# Les tecnologies de la parla.

## Lloc de trobada, difícil però necessària, entre lingüística i tecnologia

La presentació de les tecnologies de la parla que aquí pren forma de document escrit s'inscrivia en un curs sobre les aplicacions de la lingüística adreçat principalment a persones procedents d'aquest camp. L'encàrrec, però, s'havia fet a un enginyer que treballa en el desenvolupament d'aquestes tecnologies. La decisió d'acceptar-lo no em va resultar fàcil. En part perquè sóc conscient que la meva mirada sobre la realitat tecnològica està condicionada pel meu bagatge d'enginyer, diferent del d'un lingüista. Però sobretot perquè la recerca i el desenvolupament d'aplicacions en tecnologies de la parla han deixat força de banda els coneixements lingüístics. Les raons són diverses, com veurem, però la dificultat del treball conjunt de lingüistes i tecnòlegs no significa que aquest no sigui necessari per arribar més lluny en els objectius de la pròpia tecnologia. Com és evident, vaig decidir acceptar el repte, i deixar-lo reflectit en el subtítol d'aquest escrit.

És interessant constatar que, mentre que en el tractament del llenguatge escrit (llenguatge natural) hi han intervingut tradicionalment lingüistes (anomenats computacionals), treballant habitualment al costat d'informàtics, no ha estat així en el tractament del llenguatge oral, domini que ha estat conreat sobretot pels enginyers elèctrics (de telecomunicació, en diríem al nostre país). En efecte, només cal fer una ullada als termes definitoris de cada una de les dues àrees tecnocientífiques per constatar això que diem. Des del costat textual es parla amb termes de la lingüística: etiquetatge morfològic, anàlisi sintàctica, desambiguació, etc. En canvi, del costat oral les paraules-clau són mots que agraden als enginyers: codificació, reconeixement, síntesi, conversió, etc. Pel que fa als noms de les aplicacions, els dos dominis usen una terminologia més semblant, de manera que la lingüística (computacional) hi parla d'una forma més propera a l'enginyeria: extracció d'informació (multimèdia, si cal), resum automàtic, etc. Tanmateix, la diferència terminològica al nivell de les tecnologies bàsiques és prou significativa de la llunyania relativa del tractament de la parla respecte de la lingüística.

Joaquim Llisterri, un autor proper que, des del camp de la lingüística i específicament de la fonètica, pot parlar amb coneixement de primera mà d'aquest cert "destrobament" que estem considerant, fa observar el "contraste entre las constantes referencias a la necesidad de colaboración entre fonetistas y tecnólogos y la situación que se desprende de las publicaciones: si éstas reflejan acertadamente la realidad, no parece que, al menos en nuestro contexto más inmediato, sobren motivos para el optimismo en lo que a la interdisciplinariedad se refiere"[1]. I presenta estadístiques de comunicacions a congressos per constatar millor la separació existent entre l'àmbit de la fonètica i el de les tecnologies de la parla.

Aquest escrit no pretén abastar de la mateixa manera totes les tecnologies de la parla, sinó que es centra sobretot en una: el reconeixement de la parla. A part del fet que l'autor coneix aquesta tecnologia millor que les altres, el reconeixement és segurament l'àrea tecnològica que millor exemplifica la dificultat de trobament a què es refereix el subtítol.

Abans, però, d'endinsar-nos en les tecnologies ens aturarem en les finalitats de les seves aplicacions i, a continuació, observarem el fort contrast que històricament ha existit entre les expectatives generades en els potencials usuaris i la molt més minsa realitat tecnològica.

[1] Joaquim Llisterri ha buscat i fomentat, des dels anys 80, la trobada interdisciplinària amb els tecnòlegs i ha col·laborat amb enginyers en nombrosos projectes de recerca i desenvolupament. La cita està extreta de la seva excel·lent publicació titulada "El papel de la fonética en las tecnologías del habla", referenciada al final de la bibliografia, la qual ha estat una referència molt valuosa per a la redacció del present escrit.

