A SÍNTESE DE VOZ EN LINGUA GALEGA: O PROXECTO COTOVÍA

Manuel González González
Universidade de Santiago de Compostela

1. QUE SON AS TECNOLOXÍAS DA FALA

A fala é o medio de comunicación máis eficiente e máis natural que posuímos, por iso un dos principais obxectivos para a integración do mundo tecnolóxico na sociedade é o de dotar as máquinas de capacidade para emitiren e para interpretaren mensaxes orais. Este é o desafío fundamental das chamadas tecnoloxías da fala, que podemos definir como aquelas tecnoloxías que utilizan a fala como vehículo na comunicación home-máquina. O avance que se produciu no desenvolvemento dos microprocesadores, das técnicas de procesamento dixital do sinal e das telecomunicacións foron factores determinantes na expansión e auxe de tecnoloxías da fala como:

1. A *síntese de voz*. Consiste na conversión de textos escritos en textos orais dunha maneira totalmente automática, sen intervención humana.

2. O *recoñecemento de voz*. É o proceso complementario da síntese de voz, que permite converter unha secuencia oral nun texto escrito.

3. A *codificación da voz.* Ocúpase das formas de compresión da voz dixitalizada como sinal para a súa transmisión e almacenamento dunha maneira económica.

4. A *verificación da identidade de falantes*. Trátase da identificación dunha voz cun falante determinado ou da verificación desa identidade, analizando as características que individualizan a voz dunha persoa.

5. A *tradución* da lingua oral. Proceso de recoñecemento das mensaxes orais producidas nunha lingua determinada e a súa tradución a outra lingua distinta.

2. A IMPORTANCIA DAS TECNOLOXÍAS DA FALA NA SOCIEDADE DA INFORMACIÓN

A comunicación existiu sempre no seo das sociedades, pero nos últimos tempos, na chamada sociedade da información, producíronse cambios realmente importantes en distintos aspectos:

a) Fronte á comunicación das sociedades antigas, que tiña lugar fundamentalmente en escenarios que se reducían, na maior parte dos casos, ás relacións que se establecían entre os veciños dunha comunidade ou, de maneira menos frecuente,

entre membros de comunidades próximas, e, xa con carácter moito máis excepcional, entre certas persoas pertencentes a países distintos, as sociedades modernas caracterízanse precisamente pola universalidade do espazo comunicativo. Hoxe as comunicacións non teñen fronteiras, e unha empresa coa súa sede central en Xapón comunícase de maneira permanente cunha filial en Alemaña ou nos Estados Unidos de América.

b) A comunicación a distancia ata época moi recente, practicamente ata a invención e xeneralización do teléfono, era unha comunicación lenta e, polo tanto neste sentido, pouco eficiente: unha carta tardaba semanas e mesmo meses en percorrer espazos xeográficos por veces non moi distantes. Unha característica da comunicación e da transmisión de informacións no mundo actual é a inmediatez: prodúcese practicamente de xeito instantáneo ou cun gasto de tempo mínimo.

c) Un feito fundamental na sociedade en que vivimos é a importancia da economía: tendencia a lograr unha mellor relación custo-beneficio ou custo-rendemento. Por iso se comprende facilmente o éxito das tecnoloxías que teñen un custo baixo ou daquelas que, mesmo sendo onerosas no seu desenvolvemento, teñen unha explotación que resulta economicamente rendible ben pola capacidade de produción en serie ou ben porque a súa explotación resulta barata.

d) A sociedade actual está marcada fundamentalmente polo papel nuclear que está ocupando a informática, que produciu unha verdadeira revolución en moitos aspectos das nosas vidas: desde a transformación do sistema de traballo de moitas persoas, ata a incidencia na modificación da maneira de divertirse ou de relacionarse.

e) Se hai unha palabra que caracteriza os días que nos están tocando vivir, esa palabra é INTERNET, rede que nos permite acceder con facilidade a un volume de información que era impensable non hai moitos anos, que nos permite comunicarnos dun xeito económico con calquera punto do planeta, e que nos posibilita transmitir calquera información ao ciberespazo cun custo realmente moi baixo.

Todos estes son, entre outros, factores que propiciaron e facilitaron o desenvolvemento das tecnoloxías da fala en época recente. Hoxe, por exemplo, cando falamos de Internet, todos pensamos inmediatamente nun ordenador e nun teclado, porque a maioría dos usos que facemos dela é por medio de textos e ordes escritas, o que non deixa de ser bastante antinatural e, sen ningunha dúbida, antieconómico, xa que a maneira natural de expresármonos e de comunicármonos os humanos é a través da lingua oral. A utilización da lingua oral na interacción cos sistemas informáticos, á parte de máis natural, presenta evidentes vantaxes, como poden ser: a de prescindir do teclado en determinado tipo de instrumentos (como por exemplo, os asistentes dixitais), a posibilidade de utilizar o teléfono para realizar un gran número de tarefas, sen necesidade de estar fisicamente diante dun ordenador; a superación do estatismo que esixe a utilización do teclado, e a posibilidade de poder realizar outras actividades mentres se envía ou se

recibe información. A utilización da voz na maioría dos usos que fagamos de Internet será, sen dúbida, un feito que non tardará en xeneralizarse.

Realmente son innumerables as aplicacións das tecnoloxías da fala. Pensemos, por exemplo, nos ditáfonos para traballar especialmente con procesadores de textos (pero tamén con outros tipos de programas), en lectores de mensaxes electrónicas, en servizos telefónicos automáticos, na autorización oral de transaccións comerciais, en servizos que anuncian a chegada e saída de voos, trens ou autobuses dunha maneira automática; no ensino de linguas; en lectores de xornais ou doutro tipo de textos para persoas con discapacidade visual; en sistemas de operación de ordenadores por parte de persoas cegas; na subtitulación automática de programas de televisión para persoas con problemas auditivos; etc. etc.

