RECOÑECEMENTO AUTOMÁTICO DA FALA: IDEAS BÁSICAS E ALGÚNS EXEMPLOS ILUSTRATIVOS

Carmen García Mateo e Antonio Cardenal López

Departamento de Teoría do Sinal e Comunicacións
ETSE de Telecomunicación – Universidade de Vigo

1. Introdución

Desde os primeiros traballos sobre recoñecemento de díxitos realizados alá polos anos cincuenta ata os sistemas modernos de ditado e de transcrición de noticias, percorreuse un longo camiño na investigación relacionada co recoñecemento automático de fala. Do mesmo xeito que sucedeu no campo da intelixencia artificial, co que está moi relacionado, o recoñecemento de voz non respondeu totalmente ás expectativas que se depositaron nel. Aínda hoxe é complicado, e ás veces incluso frustrante, interactuar mediante voz dunha forma natural coas máquinas. Moitas veces vémonos limitados a aplicacións con vocabularios reducidos e específicos para a tarefa, ou a sistemas de ditado, que requiren unha fala coidadosa e precisan frecuentes correccións.

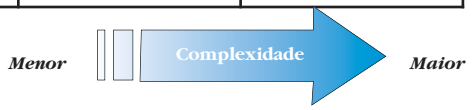
Con todo, é innegable que os avances que nesta dirección se observaron nos últimos anos foron importantes. Nestas case seis décadas de investigación asistiuse á formulación de tarefas cada vez máis ambiciosas, que aos poucos foron solucionándose de forma máis ou menos eficaz segundo as restricións impostas pola tecnoloxía do momento. De forma progresiva, o recoñecemento de voz ampliou o seu rango de acción desde as primeiras aplicacións de palabras illadas, limitadas a un único usuario e cun vocabulario moi restrinxido, ata os sistemas modernos, independentes do locutor e capaces de recoñecer fala continua con vocabularios de varias decenas de miles de palabras. Ao mesmo tempo o recoñecemento automático de voz foise incorporando á vida diaria. En aplicacións relacionadas con telefonía é común atopar sistemas de consulta de información, de reserva de billetes de avión, accesos a bases de datos, e moitos outros servizos que utilizan esta tecnoloxía como base. No ámbito doméstico, a interacción mediante voz ofrécese como un valor engadido en diversos electrodomésticos e dispone xa de sistemas comerciais de ditado, que utilizan vocabularios extensos e evitan, me-

diante técnicas de adaptación ao locutor, os tediosos procesos de adestramento que eran comúns hai só uns anos.

Xa no campo da investigación, as tarefas que se están abordando actualmente pódense definir como altamente complexas. A interacción natural home-máquina segue sendo un obxectivo prioritario, centrándose o esforzo no desenvolvemento de sistemas de diálogo eficientes máis que nas melloras dos recoñecedores, cuxas prestacións son xa suficientes. Os sistemas de recoñecemento actuais son capaces de lograr taxas de recoñecemento extremadamente altas en aplicacións complexas (fala continua e grandes vocabularios), pero só con condicións controladas, é dicir, cando a gravación é limpa e contén voz previamente segmentada e, ademais, tanto o locutor, como o vocabulario e o tipo de linguaxe son coñecidos e están ben modelados. Cando unha ou varias destas condicións non se cumpren, os recoñecedores de voz reducen as súas prestacións de forma importante (Shriberg 2005).

No Cadro 1 amósase esquematicamente o campo de traballo en función de dous parámetros: a calidade obtida medida como "taxa de recoñecemento" que mide dalgún xeito a porcentaxe de termos correctamente recoñecidos, fronte á dificultade da tarefa.

Tipo de Fala	Planeada	Espontánea
Calidade do audio	CD	Telefónica
Tamaño Vocabulario	Centos	Decenas de miles
Adaptación ao locutor	Monolocutor	Multilocutor



Cadro1: *Clasificación dos sistemas de recoñecemento de fala*

Neste artigo amosaremos cal é o estado actual da tecnoloxía empregada nos modernos sistemas de recoñecemento de fala desde o punto de vista da investigación, facendo tamén referencia a desenvolvementos e aplicacións

que xa deixaron os laboratorios para pasaren a formar parte de produtos comerciais.

2. Panorámica do estado da arte

De forma esquemática pódense identificar o seguintes campos de aplicación:

— Sistemas de comunicación home-máquina
  • Centrais telefónicas automáticas
  • Sistemas de ditado
  • Domótica
— Subtitulado automático de programas de televisión
— Transcrición automática de reunións
— Transcrición automática de conversacións telefónicas
  • Aplicacións en intelixencia civil e militar
— Sistemas de tradución voz a voz
  • Liña moi potenciada pola Unión Europea

Aínda que os avances nos últimos anos foron importantes, os sistemas de recoñecemento máis avanzados como o Byblos BBN en USA (Nguyen et alii 2002), o sistema do laboratorio LIMSI-CNRS en Francia (Lamel et alii 2004) ou o CU-HTK da universidade de Cambridge en Reino Unido (Hain et alii 2005), aínda non conseguen prestacións totalmente satisfactorias nestas tarefas. En transcrición de programas de noticias, o mellor resultado obtido na actualidade está en torno ao 10% de WER ("word error rate"), mentres que para transcrición de conversacións telefónicas a taxa de erro vese incrementada ata o 15-20%. Estes resultados lógranse con cantidades inxentes de material de adestramento e en tempos de execución varias veces por riba de tempo real, o que dificulta a súa aplicación en problemas reais e en idiomas minoritarios.

Para lograr estas prestacións, os problemas que teñen que resolverse (á parte da recollida de datos) teñen que ver co tratamento adecuado da alta variabilidade acústica, de linguaxe e de vocabulario presentes nas interaccións entre humanos. En canto ao modelado acústico, os problemas aparecen cando existe desaxuste entre o material de recoñecemento e o de adestramento. Este desaxuste pode estar producido por variación do ruído ambiental ou pola variabilidade interlocutor. Respecto ao ruído ambiental, os recoñecedores inclúen parametrizacións que incorporan parte das características do oído humano en canto a robustez fronte ao ruído (parametriza-

ción MFCC ou PLP). Outros autores propoñen a utilización de métodos de combinación de parámetros aplicando algoritmos de fusión de datos (Beyerlin 1998, Zolnay et alii 2005). Respecto da variabilidade interlocutor, a normalización da lonxitude do tracto vocal (VTLN) (Lee & Rose 1998) e o adestramento adaptativo de locutor (Giuliani et alii 2004) son os sistemas máis utilizados para evitar a súa influencia.

En canto ao modelado lingüístico, as técnicas máis empregadas para recoñecemento automático de fala (RAF) de grandes vocabularios son os modelos estatísticos tipo N-grama. Existen ferramentas de dominio público que permiten adestrar os devanditos modelos: unha das máis coñecidas é "SRILM - The SRI Language Modeling Toolkit" da firma SRI International. A partir desta ferramenta pódense explorar técnicas de selección de material textual apropiado á tarefa a recoñecer (Diéguez et alii 2005). Estamos falando entón de reducir o desaxuste entre adestramento e recoñecemento mediante a combinación de múltiples modelos de linguaxe cada un deles especializado nun "tema" diferente. Esta vía de investigación é moi interesante pois vai na liña xeral do proxecto de incorporación temperá das fontes de coñecemento en contraposición ao uso inxente de datos (voz e texto).