## Al servei de la interacció persona-màquina i entre persones

La interacció entre persona i màquina, entre usuari o usuària i sistema, és una finalitat de primer ordre per a les tecnologies de la parla. L'usuària vol que la seva parla sigui reconeguda (potser també la seva identitat) i que el sistema empri també la parla per comunicar-se amb ella. Però això no és suficient. El sistema ha de tenir la funció que podríem anomenar *comprensió* (aquest és el terme usat en l'àrea, sense cometes). En efecte, el reconeixedor només subministra una seqüència de mots i no s'ocupa de posar-hi majúscules o signes de puntuació. Res no ha estat "entès" fins aquell moment. Ha de ser el mòdul operatiu que va a continuació el que converteixi la cadena de mots en una cadena d'unitats semàntiques; i també el sistema ha de fer la conversió inversa abans de sintetitzar la "seva" parla, és a dir, ha de generar el text a partir d'una cadena semàntica estructurada. Falta encara considerar un mòdul que gestioni les unitats semàntiques d'entrada i sortida del sistema tot modelant el propòsit de la interacció: el que s'anomena gestor del *diàleg*. Estem davant d'una altra paraula que, com *comprensió*, té un contingut fortament reduccionista en aquest context tècnic. A vegades em pregunto si amb aquesta atribució a les màquines de mots comuns que fan referència a les capacitats més específicament humanes no estarem devaluant a la llarga l'ésser humà.

Quan les tecnologies de la parla se situen en els llaços d'interacció entre persones per tal de facilitar-les-hi la comunicació, es requereixen els mateixos mòduls bàsics de reconeixement i síntesi de parla. La resta dependrà de la funcionalitat concreta que es persegueixi. Per exemple, si es tracta de la traducció de parla, caldrà un mòdul que reculli el text a la sortida del reconeixedor en la llengua d'origen i lliuri al sintetitzador de parla el text en l'altra llengua. Si el mòdul és de base estadística no hi trobarem de forma explícita el que més amunt hem anomenat *comprensió*; en canvi, altres funcionalitats, com per exemple l'extracció d'informació, sí que la requereixen.

## Expectatives dels usuaris i realitat tecnològica

Probablement, les tecnologies del llenguatge oral han generat més expectatives que moltes altres àrees tecnològiques. Això es pot dir també de l'àrea que les inclou, l'anomenada intel·ligència artificial. Tanmateix, potser pel lloc eminent que ocupa la parla en la vida dels humans, tant els mitjans de comunicació com els propis investigadors han creat unes expectatives excessivament elevades, que s'han matisat, però no suprimit, amb el pas del temps. La complexitat del cervell humà fa fracassar qualsevol intent de modelar adequadament els seus processos, i en particular el del llenguatge, que n'és l'expressió cabdal, amb els limitats mitjans científics i tècnics actuals.

En tenim una bona mostra a la pel·lícula *2001 Odissea de l'espai*. En ella, l'ordinador de control *Hal* a bord de la nau espacial, es comunica amb la tripulació humana –sobretot– per mitjà de la veu, comprèn perfectament el que se li diu i genera discursos acurats. La seva veu resulta natural, expressiva, humana; i fins i tot canta. Han passat ja 10 anys des de l'any 2001 on, fent ciència-ficció, se situaven els fets de la novel·la, i encara estem molt lluny d'aquella predicció. Així és: es pot fer dictat automàtic davant un ordinador, però les equivocacions del sistema són molt més freqüents del que es voldria, la qual cosa fa que es faci servir poc, malgrat la dificultat d'entrar text amb els teclats minúsculs dels aparells mòbils; es pot escoltar la lectura automàtica d'un text informatiu curt, però si es volgués fer amb una novel·la o poesia, la qualitat de la veu la faria ben poc agradable, ja que es percep robòtica i transmet de forma maldestre les emocions; es pot demanar informació per telèfon sense intervenció d'un agent humà, però el grau de satisfacció de la usuària un cop acabada la interacció és molt millorable, perquè, per poc que demani la tasca comunicativa, el

sistema fa la impressió de ser molt curt de gambals. I no parlem ja de tasques força més complexes com la traducció automàtica oral…