3. IMPORTANCIA DAS TECNOLOXÍAS DA FALA NAS LINGUAS EN PROCESO DE NORMALIZACIÓN

Non é necesario reflexionar moito para decatarse da importancia que as tecnoloxías da fala e a chamada enxeñería lingüística van ter na sociedade do futuro. Os desenvolvementos da síntese de voz, do recoñecemento de voz, da tradución automática, dos sistemas de diálogo home-máquina, van producir unha verdadeira revolución de consecuencias en moitos aspectos imprevisibles na sociedade dos próximos decenios. É difícil de imaxinar unha sociedade desenvolvida que, en poucos anos, non teña totalmente automatizada a maioría dos servizos de información, ou que non teña dado pasos transcendentais de cara á integración de persoas con minusvalías auditivas ou visuais (que lles permitan, por exemplo, aos xordos poder seguir un programa de televisión con subtitulación automática; ou aos cegos poderen acceder mediante a síntese de voz á lectura fácil de calquera obra literaria ou dos xornais de cada día), ou que non incorpore ao seu sistema de educación todas as posibilidades que ofrecen estas novas tecnoloxías.

En non moitos anos a incorporación ou non das tecnoloxías da fala será un elemento decisivo na división entre linguas con capacidade para dar resposta ás necesidades do home do novo milenio e as que non teñan capacidade para facelo. E isto é transcendental, porque na percepción social identificarase con linguas máis útiles e linguas menos útiles.

Este feito afectará a todas as linguas, pero terá unha repercusión e transcendencia moito maior nas linguas que están en proceso de normalización. As linguas minorizadas que non sexan capaces de incorporar as novas tecnoloxías terán graves problemas de subsistencia. Permítanme que reproduza un parágrafo que escribín non hai moito tempo e que me parece que segue tendo plena vixencia:

> "Teño manifestado en moitas ocasións, dunha maneira teimuda, porque estou convencido da súa transcendencia, que as linguas nunha situación sociolingüística como a que vive o galego, que non collan o tren das novas tecnoloxías, son linguas cun horizonte moi escuro. Pola contra, as linguas minorizadas que consigan seguir o ritmo de integración dos

> avances das industrias da lingua terán dado un gran paso para a eliminación de desequilibrios xerados polo distinto potencial económico das sociedades de que son vehículo de expresión. As tecnoloxías da fala serán neste sentido, e en contra do que a primeira vista se poida pensar, un elemento democratizador, un elemento amortecedor de desequilibrios e, xa que logo, un elemento importantísimo de normalización lingüística."[1]

4. UNHA VELLA ASPIRACIÓN: QUE AS MÁQUINAS FALEN

Nas tecnoloxías da fala estamos a vivir un momento de desenvolvemento realmente transcendental. A fala é o medio normal de comunicación entre os seres humanos, pero nunca en épocas pasadas da historia se utilizara a fala para a comunicación entre os humanos e os produtos tecnolóxicos construídos polo propio home. Esta realidade, que está presente cada vez máis nas nosas vidas, e que está chamada a estalo aínda moito máis, é a consecuencia dun longo proceso de investigacións e en boa parte de frustracións levadas a cabo ao longo da historia recente.

Lograr que as máquinas falen é unha vella aspiración dos investigadores, que en diversas épocas da historia puxeron en marcha toda a súa imaxinación para trataren de lograr este obxectivo ou, polo menos, para se achegaren a el, mesmo que fose dunha maneira meramente simbólica.

Foi probablemente no século XVIII cando se levaron a cabo os primeiros intentos que podemos chamar científicos nesta dirección. No último cuarto deste século podemos salientar varias iniciativas de interese:

– O abade Mical (1730-1789) presenta as súas cabezas falantes. Tratábase de dous bustos falantes, que estaban conectados a unha máquina que era capaz de emitir dunha maneira escasamente intelixible algunhas frases, ao mesmo tempo que se producía un movemento dos labios das cabezas falantes. Mical fixo moitas presentacións do seu invento ante o público, e espertou certo interese nalgún grupo de físicos. No prospecto no que se anunciaban estas presentacións dicíase que unha das frases que emitían era: "Problème resolu en mécanique". O abade Mical acabou por romper as cabezas falantes.

– Kratzenstein (1723-1795) gaña o premio convocado en 1779 pola Academia Imperial das Ciencias de San Petersburgo para investigar sobre o tema da natureza das vogais *a, e, i, o, u* e a construción dun dispositivo capaz de emitilas. Kratzenstein demostrou que todas estas vogais podían ser producidas proxectando o aire dun fol a través de tubos de distintas formas.

– Tamén Wolfgang von Kempelen (1734-1804), conselleiro da Corte Real vienesa, construíu unha máquina falante, despois de máis de vinte anos de traballo. Pretendía que a máquina, o mesmo que o pode

1 González González, M. (2001): "A terminoloxía galega: un labor normalizador", en: X. L. Regueira & A. Veiga: *Da gramática ó diccionario. Estudios de lingüística galega.* Anexo 49 de *Verba.* Santiago, Universidade de Santiago de Compostela, páx. 166.

facer unha persoa, puidese falar en calquera lingua. Este dispositivo consistía nun tubo vocal que trataba de remedar as cordas vocais, e unha especie de gaita que emulaba os pulmóns, a boca e a cavidade nasal. Un dos principais problemas cos que bateu Kempelen foi o de atopar o material axeitado para construír a lingua, o padal, os labios e os dentes, órganos que desempeñan un papel fundamental na produción do son humano. Kempelen deixounos un libro precioso sobre esta investigación titulado *Le mécanisme de la parole, suivi de la description d'une machine parlante et enrichie de XXVII planches*, editado en Viena en 1791. A máquina descrita nesta obra é hoxe propiedade do Deutsches Museum de Múnic. Pero Kempelen deixounos tamén outra máquina de falar, en forma dun pupitre cun teclado, que se conserva no Technisches Museum für Industrie & Gewerbe de Viena. No ano 1939 presentouse unha versión electrónica do sistema deseñado por Kepelen, o VODER, cun teclado moi elaborado que permitía controlar a articulación e xeración de sons vocálicos e consonánticos.