Algúns exemplos de sistemas comerciais ou de provedores de tecnoloxía son:

— Nuance – Dragon Naturally Speaking
  • Versións *Legal, Medical e Professional*
  • Vocabulario de 25.000 palabras
  • Fala continua e illada
— IBM – Viavoice
  • 100,000 palabras
  • Fala continua e illada
— Loquendo
  • Vocabulario de 1,000,000 de palabras
— O sistema operativo Windows Vista de Microsoft inclúe recoñecemento de fala
— Verbio Technologies S.L., con sede en Barcelona (España), é unha empresa especializada no desenvolvemento de tecnoloxías da fala, basicamente síntese de voz e recoñecemento da fala (gama de produtos VERBIO).

Sen querermos ser exhaustivos e desculpando por anticipado as omisións, algúns laboratorios de fóra do estado con presenza internacional nos foros habituais de intercambios de achegas son:

— LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur), París, Francia
— University of Cambridge, UK
  • Centro que desenvolveu o software HTK (Hidden Markov Toolkit)
— SRI International, USA
— Mississippi State University, USA
— RTWH (Rheinisch-Westfaelische Technische Hochschule), Aachen, Alemaña

En España, desde o ano 1999 existe a "Rede Temática en Tecnoloxías da Fala" que é un foro común onde os investigadores en tecnoloxía da fala poidan axuntar esforzos e compartir experiencias co fin de:

—Fomentar a investigación en tecnoloxías da fala atraendo a novos mozos investigadores a este campo mediante cursos de formación, intercambios de estudantes, bolsas e premios.
—Atraer investimentos para investigación das empresas cara ás tecnoloxías da fala mediante a procura de novas aplicacións que ofrezan novas posibilidades de negocio. Estas aplicacións débense concretar en demostradores que atraian o interese das empresas.
—Avanzar na creación de lazos de colaboración e integración dos membros da Rede para manter o liderado de España na investigación do castelán, e potenciar tamén os idiomas cooficiais como o catalán, éuscaro e galego.

Este último obxectivo é de especial importancia para idiomas minoritarios ou minorizados como é o caso do galego.

Na páxina web da rede http://www.rthabla.es/ pode atoparse unha listaxe dos grupos de investigación dedicados en España ao desenvolvemento da tecnoloxía da fala e do linguaxe natural.

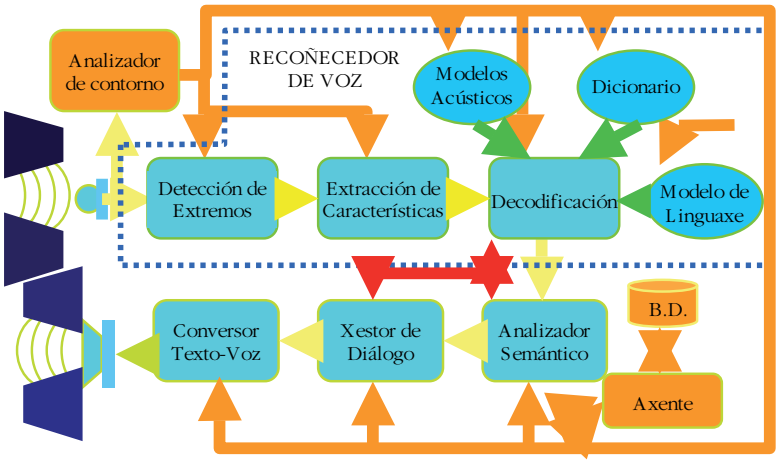
3. O PROBLEMA DO RECOÑECEMENTO DA FALA

Un sistema de recoñecemento de fala, polo xeral, forma parte dun sistema máis complexo denominado sistema de diálogo home-máquina, pois estamos a falar de aplicacións nas que os usuarios queren acadar unha certa acción por parte dun sistema automático. A interacción por voz con tales sistemas fai necesario un sistema que "interprete" a mensaxe falada. No Cadro 2 amósase o diagrama de bloque das compoñentes fundamentais de calquera sistema de diálogo actual. Esta taxonomía está estable dende hai xa uns anos en canto a tecnoloxía básica. A énfase desde o punto de vista investi-

gador estase a poñer na análise do contorno que permite que todos os bloques básicos se adapten ás condicións particulares de operación. A detección temperá de aspectos tales como nivel de ruído, calidade do son, idioma, ou incluso tipo de fala, fan posible o axuste da tecnoloxía dos bloques constituíntes, de tal xeito que a partir dun sistema básico de referencia poden construírse sistemas "a medida" de prestacións aceptables.

Nun sistema de recoñecemento de fala como o que amosamos no Cadro 2 poden distinguirse tres fases diferenciadas: **unha segmentación de audio** en segmentos homoxéneos, **unha extracción de características** ou parametrización, que consiste nunha fase de procesado do sinal mediante a cal a secuencia de voz se converte nunha serie de tramas formadas por parámetros máis ou menos representativos; e unha terceira fase de aliñamento de patróns, na que as tramas son comparadas cun conxunto de unidades predefinidas para extraer a secuencia recoñecida.

Na última etapa tamén denominada **descodificació**n atopamos dous compoñente fundamentais o **modelado acústico** e **modelado da linguaxe**. A ambos aspectos lle dedicaremos apartados específicos.



Cadro 2: *Diagrama de bloques de un sistema de diálogo home-máquina*

Segmentación de audio

Dado que os sistemas automáticos de recoñecemento de fala continua con grandes vocabularios independentes do locutor e da canle aínda non alcanzaron o nivel de prestacións requirido para esta tarefa, é necesario reducir a complexidade do problema en moitos dos factores. Facer que o reco-

ñecedor de fala (RAF) sexa dependente do locutor mediante a adaptación dos modelos acústicos mellorará as prestacións do sistema en varios puntos porcentuais. Tendo en mente a situación anterior, o obxectivo deste bloque é, dada unha gravación, detectar as quendas de locutor e pescudar se algúns deles pertencen a locutores rexistrados no sistema. Esta información sobre a identidade do locutor será comunicada ao sistema encargado de xerar as hipóteses de fala recoñecida.

Extracción de características

Por medio da parametrización preténdese extraer do segmento de voz toda a información que poida resultar significativa para guiar a tarefa posterior de aliñamento.

O proceso realízase do seguinte xeito:

—Realízase a mostraxe do segmento acústico, utilizando habitualmente unha frecuencia de 8 ou 16 KHz.
—Extráense tramas de mostras, tipicamente de 10 a 30 ms de duración, mediante un enventanado previo e permitindo o solapamento con tramas adxacentes de 10 a 15 ms para evitar o problema da non estacionariedade do proceso de voz.
—Aplícase algún tipo de transformación á trama de mostras, normalmente no dominio frecuencial, para convertela na secuencia de parámetros.