La producció fílmica *2001 Odissea de l'espai* és de l'any 1968. Podem disculpar la seva manca d'encert futurològic perquè estaven en un moment germinal d'aquestes tecnologies i es feia difícil preveure el grau de progrés només a partir de l'impuls inicial. La situació, tanmateix, s'ha anat repetint des de llavors. Encara ara, quan llegim o escoltem als mitjans de comunicació el resultat d'una consulta realitzada per periodistes a investigadors de la nostra àrea fa sovint la impressió que s'està a punt d'assolir un cim tecnològic, la qual cosa deu ser conseqüència de l'anhel dels potencials usuaris de productes que han d'atorgar capacitats orals a les màquines que ens ajuden en les tasques quotidianes.

Certament, s'ha fet força camí en els darrers 30 o 40 anys. Per exemple: s'ha passat de reconèixer els nombres del 0 al 9 dits amb el micròfon a ran de llavis, a convertir en text discursos parlamentaris amb només un 10 o 15 percent d'error a nivell de mots; el pas de text a veu es fa pràcticament sense limitacions i la intel·ligibilitat que s'aconsegueix és elevada; i partint de la gravació audiovisual d'un noticiari televisiu, la veu original, un cop aïllada, és processada en el laboratori per una sèrie de blocs en cadena (segmentador d'àudio, reconeixedor de parla i parlant, traductor text-text, conversor text-veu) fins a generar una veu de característiques similars que expressa en una altra llengua el que la cadena de processament ha pogut preservar del missatge original.

Sens dubte, l'avenç tecnològic ha estat remarcable, però tot i així, encara queda molt lluny de les expectatives. Aquesta distància és més notable en el reconeixement de la parla, possiblement per la gran complexitat que implica la percepció humana. Ens podem preguntar, doncs, què fa, en concret, que l'empresa sigui de tan abast. No em vull referir aquí a la complexitat estructural del llenguatge ni a la necessitat de desambiguació, factors ben coneguts pels lingüistes. Tampoc em voldria fixar en la diversitat de formes que pot prendre un mateix contingut semàntic, que ens trobem també en la llengua escrita (anant del llibre formal als xats en mitjans telemàtics), tot i que la parla es presenta sovint amb un grau elevat d'espontaneïtat que provoca multitud de disfuncions gramaticals i morfològiques. Voldria aturar-me tan sols en una característica que el llenguatge oral no comparteix amb el textual i que és a l'arrel de bona part de les dificultats que tenen els sistemes que processen la parla: la seva variabilitat. És a dir, un mateix missatge oral pot arribar a l'entrada del sistema digital que el processa prenent formes oscil·logràmiques variades a causa, per exemple, de les diferències existents entre parlants o entre condicions acústiques de l'entorn.

## Factors de variabilitat de tipus acústic i elèctric que són propis de la parla

El missatge que genera el cervell humà, expressable amb una seqüència de mots i de trets prosòdics, ha de passar, per materialitzar-se en veu, per un procés de codificació acústica en l'aparell fonador. Aquest sintetitza les formes d'ona de pressió de l'aire que transporten el missatge codificat, les quals es converteixen al micròfon en ones elèctriques, i són posteriorment digitalizades. Les ones generades per l'aparell fonador es transmeten per dos canals que les modifiquen, distorsionant-les i embrutant-les amb soroll: primer, el canal acústic que hi ha entre els llavis del parlant i el micròfon, i segon, el canal elèctric que va de l'entrada al micròfon a l'entrada al sistema de processament digital. Doncs bé, com ara veurem, tant la codificació acústica com el pas pels dos canals de transmissió esmentats són fonts de variabilitat que fan enormement complexa la tasca dels sistemes de processament que volen simular la capacitat humana que anomenem parla.