Podemos mencionar outros intentos de autómatas falantes semellantes de Robert Willis, de Robertson, e, xa a mediados do XIX, do profesor Faber ou de J. B. Rechsteiner. Pero isto é, con ollos de hoxe, a prehistoria da síntese de voz.

5. A CONVERSIÓN TEXTO-VOZ

5.1. QUE É UN SINTETIZADOR DE VOZ

Un sintetizador de voz é unha ferramenta que permite a conversión dun texto escrito nunha cadea oral, de xeito que a transferencia texto-voz poida ser levada a cabo cunha calidade aceptable sen a intervención directa do falante.

5.2. ESIXENCIAS MÍNIMAS QUE HA DE REUNIR UN SINTETIZADOR DE VOZ

As esixencias dun conversor texto-voz dependen, en gran medida, da finalidade para a que estea destinado: un instrumento é útil cando cumpre con eficiencia a finalidade para a que foi construído. Pero hai unha serie de requirimentos mínimos que han de cumprir todos os sintetizadores de voz. Eu sinalaría fundamentalmente dous: a fiabilidade e a intelixibilidade.

a) Parece obvio que, se o que se trata é de reproducir oralmente o que aparece nun texto escrito, aquilo que se traslada á lingua oral debe ser equivalente "fiel" do contido no texto escrito.

b) Unha segunda esixencia imprescindible é a intelixibilidade. Non se pode perder nunca de vista que a finalidade dos conversores texto-voz é fundamentalmente práctica, con aplicacións fundamentais nos sistemas de comunicación, no mundo da empresa, no ensino..., e que, polo tanto, os enunciados producidos deben ser sempre facilmente "intelixibles"; mesmo, se é posible, máis intelixibles que os enunciados producidos pola propia voz humana.

Pero, á parte desas esixencias mínimas, a síntese de voz debe aspirar a acadar o nivel máis alto posible de *naturalidade*. Cando falamos de voz sintetizada, todos temos tendencia a pensar nunha voz monótona, a tradicional voz robótica, moi distante da flexibilidade e da variedade

que presenta a voz humana. E este é aínda un problema no que queda moito camiño por percorrer, aínda que tamén é certo que nos últimos anos se produciron avances moi notables neste sentido, sobre todo nos sintetizadores baseados en corpus.

5.3. MÉTODOS DE SÍNTESE

A síntese de voz intentouse resolver por distintos procedementos, e hoxe contamos con sintetizadores articulatorios, sintetizadores de formantes e sintetizadores baseados na concatenación de unidades pregravadas:

a) A síntese articulatoria é a que primeiro se intentou historicamente, porque resultaba a máis intuitiva, ao tratar de reproducir mecanicamente o aparello fonador do organismo humano. Pero é a que hoxe presenta un nivel máis baixo de desenvolvemento, e de feito na actualidade non contamos aínda con ningún sintetizador deste tipo que reúna as esixencias mínimas para poder ser utilizado.

b) A síntese por formantes baséase nunha fonte xeradora de son e de ruído, que é modificada por unha serie de filtros que modelan as resonancias do tracto vocal, xerando os formantes que caracterizan os distintos sons dunha lingua determinada.

c) A síntese por concatenación de unidades baséase na unión de unidades dixitalizadas pregravadas. É o método máis utilizado na actualidade porque é o que ofrece unha mellor relación entre complexidade e prestacións, e porque a utilización de díxonos simplifica moito os problemas derivados da coarticulación.

6. COTOVÍA: O PRIMEIRO SINTETIZADOR DE VOZ EN LINGUA GALEGA

6.1. COMO NACEU

No ano 1995 constituíuse un grupo de traballo de carácter interdisciplinar, ao abeiro do Centro Ramón Piñeiro para a Investigación en Humanidades (daquela chamado aínda "Centro de Investigación Ramón Piñeiro"), integrado por lingüistas da Universidade de Santiago de Compostela e por enxeñeiros de telecomunicacións da Universidade de Vigo. As dúas pólas do grupo estaban coordinadas por Manuel González González e por Carme García Mateo (hoxe substituída na coordinación polo seu compañeiro do Departamento de Teoría do Sinal Eduardo Rodríguez Banga)[2]. Foi nese momento cando se iniciaron os traballos de elaboración do primeiro conversor texto-voz en lingua gale-

2 Ó longo destes anos traballaron no proxecto un amplo número de investigadores:
a) Enxeñeiros de telecomunicacións: Carme García Mateo (coordinadora da Área de Enxeñería ata o ano 2004), Eduardo Rodríguez Banga (coordinador da Área de Enxeñería desde o ano 2004), Xavier Fernández Salgado, Leandro Rodríguez Liñares, Camilo Giráldez Ruibal, Francisco Méndez Pazó, Francisco Campillo Díaz e Gonzalo José Iglesias Iglesias.
b) Lingüistas: Manuel González González (coordinador), Elisa Fernández Rei, Rut Losada Soto, Elisa Roca Rodríguez, Luís Xuncal Pereira, Ana Martínez Ínsua, Lidia Gómez García e Ana Escourido Pernas.

ga, que no primeiro momento se designou co nome de "Monchiño", en honra a Ramón Piñeiro, personaxe que daba nome ao centro que acolleu o grupo e a idea. Máis tarde, cando o grupo levou a cabo outros proxectos, acordou nomear os seus produtos con nomes de paxaros, e este conversor pasou a coñecerse co nome de CoToVía, xa que esta palabra contiña as consoantes CTV que podían ser as iniciais de C(onversor) T(exto) - V(oz).

6.2. A SÚA ARQUITECTURA

Cotovía é un conversor texto-voz baseado na técnica de concatenación de unidades, que é o sistema máis utilizado hoxe en día e o que se aproxima máis á calidade da voz humana.