As parametrizacións máis usuais utilizan os coñecidos coeficientes de predición lineal (LPC) ou os chamados Mel-Frecuency Cepstral Coeficients (MFCC), que se obteñen mediante a aplicación dun banco de filtros e unha posterior transformación DCT. Normalmente engádese a enerxía da trama e as derivadas primeira e segunda de todos os parámetros, para conformar así un vector composto por varias decenas de elementos.

Descodificación

O proceso de aliñamento de patróns é o que se denomina comunmente *descodificación*. O recoñecedor ou decodificador, recibe tramas xeradas polo parametrizador cunha cadencia constante, que depende da lonxitude da fiestra de análise. A secuencia de observacións debe ser comparada cun inventario de patróns acústicos previamente extraídos, co obxectivo de lograr un aliñamento óptimo en relación con algún tipo de medida predefinida. A natureza variable da fala aconsella representala como unha fonte estocástica

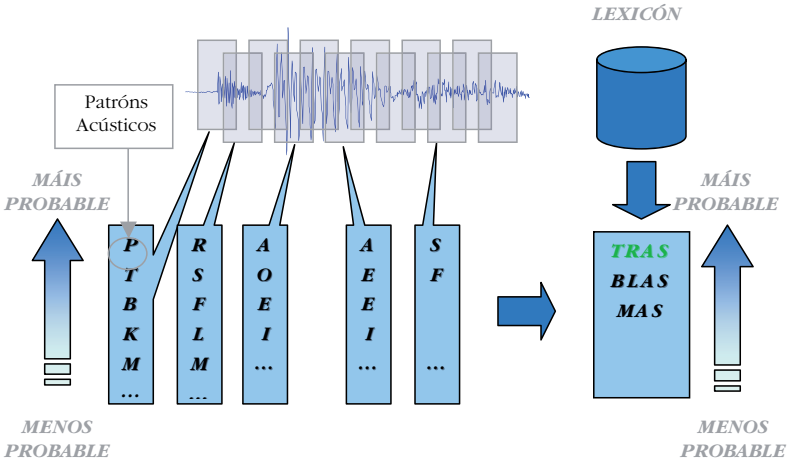
e definir esta medida de forma probabilística. Dado un conxunto de unidades $\boldsymbol{\lambda}_i$ e unha secuencia de observacións, **O**, o recoñecemento de voz exponse hoxe en día como a solución da seguinte ecuación:

$$\boldsymbol{\lambda} = \arg\max_i \{P(\mathbf{O}/\boldsymbol{\lambda}_i)\, P(\boldsymbol{\lambda}_i)\}$$

onde $\boldsymbol{\lambda}$ é a secuencia de patróns descodificada e $\boldsymbol{\lambda}_i$ percorre todas as posibles combinacións de unidades acústicas do inventario. O termo $P(\mathbf{O}/\boldsymbol{\lambda}\boldsymbol{\lambda}_i)$ representa a probabilidade acústica, e mide o axuste entre as observacións e os patróns, e $P(\boldsymbol{\lambda}_i)$ representa a probabilidade desa secuencia definida polo que se denomina "modelo de linguaxe". Os detalles da implementación desta ecuación varían substancialmente entre diferentes sistemas ou aplicacións. Dependendo de que o recoñecemento sexa fonético, de palabras illadas ou de fala continua, o factor $P(\boldsymbol{\lambda}_i)$ ten un ou outro significado ou ata pode ser directamente eliminado. Os patróns que forman o inventario tamén poden ser de tipos moi diversos dependendo da aplicación a que se destinen: fonemas ou parte deles, partes de palabras, palabras completas ou ata frases completas. Con todo a forma de representar estas unidades é a mesma en practicamente todos os recoñecedores, utilizando para iso *modelos ocultos de M*arkov (HMMs). O Cadro 3 ilustra gráficamente como é o proceso de búsca da secuencia de "unidades acústicas" mais probables dada una gravación de voz.

A tecnoloxía de recoñecemento de patróns estatísticos que nos permitiu alcanzar este nivel de desenvolvemento é relativamente antiga. A base da maioría dos recoñecedores actuais son os modelos ocultos de Markov, cuxa aplicación inicial neste campo data de principios dos anos oitenta do século pasado. Os avances observados nos últimos tempos son debidos en gran medida ao continuo crecemento nas prestacións dos computadores, pero a mellora dos diferentes elementos implicados no proceso de recoñecemento tivo tamén unha influencia fundamental.

Neste sentido un área que achegou continuos avances é o modelado acústico, ou sexa a definición e adestramento dos modelos, $\boldsymbol{\lambda}_i$. A utilización de modelos con contexto, as melloras nas técnicas de adestramento, a aparición de novos métodos de parametrización e de técnicas de normalización e adaptación, por pór algúns exemplos, conduciron a un axuste maior entre os modelos e a secuencia de voz e xa que logo a taxas de recoñecemento maiores.



Cadro 3: *Ilustración de como mediante a concatenación de modelos acústicos pódese abordar o recoñecemento dun chorro continuo de fala. As restricións na concatenacións veñen marcadas polo lexicón e adicionalmente polo modelo de linguaxe*

Doutra banda a imposición de restricións gramaticais no recoñecemento por medio de modelos de linguaxe, foi fundamental para permitir a evolución desde os recoñecedores de palabras conectadas ou gramáticas limitadas ata os sistemas orientados a fala continua con prestacións aceptables. Tamén neste área a investigación foi, e segue sendo importante, sobre todo en aspectos relacionados co seu adestramento para a obtención de $P(\boldsymbol{\lambda}_i)$, e máis recentemente coa súa adaptación automática a contornas cambiantes.

Pasamos a continuación a describir as bases algorítmicas dos modelos ocultos de Markov como ferramenta estatística que permite abordar o problema do modelado da variabilidade inter e intra locutor.

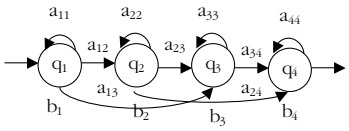
Modelado acústico: Modelos Ocultos de Markov

Unha vez segmentada a secuencia de voz e extraídos os parámetros significativos, a trama resultante debe ser comparada cunha serie de patróns de referencia. As maiores dificultades no recoñecemento de voz aparecen nesta fase e son debidas á enorme variabilidade da fala humana. Esta variabilidade ten como consecuencia que non se poida realizar esta comparación de xeito determinista.

Para ilustrar este problema con máis detalle, pode exporse o caso das vogais. Supoñendo que sabemos que un determinado segmento de voz é vocálico, a identificación da vogal pronunciada podería facerse atendendo

unicamente aos formantes. Os formantes son parámetros frecuenciais que indican a posición das resonancias fundamentais de cada vogal, producidas á súa vez polo punto de articulación. Atendendo ás dúas primeiras resonancias (o primeiro e segundo formante), constrúese o coñecido triángulo vocálico. A partir do triángulo vocálico poderían fixarse uns limiares absolutos, permitindo unha decisión inmediata. Na práctica este sistema funciona ben, pero non é infalible. A variabilidade na pronuncia (por exemplo unha pronuncia máis ou menos pechada das vogais semiabertas), podería producir erros non recuperables. Aínda que no caso das vogais o problema pode ser pequeno, podemos imaxinarnos a complexidade e dificultade que implicaría definir regras de decisión estritas para fonemas con moita maior variabilidade, como oclusivos ou fricativos.

