencia se confía nos milagres: existe unha certa esperanza —fundada nunha lei empírica, ben é certo— que predí que cada certo tempo, arredor do lustro ou os dez anos, se produce en cada campo de estudio unha revolución a cargo dalgunha persoa emprendedora e xenial que arrastra as demais á súa forma de entende-la realidade. A primeira vez que oímos este razoamento aplicado ó campo do recoñecemento continuo da fala foi hai uns oito anos. Por suposto, aínda seguimos esperando ese revolucionario.

Outra parte da responsabilidade corre a cargo das entidades públicas financiadoras, que coa súa énfase na investigación dirixida por tarefas de validación, por evitaren un vicio (a investigación non encamiñada a obter resultados que alcancen e melloren as vidas do esforzado contribuínte), dan noutro: o de considerar que se lle poden poñer prazos á creatividade científica, que é a orixe de toda tecnoloxía; para evitar que o diñeiro caia no saco roto dos investigadores e dos que poñen en práctica —que vivían, literalmente, das subvencións—, as administracións asumiron os obxectivos das empresas e subvencionan os proxectos máis fantasiosos e inverosímiles con tal de que teñan trazas de comercialización, estragando os poucos fondos que á investigación se dedican.

Por outro lado, o exceso de expectación dos consumidores failles despreza-los bos productos que de feito posúen, medindo as súas expectativas polas fantasías ideadas por uns visionarios que imaxinan máquinas nas que residen pouco menos que articulados cerebros humanos (se cadra porque a linguaxe é o que máis parece diferenciarnos dos obxectos), sen se parar a considerar que, á fin e ó cabo, unha lavadora só lava, un forno microondas só quenta, un teléfono só transmite un sinal ou que un corrector ortográfico só corrixe faltas de ortografía e non ensina a escribir nin entende o que corrixe, nin ten por qué compoñer mellores sonetos ca Quevedo.

Temos teléfonos fixos dixitais e teléfonos móbiles con procesamentos extremadamente enxeñosos, pero que non mostran máis intelixencia ou comprensión cá da persoa que está no outro extremo da comunicación; máquinas de tabaco ou de gasolina que se dirixen a nós e ascensores que nos lembran en qué piso estamos, e aínda non nos abonda: mentres non xurda o revolucionario mesías das tecnoloxías de voz teremos que seguir conformándonos, como usuarios e investigadores das tecnoloxías da fala, con tímidas melloras nos nosos sistemas de dictado, recoñecedores e sistemas de diálogo que, esperemos, consigan axudarnos a supera-las nosas barreiras comunicativas ou tecnolóxicas.

4. REFERENCIAS

LRE, *Ingeniería Lingüística. Cómo aprovechar la fuerza del lenguaje,* URL: http://www.linglink.lu/le/es/broch/harness.htm, 1999.

Deller, J. R., J. G. Proakis e J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Upper Saddle River, Nova Jersey, Prentice-Hall, 1987

Manning, C. D., e H. Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge (EUA), The MIT Press, 1999.

Ney, H., e S. Ortmanns, "Dynamic Programming Search for Continuous Speech Recognition", *IEEE Signal Processing Magazine*, vol. 6, núm. 5, 1999, páxs. 64-83.

Varile, G., e A. Zampolli (eds.), *Survey of the State of the Art in Human Language Technology*. Editado pola "National Science Foundation", o Directorado XIII-E da Comisión das Comunidades Europeas e o "Center for Spoken Language Understanding" do Oregon Graduate Institute. URL: http://cslu.cse.ogi.edu/HLTsurvey/download.html, 1995. (Existe unha versión editada como: Varile G., e A. Zampolli (eds.), *Survey of the State of the Art in Human Language Technology*. Giardini Editori and Cambridge University Press, 1997.