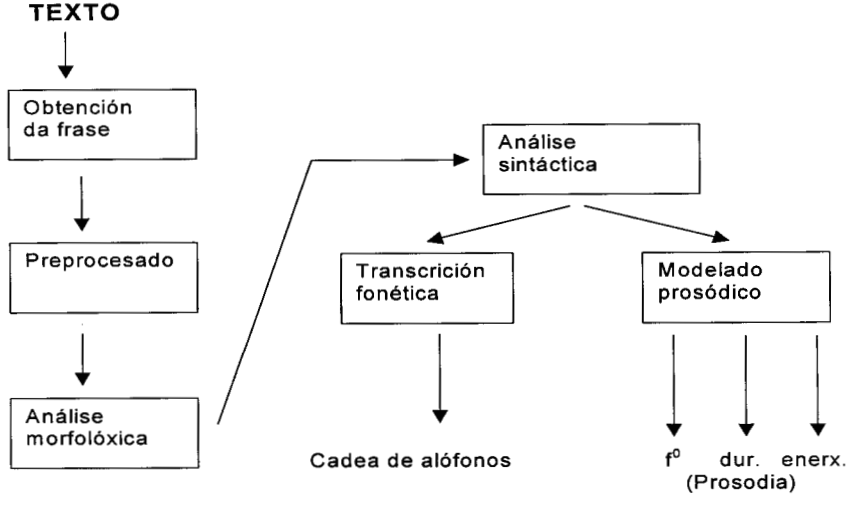
Está integrado por dous grandes módulos: un lingüístico e outro acústico.

A función fundamental do módulo lingüístico é a de xerar unha transcrición fonética, tanto segmental coma suprasegmental (e para iso debe botar man de todos os recursos de análise que lle proporcionen a información necesaria para poder levar a cabo esta transcrición dunha maneira correcta).

A función fundamental do módulo acústico é a de almacenar as unidades que, concatenadas, darán lugar á xeración da lingua sintetizada, e a de aplicar o algoritmo de concatenación das distintas unidades.

6.3. O MÓDULO LINGÜÍSTICO

O módulo lingüístico é a parte do sistema encargada de lle subministrar ao módulo acústico toda a información lingüística necesaria para que este leve a cabo a correcta selección das unidades, e aplique sobre estas as transformacións necesarias para a asignación da prosodia que corresponde.



6.3.1. A OBTENCIÓN DA FRASE. O primeiro paso que dá o sistema ao recibir o texto de entrada é seleccionar unha parte del para tratalo. Para seleccionar estas frases bótase man dos diferentes signos de puntuación e doutros caracteres como poden ser as marcas de final de parágrafo. Pero dentro desta fase lévanse a cabo tamén outras operacións:

– a separación das palabras e dos signos de puntuación que integran este fragmento;

– unha clasificación provisional das palabras, segundo sexan voces normais, abreviaturas, siglas ou acrónimos, datas, cadeas numéricas, etc.;

– unha clasificación inicial da modalidade do enunciado baseándose unicamente nos signos de puntuación: enunciativos, interrogativos, exclamativos.

6.3.2. O PREPROCESADO LINGÜÍSTICO realiza todas as actividades tendentes á normalización da frase, para facela intelixible polo sistema nas accións que debe desenvolver a partir deste momento. Nesta fase lévanse a cabo accións como:

a) Extensión completa de todo tipo de texto que presente calquera forma de abreviación: desenvólvense as abreviaturas e siglas (mediante unha lista a xeito de pequeno dicionario que contén as máis habituais). Por ex., *BSCH* converterao en *Banco Santander Central Hispano.*

b) Interpretación dos números, tanto arábigos coma romanos, aplicando unha serie de regras para a súa lectura. Por exemplo*: 2789 €* converterao en *dous mil setecentos oitenta e nove euros*.

c) Un caso particular é o de cando os números representan horas, o que esixirá ás veces unha lectura especial, sobre todo cando aparecen minutos e segundos. Por exemplo, *ás 16:32:45* cambiarase en *ás dezaseis horas, trinta e dous minutos e corenta e cinco segundos*; ou *saíu ás 14h 16' 15"* cambiarase en *saíu ás catorce horas, dezaseis minutos e quince segundos.*

d) Tamén entra neste proceso de normalización do texto da frase a busca dunha lectura o máis correcta posible dos estranxeirismos, xa que nestas palabras a correspondencia grafema-son non é a mesma que para as palabras galegas. Para solucionar este problema o preprocesado lingüístico conta cunha lista de estranxeirismos e coa súa realización fonética.

O preprocesado lingüístico debe ofrecer certa flexibilidade e facilidade de manipulación, xa que a lista de estranxeirismos, de siglas, e todo tipo de abreviaturas pode variar moito segundo o tipo de texto ao que se destine o sintetizador. Nos prototipos que están funcionando neste momento recóllense os elementos deste tipo máis habituais, pero que poden ser modificados en virtude das necesidades que deba cubrir o conversor texto-voz.

6.3.3. A ANÁLISE MORFOLÓXICA

A análise morfolóxica constitúe unha etapa clave dentro da información lingüística que necesita o sistema, xa que os seus datos serán fundamentais tanto para a posterior análise sintáctica como para a transcrición fonética.

Nesta etapa asígnase a categoría gramatical, o xénero e o número que lle

corresponden a cada palabra. En primeiro lugar extráense e etiquétanse as palabras pertencentes a series cerradas (como preposicións, conxuncións, adverbios, determinantes...). A continuación procédese, mediante a análise das raíces e dos morfemas, a determinar a categoría de palabras que pertencen a series abertas (nomes, adxectivos, verbos). A cada forma atribúenselle todas as posibles categorías que pode ter (por exemplo, *rapa* pode ser potencialmente tanto un *substantivo feminino singular*, como a *P*[3] do *presente de indicativo do verbo "rapar"*), e posteriormente será necesario realizar a desambiguación. Procédese tamén a analizar as posibles agrupacións de palabras como as locucións adverbiais, as locucións conxuntivas ou as perífrases verbais.

Para a desambiguación e a asignación da categoría definitiva a cada palabra recórrese á análise contextual, é dicir, á observación das categorías das palabras anteriores e posteriores e á aplicación dun modelo estatístico que nos permitirá asignar a categoría máis probable entre todas as posibles.