A solución pasa por unha comparación de natureza probabilística. A idea é que o sistema devolva, non unha decisión absoluta, (este fonema é unha /e/), senón a probabilidade de que a trama de parámetros corresponda a un dos fonemas modelados. Trátase xa que logo de intentar obter un modelo probabilístico do proceso de fonación humana, o que resulta unha tarefa bastante ardua.



Cadro 4: Exemplo de modelo de Markov (HMM)

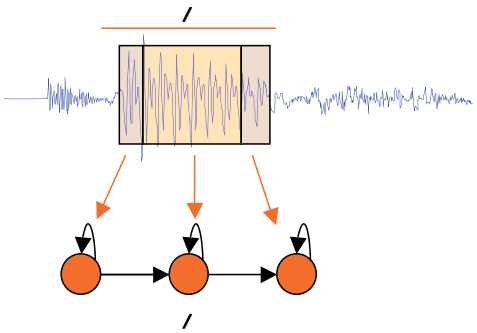
A solución máis estendida hoxe en día é o emprego de **modelos ocultos de Markov** (HMMs, "Hidden Markov Models" en inglés). Os modelos ocultos de Markov son modelos estatísticos que permiten representar de forma adecuada certos procesos estocásticos denominados *markovianos*. Os procesos markovianos teñen a característica de que a súa función de distribución depende do seu estado actual, pero non de estados anteriores, é dicir, son procesos sen memoria.

Un HMM componse de tres elementos (Cadro 4): un conxunto de estados interconectados entre si ($\{q_i\}$), unhas probabilidades de transición entre estados $\{a_{ij}\}$ e un conxunto de distribucións de probabilidades $\{b_i\}$; cada unha delas asóciase tamén a un estado. Un exemplo clásico de experimento markoviano é o seguinte (Rabiner 1989): existen un número determinado de urnas, cada unha con diversas bólas de diferentes cores; extráese unha bóla dunha urna elixida ao azar e enúnciase a súa cor; seguidamente elíxese dunha forma aleatoria (por exemplo lanzando un dado) unha nova urna para extraer a seguinte bóla; o observador coñece a secuencia de bólas, pero des-

coñece de que urna se extraeu cada unha (esta sería a parte oculta do modelo). Este experimento pode modelarse mediante un HMM, no que cada estado se correspondería a unha urna, as probabilidades de transición entre estados virían dadas polos dados empregados para decidir a nova urna e as funcións de distribución quedarían definidas polo número de bólas de cada cor en cada urna.

Se o observador do noso experimento é suficientemente curioso, pode formular as tres preguntas clásicas:

1. Se coñezo unha secuencia de observacións (bólas de cores) $\boldsymbol{O}$, suficientemente longa e sei cal é o número de urnas, ¿podo dalgún xeito deducir o valor dos parámetros $\{a_{ij}\}$ e $\{b_i\}$?
2. Se coñezo todos os parámetros do modelo ($\{q_i\}$, $\{a_{ij}\}$ e $\{b_i\}$), e unha secuencia determinada de observacións $\boldsymbol{O}$, ¿podo deducir de que urna foi extraída cada bóla?
3. Dada unha secuencia calquera de observacións, ¿cal é a probabilidade de que sexa xerada polo modelo?



Cadro 5: Exemplo de HMM aplicado a recoñecemento de fala.

A teoría de modelos de Markov dá resposta a estas tres preguntas posibilitando a súa aplicación en diferentes problemas prácticos, como é o caso do recoñecemento de voz (Picone 1990). Un exemplo típico da súa aplicación neste campo amósase no Cadro 5. Neste exemplo, cada fonema modelaríase mediante un HMMs diferente (represéntase só o fonema /o/, formado por tres estados interconectados, pero existiría outro modelo para cada fonema no inventario). O primeiro estado modelaría (como veremos, dunha forma bastante difusa), os parámetros correspondentes ao arranque do fonema, o segundo á parte estacionaria deste e o último á transición ao fonema seguinte. O número de estados e as interconexións entre estes, elíxense de xeito

heurístico, aínda que o exemplo que propomos é o máis usado para modelar fonemas ou unidades subfonéticas. Obsérvese que as funcións de distribución, modelan agora a distribución das tramas de parámetros para unha parte dun fonema determinado.

Regresando ao experimento das urnas, a equivalencia co proceso de fala sería así:

—As tramas de parámetros correspóndense ás bólas de cores.
—As urnas, a parte oculta do modelo, correspóndense á evolución do tracto vocal. Cada urna representaría no noso caso, polo menos idealmente, unha determinada configuración do tracto vocal, que á súa vez produce uns parámetros frecuenciales determinados e máis ou menos homoxéneos (por exemplo a parte estacionaria dun /o/).
—As probabilidades de transición non son agora arbitrarias, senón que dependen da lonxitude da parte estacionaria e dos arranques e finalizacións de cada fonema.

O deseño adecuado das funcións de distribución é unha parte clave do problema. Deben de ser capaces de modelar de xeito adecuado a variabilidade interlocutor (diferentes pronuncias do mesmo fonema ou variacións de prosodia) ou intralocutor (variacións entre os tractos vocálicos dos individuos, variantes dialectais, etc.) dos parámetros para unha parte dun fonema determinado.

Existen varias formas de crear estas funcións, aínda que unha das máis comúns é empregar modelos de mesturas de gausianas, nos que a función de distribución de cada estado vén dada por unha suma ponderada de gausianas:

$$b_j(\mathbf{O}) = \sum_{m=0}^{M-1} c_{jm} N(\mathbf{O}, \mu_{jm}, \sigma_{jm})$$

Onde M é o número de mesturas, **O** é a trama de observacións e μ_{jm} e σ_{jm} son os vectores de medias e varianzas para a mestura *m* da función de distribución do estado *j*.

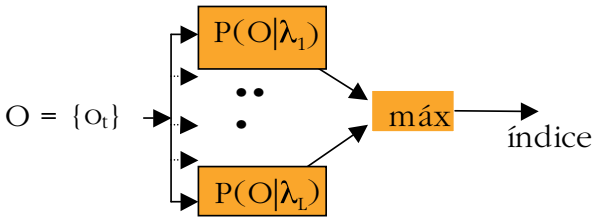
O gran problema dos HMMs é obter as probabilidades de transición e as funcións de distribución de cada estado. Para iso é necesario partir dunha base de datos de adestramento. Este é o proceso máis caro e máis custoso á hora de implementar un sistema de recoñecemento.

As bases de datos de adestramento están formadas por horas de gravacións de falantes coidadosamente elixidos para representaren fielmente a poboación á que vai destinada a aplicación de recoñecemento. Neste sentido, búscase que a base estea balanceada en canto ao sexo dos locutores, ao

seu rango de idades e á súa localización xeográfica. As palabras que pronuncian os falantes tampouco se deixan ao azar, senón que se buscan frases e palabras foneticamente ricas, para cubrir todo o inventario fonético o mellor posible.