La codificació acústica que es realitza en l'aparell fonador del parlant es veu afectada, lògicament, per les seves característiques, permanents o circumstancials: dialecte, edat, sexe, condicions psicofísiques (manera d'expressar-se, estat emocional, presència de soroll ambiental molest,…), condicions fisiològiques de la fonació (encostipat, disfuncions articulatòries,…), etc. Aquestes característiques del parlant es reflecteixen tant en la prosòdia com en el timbre de la seva veu. La variabilitat que provoquen en la parla és una causa important de la dificultat per aconseguir les baixes taxes d'error de reconeixement que demanen la majoria d'aplicacions. I aquesta gran variabilitat és també un escull que troba la síntesi per aconseguir una parla que sigui percebuda per l'oient com a natural i expressiva. No cal dir que són les característiques de caràcter permanent les que permeten identificar automàticament un determinat parlant, mentre que les circumstancials en dificulten la tasca.

Tant el canal per on passa l'ona acústica com el canal de transmissió elèctrica modifiquen el senyal que transporta els símbols lingüístics. El primer canal ve condicionat per l'entorn arquitectònic, que determina el grau de reverberació, i per l'ambient acústic, que pot afegir-hi altres veus o sons diferents de la parla. Tant la reverberació com el soroll ambiental poden produir una forta davallada en les prestacions dels sistemes de reconeixement de la parla, passant fàcilment d'un 10% d'error en paraules amb el micròfon davant de la boca del locutor a un 50% o més. Parlem de sistemes de reconeixement de la parla, però el mateix es podria dir de qualsevol altra forma de reconeixement: de parlant, d'emocions, d'idioma, etc. La variabilitat produïda per la diversitat de canals elèctrics, en general, no sol ser tant molesta per als diversos tipus de reconeixedors com ho és la dels canals acústics.

Si es volen prestacions òptimes s'ha de recórrer a sistemes adaptats al canal elèctric i acústic on treballa el sistema; per aquesta raó, referint-nos ara només al reconeixement de la parla, s'han enregistrat i etiquetat corpus orals independentment per a micròfon d'ordinador (en català, per exemple, *FreeSpeech*[2]) o telèfon (*SpeechDat*), tot i que ambdós casos comparteixen el fet que la distància dels llavis del parlant al micròfon del sistema de gravació és curta. I també s'han produït corpus orals enfocats a reconeixement de la parla en entorns on els micròfons estan separats dels parlants, com l'entorn de cotxe (*SpeechDatCar*), o en un conjunt d'entorns diversos com oficina, sala d'estar de llar, cafè, exterior,… (*Speecon*); per enregistrar aquests darrers corpus s'han usat simultàniament micròfons a diferents distàncies, sempre amb l'objectiu de recollir com més millor la variabilitat del canal acústic.

## El reconeixement de la parla com a exemple de tractament de la variabilitat

Com s'afirmava més amunt, el reconeixement de la parla és un dels millors exemples de la distància existent entre les expectatives dels usuaris i la realitat de les tecnologies de la parla. En ell s'hi ajunten tots els factors de dificultat suara esmentats, tant els propis de la complexitat del llenguatge com els atribuïbles a la diversitat, ja sigui d'estils de parla o –com s'ha comentat en el darrer apartat– de tipus acústic i elèctric.

El primer que ha de fer el lingüista computacional o tècnic que vulgui establir els requeriments del sistema de reconeixement en una determinada aplicació és respondre a les següents qüestions relacionades amb els factors de dificultat esmentats més amunt:

[2] Tots els acrònims en cursiva que apareixen en aquest paràgraf corresponen a bases de dades que han estat produïdes per al català seguint estàndards desenvolupats en projectes financiats per la Unió Europea. En general, hi són representats els quatre grans dialectes (central, nordoccidental, valencià i balear), tot i que en graus diversos.