Tamén neste etapa se procede á división en sílabas e á asignación do acento nas palabras. O algoritmo de división silábica está baseado na detección do núcleo silábico, que é sempre unha vogal. As fronteiras silábicas establécense observando a consoante ou consoantes que rodean o núcleo silábico e tendo en conta posibles ditongos e tritongos. Este algoritmo é suficientemente potente como para tratar non só as secuencias de fonemas que aparecen nas voces patrimoniais, senón tamén as menos frecuentes que poden aparecer en cultismos e estranxeirismos.

No proceso de análise morfolóxica desempeña un lugar fundamental o analizador verbal (Laverca)[3], que se encarga de detectar as diferentes formas verbais conxugadas que aparecen no texto, e de asignarlle a cada unha o tempo, modo, persoa e número, que lle corresponde. Pero este analizador realiza outras funcións importantísimas:

a) segrega do verbo as formas enclíticas (sexan estas formas átonas do pronome persoal ou as chamadas segundas formas do artigo) e reconstrúe a forma de base do verbo, tal como sería sen a transformación que sofre pola amálgama dos elementos enclíticos (por ex., na forma *cómelo* xebra a forma da P[2] do presente de indicativo *comes* do pronome persoal enclítico *lo*).

b) aplica unha serie de regras de atribución de timbre ás formas verbais que teñen vogais de grao medio *e* ou *o*. No caso anterior sabe que unha forma como *comes* debe ter unha vogal tónica de timbre aberto: *kòmes*.

3 Con este analizador verbal publicamos un *Diccionario de verbos galegos*, co subtítulo de *Laverca*, (Edicións Xerais de Galicia, Vigo, 2002) que leva incorporado un CD-ROM cun programa informático que permite a análise automática de calquera forma verbal, a extracción das formas verbais dun arquivo de texto e a súa análise, e a conxugación automática de todos os verbos da lingua galega. Esta obra foi recoñecida co Premio da Crítica do ano 2003, na modalidade de investigación.

6.3.4. A ANÁLISE SINTÁCTICA trata de establecer a estrutura do enunciado, examinando as oracións, as proposicións existentes e o tipo de relación existente entre estas, os sintagmas que integran as proposicións e as palabras que constitúen cada sintagma. Neste proceso é de grande axuda o exame dos signos de puntuación e de palabras clave como as conxuncións ou locucións conxuntivas, os pronomes relativos, etc.

Na etapa de análise sintáctica tamén se realiza un proceso de reacentuación, e procédese á eliminación do acento nas palabras función, así como á inserción de pausas. A información para a inserción das pausas vén dada, na maior parte dos casos, polos signos de puntuación, pero cando nos encontramos con proposicións que exceden determinada lonxitude cómpre inserir pausas adicionais entre sintagmas, do mesmo xeito que o faría un falante normal. Estas pausas responden, fisioloxicamente, á necesidade de inspirar e, de non se consideraren na voz sintetizada, produciríase na persoa que escoita unha sensación de artificialidade ou de fala fatigosa.

6.3.5. A TRANSCRICIÓN FONÉTICA. Nesta etapa transfórmase o texto que ata este momento aparecía escrito en caracteres propios da escritura normal nunha cadea de caracteres fonéticos que representan os alófonos, e que servirán para seleccionar as unidades acústicas sobre as que se construirá a fala sintetizada. A transcrición fonética automática nunha lingua coma o galego é bastante máis simple ca, por exemplo, no inglés, porque no galego existe unha correlación moito maior entre grafemas e fonemas da que existe naquela lingua; pero é bastante máis complexa ca no castelán, lingua na que esta correlación á que nos acabamos de referir é aínda maior ca no galego. Sen pretender entrar en detalles dos problemas que presenta a transcrición fonética automatizada en lingua galega, calquera persoa pode decatarse da complexidade que ten asignarlles o timbre correcto en cada caso ás vogais medias *e*, *o*; ou adxudicarlle a transcrición correcta ao grafema *x*, que unhas veces representa a consoante fricativa postalveolar xorda [ʃ] (*xente*) e outras o grupo consonántico [ks] (*xenofobia*).

6.3.6. O MODELADO PROSÓDICO ten como finalidade asignarlle a cada alófono a súa correspondente duración, enerxía e evolución da frecuencia fundamental. Esta etapa é a que vai determinar en boa medida a naturalidade da voz sintetizada. Utilízanse a sílaba e o grupo acentual como unidades prosódicas básicas e realízase a asignación ás vogais de padróns de duración e frecuencia fundamental dependendo do tipo de proposición, da posición do grupo acentual dentro da proposición (tres tipos: inicial, medio e final) e da posición do acento dentro do grupo acentual (oxítono, paroxítono e proparoxítono). Para as consoantes só se realiza o modelado da duración, xa que o contorno de F^o obtense por interpolación entre os valores das vogais.

Os padróns de duración e frecuencia fundamental foron obtidos a partir dun conxunto de frases deseñadas para tal fin. O procedemento consistiu en aproximar o contorno de frecuencia fundamental nas

vogais mediante dous ou máis segmentos, resintetizar a frase co novo contorno, e decidir tras escoitar a frase modificada se a aproximación era axeitada ou non. Exploráronse tamén outros modelos prosódicos, coma o de Fujisaki, que permite descompoñer o contorno de frecuencia fundamental en dúas contribucións distintas: a compoñente de frase, que depende do tipo de proposición, e a compoñente acentual, que é función da posición dos acentos.

Na versión de Cotovía baseada en corpus considérase máis dunha curva de frecuencia fundamental, tirando maior rendemento das distintas realizacións de cada unidade que existe no corpus, e xerando unha prosodia que depende tamén das propias unidades. Con isto lógrase unha maior naturalidade e decrece a distorsión típica dos algoritmos de modificación prosódica, xa que, na maioría dos casos, non será necesaria. Así e todo, non debemos esquecer que na prosodia interveñen tamén factores non lingüísticos, como a actitude do locutor, a intencionalidade, o estado físico e emocional... que dificilmente serán plasmables nun sintetizador de voz, polo que a naturalidade da voz artificial sempre vai estar lonxe da da voz natural.