As bases de datos de adestramento están ademais etiquetadas. Todas as gravacións teñen que ser transcritas polo menos a nivel de palabra, (se é posible tamén foneticamente), pero en todo caso marcando todos os ruídos, dúbidas, tartamudeos e en xeral calquera tipo de fenómeno paralingüístico que puidese afectar ao proceso de adestramento. Como se pode comprender facilmente, este proceso é enormemente custoso e encarece a construción das bases de datos de recoñecemento. Afortunadamente, existen organismos dedicados á distribución de bases de datos comerciais (ELDA, LDC) onde se poden adquirir bases de datos para un bo número de idiomas.

Se se dispón dunha base de datos pode realizarse xa o adestramento dos modelos. O proceso consiste en parametrizar todas as gravacións dispoñibles e separar as tramas de parámetros correspondentes ao modelo λ_i que se pretende adestrar (por exemplo, considerar todas as realizacións do fonema /e/). Supondo que xa se decidiu o número de estados e a interconexión entre estes para o modelo correspondente, deben obterse agora os parámetros e $\{a_{ij}\}$ e $\{b_i\}$. Os algoritmos que se empregan para isto denomínanse de máxima verosimilitude e intentan maximizar a probabilidade a posteriori de que a secuencia de parámetros sexa xerada polo modelo ($P(\lambda_i | O)$). O algoritmo de Baum-Welch (Welch 2003) é o máis empregado para esta fase de adestramento.



Cadro 6: Ilustración do proceso de recoñecemento.

Unha vez adestrados os modelos, pode pasarse á fase de recoñecemento propiamente dita. Nesta fase as tramas de parámetros son comparadas cos modelos que se adestraron.

Tipicamente disporase de N modelos, $\lambda i,\ i = 1 \ldots N$ de entre os que se elixirá aquel que maximice a probabilidade a posteriori $P(\mathbf{O}/\boldsymbol{\lambda}\boldsymbol{\lambda}_i)$. Este proceso denomínase habitualmente descodificación e lévase a cabo mediante o algoritmo clásico de Viterbi.

Na práctica, o problema é máis complicado. Se os modelos representan fonemas, a aplicación do algoritmo de Viterbi daría como resultado unha secuencia de fonemas da que, debido aos erros de transcrición, resultaría imposible deducir a palabra pronunciada polo falante.

Existen varias solucións a este problema. Se por exemplo desexamos realizar un recoñecedor de díxitos, é posible obter un modelo para cada un deles, de maneira que a aplicación de Viterbi nos daría un resultado correcto (obviando algúns problemas relacionados coa segmentación da secuencia de voz e coa posible aparición de ruídos). Con todo, isto require dispor dunha base de datos de adestramento que teña suficientes realizacións de cada un dos díxitos.

Esta solución non é viable para a maioría das aplicacións, nas que o vocabulario non é predicible. Máis habitual é adestrar modelos fonéticos e crear un vocabulario de recoñecemento a base de concatenar modelos de fonemas entre si. De feito, non adoitan utilizarse fonemas simples (/a/,/e/,etc.), xa que estes non teñen en conta a coarticulación entre fonemas veciños. No seu lugar empréganse fonemas con contexto, é dicir, modélase cada fonema tendo en conta o fonema dereito ou esquerdo dentro da palabra. Os modelos de fonemas con información contextual a ambos os dous lados denomínanse *trifonemas*. Por exemplo: **a-b+r** é o trifonema que modela o fonema /b/ cando vai precedido por unha /a/ e seguido por unha /r/. Os trifonemas ofrecen mellores prestacións, pero teñen a desvantaxe de requirir gran cantidade de material de adestramento (son moitos máis modelos a adestrar) e de necesitar maior potencia de cálculo na fase de recoñecemento.

Modelado da linguaxe

Como vimos, o modelado acústico mediante modelos de Markov é a base do recoñecemento de voz, permitindo a identificación dos fonemas pronunciados polo falante. O modelado acústico en por si é suficiente para aplicacións de palabras illadas ou para conxuntos limitados e pouco extensos de frases predefinidas. Entre outros exemplos deste tipo de escenarios, podemos expor as aplicacións de automatización de centrais telefónicas, nas que o usuario pronuncia o nome e apelido da persoa con quen quere falar; ou os sistemas automáticos de acceso a bases de datos, nos que o tipo e formato das preguntas posibles está limitado e é coñecido.

En aplicacións máis complexas é imposible limitar o recoñecemento a un conxunto de frases ou de palabras illadas. Nese caso a única solución factible é definir un vocabulario de recoñecemento, tipicamente adaptado á aplicación, e permitir a aparición de calquera combinación de palabras dese vocabulario. Neste caso, o problema principal é que o recoñecedor debe

decidir, baseándose unicamente nas probabilidades acústicas, cal é a mellor segmentación da secuencia de voz en palabras do vocabulario. Con todo, a fala presenta diversos fenómenos que complican esta tarefa ata tal punto que producen que o emprego unicamente de probabilidades acústicas nestes escenarios fracase de forma estrepitosa. Entre eses fenómenos podemos citar os seguintes:

—Variantes dialectais, sociolectais, contextuais e idiolectais.
—Coarticulación entre palabras.
—Palabras homófonas.

Todos estes fenómenos afectan tamén aos recoñecedores de palabras illadas, pero debido a que o sistema espera unha soa palabra ou un conxunto limitado e coñecido, o efecto é moito menos severo.

Por exemplo, supoñamos un recoñecedor de díxitos que contén a palabra "*dez*", un falante con seseo pronunciará "*des*". Con todo, o desaxuste acústico non será suficiente para que a palabra sexa confundida con "*un*", "*dous*", "*tres*", etc., co que é probable que non se produza ningún erro. Con todo, nun escenario de fala continua, o recoñecedor poderá confundir esa palabra co comezo de "*descoñecido*", "*desanimado*" ou de calquera outro vocábulo similar e dependendo da palabra pronunciada a continuación (por exemplo "*coñecido*" ou "*animado*") producir con certa probabilidade unha transcrición errónea.

Algúns destes efectos poden ser paliados dotando ao sistema da posibilidade de empregar pronuncias alternativas para cada palabra do dicionario, incluíndo así as variantes dialectais e sociolectais máis xeneralizadas. Aínda que esta solución pode ser algo efectiva en escenarios sinxelos, en xeral non evita nin os problemas relacionados coa coarticulación, nin as confusións debidas a palabras homófonas. O efecto final de todos estes problemas podemos ilustralo mediante o seguinte exemplo, no que se mostra a transcrición orixinal e a obtida mediante un recoñecedor de voz cun vocabulario de 20.000 palabras:

Orixinal: "*a venda alcanza nestes momentos cotas similares"*
Transcrición: "*a venda alcanzan es tes mo home en tras co ata cine dar es*"

Neste caso, o falante pronuncia con certo grao de ceceo, o que explica en parte a confusión entre "*similares*" e "*cine dar es*". En canto ás transcricións "*alcanzan es tes*" e "*alcanza nestes*" pode verse que son moi similares dende o punto de vista fonético, de maneira que para o recoñecedor resultan case indistinguibles.