- Vocabulari: grandària i tipus (dígits, comanes, llista de noms, vocabulari complet d'un domini semàntic,...).
- Domini semàntic: obert o restringit a un àmbit concret (notícies, sanitat, justícia,...).
- Estil de parla: lectura, discurs, conversa informal,...
- Conjunt d'usuaris: un de sol, sistema adaptat a cada usuari o obert a qualsevol usuari
- Entorn acústic: oficina, sala d'estar, vestíbul, bar, carrer,...
- Transmissió elèctrica: micròfon, telèfon, veu IP,...

L'eficàcia dels sistemes actuals de reconeixement de la parla rau en l'ús de models estadístics generats amb mètodes d'aprenentatge automàtic. Els sistemes basats en coneixement lingüístic (fonològic, lèxic, sintàctic, semàntic,...) que s'havien començat a desenvolupar als anys 70 del segle passat no van aconseguir tractar bé la variabilitat de la parla i des dels anys 80 han ocupat un lloc menor en el desenvolupament tecnològic. De fet, la lingüística descriu el llenguatge amb regles, però la parla espontània té una forta dosi d'irregularitat. A més, el coneixement lingüístic s'ha basat tradicionalment en els comportaments típics i el material d'estudi s'adquiria sota condicions ben controlades, sense atendre gaire, doncs, a la diversitat.

Els sistemes de reconeixement de la parla converteixen el senyal de veu en una seqüència de paraules. El seu objectiu principal, per tant, és obtenir un text amb el mínim error a nivell de paraula. L'error es mesura sumant el nombre de substitucions, insercions i esborraments de paraules respecte de la seqüència correcta. Quan el vocabulari és petit, cada paraula pot tenir un model estadístic propi. En vocabularis grans (actualment, un vocabulari de 50 o 100 mil mots és habitual en els sistemes més posats al dia), s'ha de recórrer a unitats més curtes, típicament els fonemes (i al·lòfons). Els models fonèticoacústics de les paraules del vocabulari es formen combinant, d'acord amb la transcripció fonètica, els models d'aquestes unitats fonètiques elementals. La taxa de reconeixement millora substancialment si les unitats són contextuals, ja que incorporen la variabilitat causada per la coarticulació. Actualment, doncs, s'opta per l'ús de fonemes en context. Per exemple, /m-a-s/ és el fonema /a/ amb context esquerre /m/ i context dret /s/. Donat que el nombre d'unitats, que depèn de la llengua, arriba fàcilment a ser d'uns quants milers, es requereix un corpus d'entrenament que sigui suficientment gran per estimar l'enorme conjunt de paràmetres estadístics que hi ha involucrats. En efecte, com que habitualment s'empren models de Markov ocults, amb desenes o fins i tot centenars de distribucions gaussianes per unitat, i cada gaussiana sol implicar uns 80 paràmetres (les mitjanes i variàncies de les característiques de la veu), els paràmetres es compten per milions.

En el reconeixement s'usen també els anomenats *models de llenguatge*, que estableixen les connexions entre els mots del vocabulari. Poden ser simples grafs per a tasques de reconeixement d'abast reduït (per exemple: les hores del dia o els nombres naturals). Però quan, com sol passar, la tasca és més complexa i el vocabulari és extens, s'usen els anomenats *n-grames*, que caracteritzen la probabilitat d'observar cada conjunt ordenat de *n* paraules. Típicament, *n* és 3, de manera que a cada paraula situada entre dues més (per exemple: /casa/ entre /una/ i /pairal/) se li assigna una probabilitat que es troba observant el grau d'aparició de la tríada en enormes bancs de dades textuals recollits del domini semàntic corresponent a la tasca de reconeixement. Observi's que això resulta ser una forma implícita de modelar les restriccions sintàctiques i semàntiques. Forma ben barroera, ja que no té en compte cap tipus de coneixement, i a més a més limitada pel fet de no permetre capturar les relacions entre paraules distants dins de l'oració.