6.4. O MÓDULO ACÚSTICO

O módulo acústico é a parte do conversor onde está almacenada a materia fónica, os segmentos que se utilizarán para a concatenación que xerarán a fala sintetizada, os modelos prosódicos e os algoritmos de concatenación de unidades.

6.4.1. PROBAS PREVIAS PARA A ELECCIÓN DE LOCUTOR

Unha decisión importante na elaboración deste tipo de conversores baseados na concatenación de unidades é a escolla do locutor ou locutores que han gravar as unidades lingüísticas consideradas e o corpus para a modelaxe da prosodia. Debemos ter en conta que a voz resultante da síntese soará, en boa parte, dunha maneira semellante á voz do doador das unidades coas que se vai construír.

A selección dos locutores (tanto masculinos coma femininos) fíxose despois dunha avaliación das características da súa voz. Por exemplo, antes da selección definitiva do locutor masculino para a extracción das unidades de logátomos fíxoselles unha proba de gravación dun corpus reducido a seis locutores distintos previamente seleccionados pola súa boa dicción, de voz grave, xa que este tipo de voces é máis resistente ás modificacións prosódicas e ás manipulacións ás que son sometidas as unidades. Sintetizouse o texto de cada un dos locutores e, despois de analizar o resultado, escolleuse o locutor que posuía a calidade de fala máis satisfactoria.

6.4.2. DESEÑO DAS UNIDADES

Para a construción de Cotovía optouse, como xa se dixo, por un sistema baseado na concatenación de dífonos, que presenta vantaxes notables sobre todo na superación do problema da variabilidade producida polos fenómenos da coarticulación. Na fala natural as realizacións dos fonemas están fortemente condicionadas

pola presenza dos outros cos que está en contacto, que fan que na articulación a posición da lingua e dos labios cambie dun a outro dunha maneira gradual. A representación da coarticulación, fundamental tamén para a *naturalidade* da fala sintetizada, soluciónase en gran medida traballando con difonos, dado que mediante ese método nos aseguramos de que temos recollidas en todos os casos as transicións duns sons a outros.

É fundamental facer un bo deseño do corpus que se vai gravar, que sexa o máis económico posible, pero que conteña o número suficiente de unidades para dar resposta ás necesidades da voz sintetizada. Foi necesario decidir que alófonos se ían ter en conta, á vista do seu rendemento e necesidade, sempre mediante a comprobación perceptiva. Por exemplo, non se terá en conta a realización sonorizada [ʂ] do fonema fricativo alveolar en posición implosiva, senón unicamente a realización xorda [s̬], xa que ao ser a sonorizada unha realización sempre condicionada polo contexto (pola presenza dunha consoante sonora a continuación), ao introducir nestes casos o difono seguinte composto pola segunda parte do fonema *s* e a primeira parte dun fonema sonoro, a realización sonorizada desta segunda parte é suficiente para percibir a realización como natural. Por exemplo, para sintetizar a palabra *esgotar*, no grupo consonántico *sm* hai que unir unha unidade que constará da segunda parte do *e* e da primeira parte do *s* a outra unidade que conterá a segunda parte do *s* e a primeira parte do *g*. Pois ben, no sintetizador de logátomos, só aparecerá sonorizada a parte do *s* correspondente á segunda unidade, pero non a correspondente á primeira: en realidade faise a fusión dun fragmento [ʂ] (xordo) con outro [ʂ] sonorizado. Pola mesma razón non se tiveron en conta as vogais nasalizadas, certas realización do arquifonema /N/ nin outros alófonos do fonema fricativo /s/.

No deseño do corpus tampouco se considerou a necesidade de introducir unidades con semivogais e semiconsoantes, xa que estas poden ser obtidas a partir da manipulación de realizacións das vogais *i* e *u*.

Foi, en cambio, imprescindible considerar separadamente os alófonos aproximantes e oclusivos dos fonemas /b/, /d/, /g/. E acordouse incluír o alófono [ħ] do subsistema con gheada, que non se vai reproducir no conversor, pero que se utilizará na pronunciación de palabras estranxeiras como *hippy* ou *hándicap*.

Nalgúns casos optamos por incorporar unidades superiores aos difonos. É de sobra coñecido na literatura da síntese de voz a dificultade que ofrece a segmentación axeitada das consoantes vibrantes, de maneira especial cando estas forman un grupo tautosilábico con outra consoante; de aí que, nestes casos, se optase por incorporar trífonos, que evitasen a partición da vibrante, en secuencias do tipo *pra*, *tre*, *ard*, etc.

6.4.3. A OBTENCIÓN DAS UNIDADES

Para a obtención das unidades estudáronse dúas posibilidades:

a) Tomalas da fala natural, que ten a vantaxe de que a pronunciación é máis

distendida e natural, pero que ten tamén inconvenientes moi serios, como realizacións excesivamente relaxadas, problemas de segmentación á hora de etiquetar, e dificultade para atopar determinados contextos en palabras do léxico común galego.

b) Extraer as unidades de logátomos, é dicir, de palabras artificiais sen significado. Este método ten as vantaxes de que o contexto pode ser especialmente seleccionado e coidado, as realizacións adoitan ser realizadas dunha maneira máis uniforme, e a etiquetaxe das unidades é máis doada e máis exacta ca na fala natural. Pero ten o inconveniente de que a lectura do corpus é realmente difícil e fatigosa, e ademais esixe un locutor adestrado, que sexa quen de reproducir as realizacións que se lle esixen.

Na primeira versión de Cotovía optouse por traballar con unidades extraídas de logátomos. A base de datos contén unha soa realización dunhas 1.100 unidades, entre difonos e trífonos; nunha primeira etapa fíxose só con voz masculina, pero máis tarde, cando pareceu conveniente que o sintetizador funcionase tamén con voz feminina, recolléronse outras tantas unidades dunha informante feminina. O acento do logátomo recae na vogal da unidade extraída para evitar o relaxamento da realización. As unidades que se extraen son cortadas pola súa parte máis estacionaria, para conseguir a maior estabilidade posible.