Para evitar este tipo de problemas, necesítase un mecanismo que sexa capaz de incluír restricións de tipo sintáctico (e se é posible, tamén semántico), no proceso de recoñecemento. Isto faise empregando os denominados **modelos de linguaxe** (Manning & Schütze 1999). Un modelo de linguaxe eficiente debería, dalgún xeito, ser capaz de indicar que a secuencia de palabras *"a venda alcanzan"* é incorrecta e "*a venda alcanza*" correcta, potenciando a segunda fronte á primeira.

O problema de obter un modelo de linguaxe é, no entanto, bastante complicado e foi motivo de investigacións continuadas durante as últimas décadas. Obter un modelo universal que sexa quen de representar toda a riqueza e variabilidade dunha lingua parece aínda utópico.

O método que intuitivamente parece máis natural e sinxelo para obter un modelo de lingua é o emprego de categorías gramaticais. Neste tipo de modelos, a cada palabra asígnaselle a súa categoría (por exemplo: *os*, artigo masculino plural) e mediante unha serie de regras defínense as secuencias de categorías permitidas pola lingua. Estes sistemas presentan varios problemas prácticos. Por unha banda a riqueza da lingua é tal que non é posible concretar toda a súa variabilidade nun conxunto limitado de regras definidas de forma empírica. Por outro, existe un problema grave con palabras ambiguas, que poden ter unha ou outra función en diferentes posicións da frase. Por estas razóns, as regras defínense habitualmente de xeito estatístico, é dicir, pártese dun corpus de texto de adestramento no que cada palabra se categoriza manualmente e extráense regras estatísticas, definidas mediante a probabilidade de aparición dunha secuencia determinada de categorías. A necesidade de realizar manualmente a categorización inicial do corpus dificulta a construción destes modelos e limita a súa aplicación. Ademais, os modelos baseados en categorías non son capaces de xeneralizar regras semánticas, de xeito que, para un modelo deste tipo, *"a casa corre polo campo"* é unha frase tan correcta como "*o can corre polo campo*". Para rematar, a aplicación destes modelos implica unha categorización inicial da palabra, o que dificulta a ampliación do vocabulario fóra daquel definido polo corpus de adestramento.

Todas estas razóns provocan que os modelos de categorías non se utilicen de xeito xeneralizado nos sistemas de recoñecemento. No seu lugar empréganse modelos estatísticos baseados en palabras. Os modelos baseados en palabras son puramente estatísticos, no sentido de que non identifican a categoría ou función dunha palabra, senón que se limitan a estudar a súa frecuencia de aparición no texto en relación coas palabras veciñas. Para estes modelos cada variante dun vocábulo (masculino-feminino, plural-singular, conxugacións de verbos, etc.) é un elemento novo que debe considerarse por separado, polo que xeneralizan peor que os modelos baseados en

clases. Pero a súa gran vantaxe é que son sinxelos de xerar e adáptanse ben ao texto de adestramento.

Os modelos de linguaxe estatísticos baseados en palabras constrúense mediante o contado de palabras. Sinxelamente obtense o número de veces que unha secuencia de palabras aparece no texto de adestramento, asignándoselle unha probabilidade $P(w_1\, w_2 \ldots w_N)$.

Idealmente, o modelo de linguaxe debería de cuantificar a relación estatística entre dúas palabras calquera do texto de adestramento, independentemente da súa separación. Isto obrigaría a obter a probabilidade de cada palabra do vocabulario, cada combinación de dúas, tres e en xeral de N palabras, que na nomenclatura dos modelos de linguaxe denomínanse bigramas, trigramas e n-gramas respectivamente. A probabilidade de que o 5-grama (ou pentagrama segundo os autores) *"o can corre polo"* vaia seguida da palabra "*campo*" viría dada pola seguinte expresión:

$$P(campo \mid o\ can\ corre\ polo)$$

Que é a probabilidade condicionada de que apareza a palabra "*campo"*, sabendo que se pronunciaron as palabras "*o can corre polo"*. Esta probabilidade calcúlase contando o número de veces que a secuencia *"o can corre polo campo"* aparece no texto de adestramento e dividindo polo número de veces que aparece "*o can corre polo*". En xeral:

$$P(w_N \mid w_1 w_2 \cdots w_{N-1}) = \frac{C(w_1 w_2 \cdots w_N)}{C(w_1 w_2 \cdots w_{N-1})}$$

Onde $C(w_1\, w_2\, w_3)$ é o número de ocorrencias do trigrama $[w_1\, w_2\, w_3]$ no texto de adestramento. Para obter a probabilidade dun n-grama determinado, partindo dun punto de repouso (unha pausa, ou desde o principio do discurso), aplicaríase a seguinte expresión:

$$P(w_1\, w_2 \ldots\, w_N) = P(w_1)\ \mathrm{P}(w_2 \mid w_1)\ \mathrm{P}(w_3 \mid w_1 w_2)\ \mathrm{P}(w_4 \mid w_1 w_2 w_3) \ldots \mathrm{P}(w_N \mid w_1 w_2 \ldots w_{N-1})$$

Do explicado anteriormente é fácil entender que a probabilidade condicionada é difícil de obter para un N relativamente grande. Por exemplo, para un vocabulario de 20.000 palabras, existen 4×10^8 combinacións posibles de dúas palabras, 8×10^{12}de tres palabras e $1{,}6 \times 10^{17}$ combinacións de catro palabras. Isto implica que nun texto de adestramento de varios millóns de palabras, non aparecerán máis que unha fracción minúscula de todas as posibles combinacións de catro palabras e, ademais, a maioría das combinacións que aparezan repetiranse apenas unha ou dúas veces. Estas cantidades non son o suficientemente significativas para estimar unha boa probabilidade.

Por esta razón, os modelos de linguaxe adoitan basearse en n-gramas de orde tres ou catro, e só se empregan n-gramas de orde superior en casos nos que o texto de adestramento é moi grande (miles de millóns de palabras). Empregar un modelo de linguaxe baseado en trigramas, por pór un exemplo, supón aceptar a seguinte hipótese:

$$P(w_N | w_1 w_2 \ldots w_{N-1}) = P(w_N | w_{N-2} w_{N-1})$$

É dicir, equivale a aceptar que a probabilidade da palabra actual só depende das dous anteriores, o que na maioría dos casos resulta obviamente falso.

O problema principal que se presenta nos modelos de linguaxe estatística é cómo tratar o caso dos n-gramas non vistos. Se un n-grama non apareceu no texto de adestramento, o modelo asígnalle unha probabilidade cero. Isto implica que na fase de recoñecemento este n-grama será ilegal. Nun sistema de recoñecemento de fala continua deben permitirse todas as combinacións posibles do vocabulario, só que asignando unha probabilidade pequena a aquelas pouco probables. Existen diversos métodos de estimación dos n-gramas non vistos, que de forma moi xenérica se basean en reservar parte da masa probabilística a aqueles n-gramas que non aparecen no texto de adestramento e que xenericamente se denominan métodos de descontos. Un algoritmo sinxelo que nos permite ilustrar este tipo de técnicas é supor que cada n-grama (incluídos os non vistos) aparecen unha vez máis do que aparecen en realidade. Supoñamos que o texto de adestramento está composto das seguintes frases: "*luís le libros*", "*luís le periódicos*", "*manuel le periódicos*". Posto que *C(manuel le libros) = 0*, a probabilidade asignada polo modelo a esa secuencia será cero. Se empregamos o método descrito, as contas quedan modificadas de maneira que: *C(manuel le) = 5* e *C(manuel le libros)=1*, co que *P(libros | manuel le) = 1/5*. En xeral a probabilidade para un n-grama calquera de orden N quedará como segue:

$$P(w_N \mid w_1 w_2 \cdots w_{N-1}) = \frac{C(w_1 w_2 \cdots w_N) + 1}{C(w_1 w_2 \cdots w_{N-1}) + K}$$

Onde K é o tamaño do vocabulario.