Amb tot el que s'ha comentat, es constata que el paper de la lingüística és ben minso en aquesta forma actual de treballar en el reconeixement de la parla. De fet, es redueix a definir l'inventari de fonemes i al·lòfons de la llengua, a transcriure fonèticament les paraules del vocabulari (a vegades, s'admeten diferents transcripcions per tal de modelar les variants dialectals o altres) i, per últim, a participar en la producció del corpus d'entrenament dels models foneticoacústics). Fins i tot la transcripció manual o l'ús de regles de transcripció es fan innecessaris si el tècnic crea un transcriptor automàtic entrenat a partir d'un corpus transcrit manualment.

Hi ha un altre aspecte del reconeixement on s'usa un cert coneixement fonètic: l'agrupament de contextos. Aquesta operació es fa de forma automàtica durant l'obtenció dels models foneticoacústics en l'entrenament del sistema. Consisteix a agrupar els contextos de cada fonema que són més semblants estadísticament a fi i efecte de reduir el nombre d'unitats fonètiques. Com que el corpus d'entrenament és limitat, quan s'agrupen dues unitats els seus nombres d'aparicions en el corpus se sumen, de manera que el model de la unitat resultant pot ser més ben entrenat. Aquesta tècnica cerca, doncs, un compromís entre invariància acústica de la unitat i estimació fiable del seu model estadístic. Per dur-ho a terme, es construeix un arbre de contextos per a cada unitat, on a cada node se li associa una pregunta de caire fonètic com, per exemple, "el context dret, és consonant nasal?"

## El paper de la lingüística en altres tecnologies de la parla

Els sistemes de conversió de text en parla que actualment ofereixen més qualitat sintetitzen la veu mitjançant la concatenació d'unitats fonètiques elementals que engloben les transicions entre sons. Anàlogament al reconeixement, la síntesi va recorrent cada cop més a corpus orals, d'on se seleccionen automàticament les unitats fonètiques més apropiades des del punt de vista acústic i prosòdic. Malgrat que el desenvolupament de la conversió de text en parla ha requerit tradicionalment una forta presència de coneixement lingüístic, aquesta s'ha reduït a causa de la selecció feta amb criteris matemàtics. I encara es redueix més en un enfocament més recent de la síntesi que, com el reconeixement, es basa en aprenentatge automàtic amb models de Markov ocults i distribucions gaussianes.

Per establir una comunicació persona-ordinador que vagi més enllà de la transcripció textual del que es diu o de la recepció d'ordres, el sistema ha de ser capaç de gestionar la successió d'interaccions amb l'interlocutor humà. Aquest *diàleg* requereix també un model, la complexitat del qual depèn de l'abast semàntic de la comunicació. Actualment, el *diàleg* pot ser relativament flexible en dominis semàntics limitats com, per exemple, la consulta d'horaris i preus de viatges en tren o avió. El coneixement del lingüista intervé en diversos punts del desenvolupament (producció i anàlisi de corpus, disseny d'estratègies,…), tot i que també en aquesta tecnologia es tendeix a incrementar l'aportació de l'aprenentatge automàtic dels models.

En la traducció automàtica de la parla predominaven fins fa poc els sistemes basats en regles, que utilitzen coneixement lingüístic. Però en els darrers anys ha irromput amb força l'enfocament estadístic, que usa models generats automàticament, i en l'actualitat és l'àrea predominant en la recerca. Com passa en altres tecnologies de la parla, amb aquest enfocament l'aportació lingüística queda bastant relegada a la producció del corpus que serveix per desenvolupar el sistema, que en aquest cas ha de ser un corpus paral·lelitzat.

En aplicacions de diàleg, modalitats no estadístiques de traducció de parla, etc, es requereix la *comprensió* del missatge oral, la qual cosa significa que el sistema ha de convertir la cadena de paraules lliurada pel reconeixedor en una cadena de símbols semàntics. Altre cop,

aquí es recorre sovint a l'estadística, de manera que la incorporació de coneixement lingüístic és limitada. En canvi, en l'operació inversa, és a dir, la generació de text a partir d'una descripció semàntica, s'usa més el coneixement, ja que se sol dur a terme d'acord amb les regles sintàctiques i morfològiques de la llengua.