O modelo máis recente funciona con unidades extraídas de corpus. O corpus actual conta cuns 120.000 semifonemas, clasificados en 970 grupos segundo o fonema de que se trate, se é a metade dereita ou a esquerda, e segundo o fonema co que esta metade está en contacto. Esta versión de Cotovía, que ten un custo computacional moito máis alto ca a baseada en logátomos ofrece un nivel moito maior de naturalidade ca a primeira.

6.4.4. A BASE DE DATOS PROSÓDICA

A prosodia é un elemento fundamental tanto pola función lingüística que desempeña (unha frase interrogativa non é o mesmo ca unha enunciativa) como polo efecto que exerce sobre a naturalidade da voz sintetizada. O obxectivo ideal é elaborar un conversor que presente un bo número de padróns entoativos que lle confiran naturalidade e variedade ao discurso producido, xa que unha das sensacións máis frecuentes cando se escoita un conversor é a monotonía da súa fala, o que provoca o cansazo e consecuente rexeitamento do oínte.

Para extraer os padróns entoativos, elaborouse un corpus, gravado por varios locutores, que recolle unha serie de exemplos de cada unha das modalidades oracionais predeterminadas.

Na versión máis recente do sintetizador, o corpus de frases consta de 1.300 enunciados, dos que 800 foron deseñados manualmente para reflectir aquelas estruturas prosódicas consideradas máis frecuentes e máis relevantes dentro do galego. Nesta parte téñense en conta as distintas modalidades (enunciativas, interrogativas, exclamativas, suspensivas), a distri-

bución dos grupos fónicos dentro do enunciado, e os grupos acentuais de cada grupo fónico. As outras 500 frases están tomadas de texto libre, que enriquece a variedade de estruturas prosódicas.

6.4.5. GRAVACIÓN, SEGMENTACIÓN E ETIQUETAXE DAS UNIDADES

As frases e os logátomos do corpus graváronse nun DAT e foron transferidas a un PC a través dunha tarxeta de son. Foi necesario un cambio de formato para permitir a edición nunha estación *Sun* con sistema operativo *Unix*.

A etiquetaxe efectuouse co programa SFS (*Speech File System*) do University College of London, que permite a edición de ficheiros de voz, a introdución de marcas nestes, así como a obtención de espectrogramas e curvas de frecuencia fundamental.

En cada unha das frases introducíronse marcas ao comezo e final de cada dífono, e unha marca adicional na fronteira entre os sons. Para a notación utilizouse a adaptación feita por nós mesmos para o galego do SAMPA (*Speech Assessment Methods Phonetic Alphabet*).

Os criterios de marcaxe seguidos atenden á estrutura de cada un dos sons que forman parte da unidade, e varían segundo se trate dunha vogal, dunha oclusiva, dunha africada, dunha fricativa ou dunha aproximante.

O programa que extrae as unidades das frases leva a cabo unha serie de tarefas, como:

– a lectura das marcas das frases portadoras das unidades que se utilizarán para a síntese;

– a localización e extracción das unidades;

– o decimado e filtraxe paso-baixo das unidades para baixar a frecuencia de mostraxe de 16 KHz a 8 KHz;

– a construción dos ficheiros que conteñen a información da disposición das unidades nas bases que as aloxan.

6.4.6. A CONCATENACIÓN DE UNIDADES

Os conversores texto-voz baseados na concatenación de unidades obtidas da fala natural necesitan utilizar algunha técnica que permita a unión das distintas unidades sen que se dean transicións bruscas, e que permita modificar as características prosódicas destas unidades. Para isto existen diversos métodos: *métodos baseados en predición lineal*, *métodos de suma solapada sincrónica coa frecuencia fundamental* (con diversas variantes coma os coñecidos como PSOLA[4], ou o TD-PSOLA[5]), *métodos baseados en modelado sinusoidal*. En Cotovía partiuse dun método baseado en modelado sinusoidal, que foi evolucionando para aproveitar os aspectos máis vantaxosos doutras abordaxes.

4 Pitch Synchronous Overlap-Add.
5 Time Domain Pitch Synchronous Overlap-Add.

7. CONCLUSIÓN

As tecnoloxías da fala van condicionar, sen ningunha dúbida, o papel que desempeñarán as linguas no futuro. As linguas que non acaden un desenvolvemento mínimo neste campo terán dificultades moi serias para poderen responder ás necesidades que a sociedade do séc. XXI lles vai esixir. Nos últimos anos estase traballando na elaboración de distintas ferramentas para a lingua galega, tanto de síntese de voz, coma de recoñecemento de voz, sistemas de diálogo home-máquina, tradución automática, etc. Destes, o ámbito máis avanzado quizais sexa o da conversión texto-voz, para o que neste momento contamos con dous sistemas en funcionamento, un elaborado por Telefónica I + D e o outro coñecido como Cotovía, realizado no Centro Ramón Piñeiro para a Investigación en Humanidades por un equipo interdisciplinar integrado por profesores e investigadores da Escola Técnica Superior de Enxeñeiros de Telecomunicacións da Universidade de Vigo e da Facultade de Filoloxía da Universidade de Santiago. Deste último proxecto están funcionando dúas versións, unha baseada en logátomos e a outra baseada en corpus, coa opción de seleccionar voz masculina ou voz feminina, e que poden ser consultadas no enderezo electrónico http://www.gts.tsc.uvigo.es/cotovia/, ou a través da páxina web do Centro Ramón Piñeiro para a Investigación en Humanidades http://www.cirp.es

BIBLIOGRAFÍA

FERNÁNDEZ REI, E.; GONZÁLEZ GONZÁLEZ, M. (1998): "Un sintetizador de voz para el gallego", *Travaux de linguistique hispanique* (sous la direction de Gilles Luquet), 65-76, Paris, Presses de la Sorbonne Nouvelle.