Un dos problemas deste tipo de métodos de descontos é que supón a mesma probabilidade para todos os n-gramas non vistos. O método asignaría por exemplo, a mesma probabilidade a *P(patacas | manuel le)*, *P(libros | manuel le)* ou *P(prospectos | manuel* le), supondo que os tres trigramas non aparezan no texto de adestramento.

Parece intuitivo pensar que a asignación das probabilidades dos n-gramas non vistos debe facerse tendo en conta as probabilidades dos n-gramas de orde inferior. Para o exemplo anterior é de supor que *P(prospectos)* < *P(libros)* e que *P(le patacas)* < *P(le prospectos)* < *P(le libros)*, polo que a probabilidade *P(libros | manuel le)* debería ser maior que as outras dúas. Esta idea básica leva ao desenvolvemento dos denominados métodos de suavizado e suavizado por interpolación dos n-gramas non vistos, entre os que podemos citar os métodos se Witten-Bell, Good-Touring e Kesser-Ney que son os utilizados de forma xeneralizada hoxe en día en recoñecemento de voz.

Na práctica, os métodos de descontos aplícanse mediante os chamados factores de *back-off,* que se ocupan de relacionar as probabilidades dos n-gramas non vistos cos dos de orde inferior. Para o caso de trigramas a probabilidade condicionada dun trigrama impleméntase do seguinte xeito:

$$\hat{P}(w_3 \mid w_1 w_2) = \begin{cases} P(w_3 \mid w_1 w_2) & si\ C(w_1 w_2 w_3) \neq 0 \\ \alpha(w_1 w_2)\hat{P}(w_3 \mid w_2) & si\ C(w_1 w_2 w_3) = 0\ y\ C(w_1 w_2) \neq 0 \\ \hat{P}(w_3 \mid w_2) & si\ C(w_1 w_2 w_3) = 0\ y\ C(w_1 w_2) = 0 \end{cases}$$

Onde $\alpha(w_1 w_2)$ é o factor de back-off asociado ao bigrama $[w_1 w_2]$

4. Exemplo de proxectos de investigación

A transcrición de noticiarios televisivos constitúe un marco de traballo idóneo para medir as prestacións dun recoñecedor de voz. A gran diversidade de locutores, estilos de fala e temas tratados ao longo dun programa de noticias supón unha esixente proba para un recoñecedor, obrigándoo a ser capaz de funcionar de xeito robusto para un abano de situacións diferentes. É por iso que esta tarefa centrou unha boa parte da investigación en recoñecemento de voz na última década.

O sistema Transcrigal da Universidade de Vigo (García Mateo et alii 2004) foi deseñado para a transcrición de noticiarios en lingua galega, que se caracterizan pola presenza frecuente de persoas que empregan o idioma castelán. O bilingüismo inherente a esta tarefa constitúe unha nova variable a tratar, que aumenta a complexidade e o interese do sistema. Para abordar esta variabilidade, é conveniente recorrer a esquemas adaptados, tanto no relativo aos modelos acústicos, como no modelo de linguaxe. O proxecto inclúe tanto a propia máquina de recoñecemento para grandes vocabularios, como o desenvolvemento dun módulo de segmentación de locutor, a adaptación de modelos de linguaxe e acústicos en contorna multipase e a extracción de tema.

O desenvolvemento abordouse desde o punto de vista de sistema bilingüe que debe ter a capacidade de adaptarse a distintos tipos de linguaxe: fala planeada, espontánea, etc. Co emprego destas técnicas conséguese unha taxa de erro de palabra media do 25% aproximadamente sobre TranscrigalDB (Diéguez 2005). A título de comparación, podemos comentar que o sistema de LIMSI consegue un 20% de WER, con máis ou menos a mesma cantidade de material (voz e texto) de adestramento para castelán de México/USA.

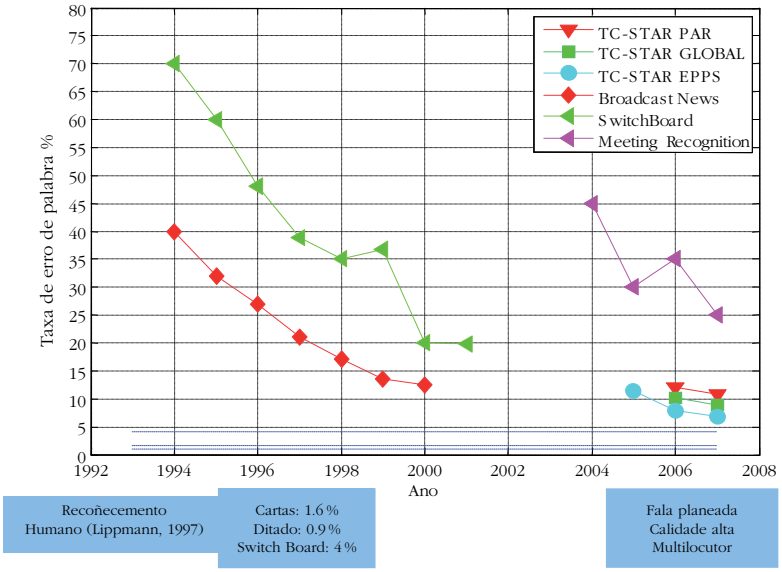
A tecnoloxía desenvolvida permite a adaptación a outras tarefas noutros idiomas. Neste sentido, participamos na campaña de avaliación do proxecto TC-STAR no ano 2006, orientada ao recoñecemento automática de fala parlamentaria gravada do Parlamento Europeo e das Cortes Españolas. Os detalles da nosa implementación poden atoparse en (Docío et alii 2006), se ben amosamos na Táboa 1 un resumo das prestacións acadadas. A principal conclusión que se extrae dos resultados conseguidos é a necesidade de adaptación tanto dos modelos acústicos como de linguaxe á tarefa e ao locutor en cuestión.