## Conclusió i perspectives de futur

Fins aquí s'ha exposat, breument i de forma genèrica, l'estat actual de les tecnologies de la parla i les seves aplicacions. Sense pretendre el tractament equilibrat de les diverses tecnologies, s'ha seguit un fil argumental que intentava respondre a la pregunta: ¿per què el coneixement lingüístic ha anat quedant progressivament al marge dels desenvolupaments tecnològics relacionats amb la parla? Ara, en acabar, és el moment de valorar aquesta tendència i apuntar alternatives.

En la nostra àrea tecnològica hi ha un convenciment força estès que un ús apropiat i més intensiu del coneixement fonètic –i, en general, lingüístic– hauria de beneficiar les prestacions de les pròpies tecnologies. D'una banda, es tractaria de disposar de sistemes de tractament de la parla que fossin més permeables a la inclusió de coneixement. D'altra banda, l'enfocament de les investigacions lingüístiques, i de les fonètiques en particular, hauria d'integrar la variabilitat de la parla i orientar-se a la producció de descripcions i models de tipus més quantitatiu, amb formalismes compatibles amb els dels sistemes de tractament de la parla.

Diu Steve Greenberg: "speech technology can proudly point to its apparent success with speech recognition and concatenative synthesis in defense of its machine-learning-centric approach […] imperfect science is capable of providing an effective foundation for technology –as long as the demands of the market are not exceedingly stringent or profund"[3]. La recerca, condicionada per la necessitat de finançament, busca resultats a curt termini. L'aprenentatge automàtic a partir de corpus de grans dimensions ha ofert i segueix oferint progressos tecnològics immediats. Es basa en la ciència matemàtica, però, per la seva pròpia naturalesa, li resulta difícil incloure resultats provinents d'altres disciplines, com la lingüística o la neurociència. A això es deu referir Greenberg quan parla de la "ciència imperfecta". Ara bé, tal com ell mateix suggereix en el text citat, els sistemes resultants no poden atendre a les demandes més exigents en termes de prestacions.

Costa de creure que seguint el paradigme actual, és a dir, augmentant indefinidament els corpus i la potència de càlcul, tot usant eines matemàtiques cada cop més sofisticades, resulti possible anar superant les fortes limitacions que tenen els sistemes de tractament de la parla. I encara que això fós possible, els recursos que s'hi haurien d'esmerçar podrien ser excessius en la majoria de situacions, ja que l'enfocament basat en aprenentatge obliga a produir nous corpus cada vegada que es canvia de llengua, de domini, d'entorn acústic, etc. Si bé és cert que la progressiva acumulació de documents textuals i sonors en servidors connectats a la Xarxa està possibilitant, per exemple, la incipient aparició d'una nova generació de sistemes de reconeixement de la parla relativament robustos a les fonts de variabilitat esmentades anteriorment, el paradigma predominant afavoreix les llengües o els entorns que disposen de més dades a la Xarxa, mentre que un enfocament més basat en coneixement probablement no crearia tants desequilibris.

---

[3] Greenberg, S. (2005). "From here to utility – Melding phonetic insight with speech technology". In W. Barry & W. Domelen (Eds.), *Integrating phonetic knowledge with speech technology*. Dordrecht, Netherlands: Kluwer, pp. 107-132. (http://silicon-speech.com/Media/PDF/2006_Greenberg_MultiTierTheory.pdf). Citat per J. Llisterri en l'article esmentat prèviament.

La prosòdia ens pot servir d'exemple. D'una banda, és absolutament necessària per aconseguir veus artificials que sonin de forma natural i expressiva, és important també per captar continguts pragmàtics en les interaccions orals amb un ordinador, i hauria de deixar de tenir un paper quasi nul en les tecnologies perceptives (reconeixement de parla i parlant, traducció, etc.) perquè aporta informació rellevant sobre el missatge i el parlant. D'altra banda, hi ha multitud de treballs de recerca en prosòdia realitzats per fonetistes, gairebé sempre al marge dels tecnòlegs. Si es disposés de models que incorporessin adequadament coneixement prosòdic en els sistemes de tractament de la parla, probablement es guanyaria en prestacions dels sistemes, i a més a més, la necessitat de corpus disminuiria.