FERNÁNDEZ SALGADO, X.; RODRÍGUEZ BANGA, E. (1998): "Análisis de duraciones para su aplicación en un Conversor Texto-Voz", *Actas del Congreso URSI 98*. Pamplona

GARCÍA MATEO, C. (2002): "Recursos e actividades necesarias para desenvolver tecnoloxía da fala en galego". En: BUGARÍN LÓPEZ, Mª. Xesús e outros (2002): *Actas da VIII Conferencia internacional de linguas minoritarias* (Santiago de Compostela, 22, 23, 24 de novembro de 2001). Santiago, Xunta de Galicia, páxs. 151-156.

GARCÍA-MATEO, C. and GONZÁLEZ-GONZÁLEZ, M. (1998): "An Overview of the Existing Language Resources for Galician", *LREC Workshop: Language Resources for European Minorities Languages.* Granada. 28-30 May 1998.

GONZÁLEZ GONZÁLEZ, M. (2002): "Laverca: diccionario de verbos gallegos con voz sintetizada", en DÍAZ GARCÍA, J. (ed.): *Actas del II Congreso de fonética experimental*, 209-214. Sevilla, Universidad de Sevilla.

GONZÁLEZ GONZÁLEZ, M.; LOSADA SOTO, R. FERNÁNDEZ REI, E. (1999): "O galego e as tecnoloxías da fala: o caso do sintetizador de voz", *Actas do V Congreso Internacional de Estudios Galegos*, 2, 703-716, Trier, Edicións do Castro / Galicien-Zentrum der Universität Trier.

GONZÁLEZ GONZÁLEZ, M.; GARCÍA MATEO, C.; RODRÍGUEZ BANGA, E.; FERNÁNDEZ REI, E. (2002): *Diccionario de verbos galegos*. *Laverca* (Contén CD-ROM co Programa Laverca 1.0. Vigo, Edicións Xerais de Galicia.

GONZÁLEZ REI, B. (2000): "Diseño de una base de datos tipo SpeechDat para el idioma gallego", *Procesamiento del Lenguaje Natural*, nº 24.

LOSADA SOTO, R. Mª (1996): "Adaptación do *SAMPA* para a lingua galega", *I Congreso Internacional A Lingua Galega. Historia e actualidade* (16-20 setembro 1996), en prensa.

RODRÍGUEZ BANGA E.; FERNÁNDEZ SALGADO, X.; BALBOA ANDRES, A.; CHAPELA VILLANUEVA, P. (1998): "Modelado de la entonación en un conversor texto-voz mediante el modelo de Fujisaki", *Actas del Congreso URSI 98*. Pamplona.

RODRÍGUEZ BANGA, E.; FERNÁNDEZ SALGADO, X.; FERNÁNDEZ REI, E.; GONZÁLEZ GONZÁLEZ, M. (1998): "Análisis lingüístico para un conversor texto-voz en lengua gallega", *Novática Revista de la Asociación de Técnicos de Informática*, núm. 133. Maio-xuño 1998, páxs. 40-45.

RODRÍGUEZ BANGA, E.; GARCÍA MATEO, C.; FERNÁNDEZ SALGADO, X. (2001): "Concatenative Text-to-Speech Synthesis based on Sinusoidal Modelling" in *Improvements in Speech Synthesis*. John Wiley and Sons, Ltd., páxs. 39–51.

♦♦♦

Manuel GONZÁLEZ GONZÁLEZ: "A síntese de voz en lingua galega: o proxecto Cotovía", *Revista Galega do Ensino*, núm. 44, novembro 2004, pp. 199-215.

Resumo: a fala é o medio de comunicación máis eficiente e máis natural que posuímos, por iso un dos principais obxectivos para a integración do mundo tecnolóxico na sociedade é o de dotar as máquinas de capacidade para emitiren e para interpretaren mensaxes orais. Un dos instrumentos máis avanzados é o sintetizador de voz. Un sintetizador de voz é unha ferramenta que permite a conversión dun texto escrito nunha cadea oral, de xeito que a transferencia texto-voz poida ser levada a cabo cunha calidade aceptable sen a intervención directa do falante. Este artigo describe o primeiro sintetizador de voz en lingua galega. Cotovía é un conversor texto-voz baseado na técnica de concatenación de unidades, que é o sistema máis utilizado hoxe en día e o que se aproxima máis á calidade da voz humana.

Palabras clave: dixitalización, novas tecnoloxías, sociedade da información, sintetizador de voz, tradución, procesamento lingüístico, modelado.

Resumen: el habla es el medio de comunicación más eficiente y más natural que poseemos, por eso uno de los principales objetivos para la integración del mundo tecnológico en la sociedad es el de dotar a las máquinas de capacidad para emitir e para interpretar mensajes orales. Uno de los instrumentos más avanzados es el sintetizador de voz. Un sintetizador de voz es una herramienta que permite la conversión de un texto escrito en una cadena oral, de manera que la transferencia texto-voz pueda ser llevada a cabo con una calidad aceptable sin la intervención directa del hablante. Este artículo describe el primer sintetizador de voz en lengua gallega. *Cotovía* es un conversor texto-voz basado en la técnica de concatenación de unidades, que es el sistema más utilizado hoy en día y el que se aproxima más a la calidad de la voz humana.

Palabras clave: digitalización, nuevas tecnologías, sociedad de la información, sintetizador de voz, traducción, procesamiento lingüístico, modelado.

Abstract: speech is the most efficient and most natural communication mean. Consequently, one of the main aims for the integration of the technological world into society it is to give machines the capability for sending out and to interpret oral messages. One of the most advanced instruments it is the voice synthesizer. A voice synthesizer is a tool that allows the conversion of a written text into an oral chain, so that the text-voice transference could be done with acceptable quality and without the direct intervention of the speaker.
This article describes the first voice synthesizer in Galician language. *Cotovía* is a text-voice converter based in the technique of the unities concatenation —the most used system nowadays, as it is the most similar one to human voice.

Key words: digitalisation, new technologies, information society, voice synthesizer, translation, linguistic processing, modelling.

– Data de recepción da versión definitiva deste artigo: 16-07-2004.