L'augment de la interdisciplinarietat en la formació dels investigadors afavoriria la convergència de les "dues cultures", la tecnològica i la lingüística. No hi ha gaire exemples en aquest sentit. En el nostre entorn europeu tenim l'*European Masters in Language and Speech*[4], impartit per diverses universitats i que té l'aval de les dues grans associacions professionals de l'àrea tecnològica (una de parla i una de llenguatge natural). A la Universitat Politècnica de Catalunya s'hi han format en els darrers anys algunes desenes d'enginyers que, per obtenir adicionalment aquesta titulació europea no oficial, han hagut de cursar, a part de les assignatures tècniques, 9 crèdits de l'àrea lingüística i 3 de percepció del llenguatge, assistint a assignatures de dues universitats de Barcelona. Un segon exemple de formació interdisciplinària, aquest partint d'una base lingüística, es el *Máster de Fonética y Fonología*[5] de la UIMP, organitzat pel *Laboratorio de Fonética* del CSIC, ja que inclou una especialitat (de les quatre que ofereix) en tecnologies del llenguatge, la majoria de professors de la qual provenen del camp tècnic.

Per acabar tot obrint una mica la perspectiva, esmentarem la tendència actual en recerca a passar del tractament del llenguatge verbal estricte al tractament de la comunicació humana en totes les seves components, incloent així els moviments facials (especialment de boca i pupil·les), el llapis electrònic, o els gestos amb les mans, el cap i tot el cos (en particular els llenguatges de signes). Sembla clar que la multimodalitat rebrà cada cop més atenció en el camp de les tecnologies de la comunicació humana, especialment la combinació d'imatge, so i text. Tant en el vessant perceptiu (per exemple, el reconeixement de parla usant la veu i la imatge de la cara), com en el productiu (per exemple, en l'ùs d'imatges facials expressives sincronitzades amb el conversor de text en parla).

## Breu bibliografia per aprofundir en el tema

Una introducció, en català i força actualitzada, del tema de les tecnologies de la parla, les seves aplicacions i els seus productes, que inclou també àmbits especialitzats i en particular el de les necessitats especials:

– J. Llisterri (2009). "Les tecnologies de la parla". *Llengua, Societat i Comunicació*, 7, pp. 11-19. (http://www.ub.edu/cusc)

Per continuar llegint, en la línia del que s'ha exposat aquí, sobre els aspectes tècnics de la representació de la veu i el modelatge foneticoacústic en el reconeixement de la parla:

– J.B. Mariño, C. Nadeu (2004). "La representación de la voz para el reconocimiento del habla". *Tecnologías del texto y del habla*, M.A. Martí y J. Llisterri Eds. Barcelona: Edicions de la UB – Fundación Duques de Soria, pp. 187-224.

Un rar exemple de llibre tècnic d'autor lingüista, que conté alguns temes de les tecnologies de la parla:

[4] http://www.cstr.ed.ac.uk/emasters/
[5] http://www.estudiosfonicos.cchs.csic.es/

– J. Coleman (2005). *Introducing Speech and Language Processing*. Cambridge University Press.

I per a una descripció molt més detallada i completa de les tecnologies, si es disposa de suficient bagatge matemàtic, es pot consultar el llibre:

– X.D. Huang, A. Acero, H.W. Hon (2001). *Spoken Language Processing: A Guide to Theory, Algorithms and Systems*. Prentice Hall.

Per acabar, l'article que s'ha citat i usat profusament en el present escrit:

– J. Llisterri (2007). "El papel de la fonética en las tecnologías del habla". *Actas do 3o Congreso Internacional de Fonética Experimental*. Santiago de Compostela, 24-26 d'octubre de 2005, pp. 23-37.
(http://liceu.uab.es/~joaquim/publicacions/Llisterri_05_Fonetica_Tecnologias_Habla.pdf)