A título de exemplo e con datos recollidos de distintas fontes, amósase no Cadro 7 un histórico da evolución das prestacións de distintos sistemas de recoñecemento. Pode observarse como a taxa de recoñecemento varía de forma importante segundo as condicións da tarefa. No caso de *meeting recognition*, a tarefa consiste no recoñecemento multilocutor en reunións de traballo, polo que a fala é de tipo espontáneo: aquí obtivéronse taxas relativamente baixas. No proxecto TC-STAR inténtase o recoñecemento de fala parlamentaria, empregando dous escenarios diferentes: as Cortes Españolas (PAR) e as sesións plenarias do Parlamento Europeo (EPPS). En ambos os casos a fala é planeada polo que os resultados son moito mellores que en *meeting recognition*. Con todo, as sesións do Parlamento Europeo presentan un fala moito máis planeada, que o Parlamento Español, posto que non hai réplicas, nin discusións ou interrupcións; de aí, que o recoñecemento sexa bastante mellor. Como comparación, as liñas continuas do Cadro 7 amosan os resultados obtidos por Lippman en 1997 (Lippman 1997), quen analizou as taxas de recoñecemento de voz alcanzadas por humanos obtendo un 4% de taxa de erro de palabras para a tarefa SwitchBoard, e un 1,6 e 0.9% para textos de cartas e ditado en xeral respectivamente.

		MODELOS		
		Universais	Homes	Borrell
Test	Borrell	81.19%	82.74%	89.38%
	Todo	80.56%	80.39%	76.59%

Táboa 1: *Exemplo de prestacións (tasa de recoñecemento) do sistema da UVIGO na tarefa de transcrición do Parlamento europeo (proxecto TC-STAR ano 2006)*

A conclusión que se extrae é que estamos cerca de acadar os resultados que acadan os humanos. A pregunta segue sendo, ¿é porque non se usan? A resposta é complicada: digamos que a "transcrición automática" alcanzou a calidade máxima, pero a "comprensión automática", que permite corrixir erros, non acadou aínda os niveis requiridos en moitas aplicacións prácticas.



Cadro 7: *Resumo do histórico da evolución das prestacións dos sistemas de recoñecemento de fala. As liñas rectas amosan os resultados obtidos por Lippman en 1997 (Lippman 1997) para dúas tarefas diferentes, quen analizou as taxas de recoñecemento de voz alcanzadas por humanos*

5. Discusión e liñas futuras

Aínda falta tempo para que os ordenadores nos entendan. Quizais o problema maior sexan os sistemas de diálogo, non os de recoñecemento de fala, aínda que estes últimos necesitan aínda traballo. Algúns posibles camiños poden ser:

- Inclusión de modelos gramaticais
- Recoñecemento multimodal

Unha das características fundamentais das tecnoloxías da fala é a enorme dependencia que presentan co idioma de que se trate. En case calquera sistema adóitase facer unha división entre os bloques de procesado de sinal e os bloques de procesado de linguaxe. Para o desenvolvemento destes últimos é necesario dispor de coñecementos lingüísticos do idioma en cuestión. Tamén é indispensable dispor dunha serie de recursos orais e de texto no(os) idioma(s) de que se trate. Estas bases de datos presentan unhas características particulares que fai que en moitos casos non sexa posible reusar material xa existente e por iso é necesario embarcarse en procedementos de deseño, captura e etiquetaxe moi laboriosos e custosos. A non dispoñibilidade destes recursos é unha das causas do baixo nivel de desenvolvemento da tecnoloxía de fala en idiomas minoritarios como o galego. Só co concurso da axuda institucional é posible hoxe en día abordar este tipo de proxectos.

Referencias bibliográficas

Beyerlein, P. (1998): "Discriminative model combination", in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol 1, 481-484. http://ieeexplore.ieee.org/iel4/5518/14820/00674472.pdf?isnumber=14820&prod=CNF&arnumber=674472&arSt=481&ared=484+vol.1&arAuthor=Beyerlein%2C+P.

Docío Fernández, L.; A. Cardenal López & C. García Mateo (2006): "TC-STAR 2006 automatic speech recognition evaluation: The UVIGO System", in *TC-STAR Workshop on Speech-to-Speech Translation*. http://www.elda.org/tcstar-workshop_2006/pdfs/asr/tcstar06_docio-fernandez.pdf.

Diéguez Tirado, J.; C. García Mateo; L. Docío Fernández & A. Cardenal López (2005): "Adaptation strategies for the acoustic and language models in bilingual speech transcription", in *Proceedings of ICASSP' 05*, vol. 1, 833-836. http://ieeexplore.ieee.org/iel5/9711/30650/01415243.pdf?isnumber=30650&prod=CNF&arnumber=1415243&arSt=+833&ared=+836&arAuthor=+Dieguez-Tirado%2C+J.%3B++Garcia-Mateo%2C+C.%3B++Docio-Fernandez%2C+L.%3B++Cardenal-Lopez%2C+A.

ELDA: *Evaluations and Language esources Distribution Agency*. http://www.elda.org/

Fiscus, J. G. (1997): "A post-procesing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)", in *Proceedings of 1997 IEEE workshop on Automatic Speech Recognition and Understanding*, 347-354. http://ieeexplore.ieee.org/iel4/5256/14244/00659110.pdf?

isnumber=14244&prod=CNF&arnumber=659110&arSt=347&ared=354&arAuthor=Fiscus%2C+J.G.

García Mateo, C.; J. Diéguez Tirado; L. Docío Fernández & A. Cardenal (2004): "Transcrigal: a bilingual system for automatic indexing of broadcast news", in *Proceedings of IV International Conference on Language Resources and Evaluation*. Lisboa.

Giuliani, D.; M. Gerosa & F. Brugnara (2004): "Speaker normalization through constrained MLLR based transforms", in *Proceedings of the International Conference on Spoken Language Processing,* vol. 4, 2893-2897.

Hain, T.; P. C. Woodland; G. Evermann; M. Gales; X. Liu; G. Moore; D. Povey & L. Wang (2005): "Automatic transcription of conversational telephone speech", in *IEEE Transactions on Speech and Audio Processing* 13-6, 1173-1185. http://ieeexplore.ieee.org/iel5/89/32511/01518917.pdf?isnumber=32511&prod=JNL&arnumber=1518917&arSt=+1173&ared=+1185&arAuthor=Hain%2C+T.%3B+Woodland%2C+P.C.%3B+Evermann%2C+G.%3B+Gales%2C+M.J.F.%3B+Xunying+Liu%3B+Moore%2C+G.L.%3B+Povey%2C+D.%3B+Lan+Wang.

Lamel, L.; J. L. Gauvain; G. Adda; M. Adda-Decker; L. Canseco; L. Chen; O. Galibert; A. Messaoudi & H. Schwenk (2004): "Speech transcription in multiple languages", in *Proceedings of ICASSP '04*, vol. 3, 757-760. http://ieeexplore.ieee.org/iel5/9248/29345/01326655.pdf?isnumber=29345&prod=CNF&arnumber=1326655&arSt=+iii&ared=+757-60+vol.3&arAuthor=Lamel%2C+L.%3B+Gauvain%2C+J.L.%3B+Adda%2C+G.%3B+Adda-Decker%2C+M.%3B+Canseco%2C+L.%3B+Chen%2C+L.%3B+Galibert%2C+O.%3B+Messaoudi%2C+A.%3B+Schwenk%2C+H.

LDC: *Linguistic Data Consortium*. http://www.ldc.upenn.edu/

Lee L. & R. C. Rose (1998): "A frequency warping approach to speaker normalization", *IEEE Transactions on Speech and Audio Processing* 6-1, 49-60. http://ieeexplore.ieee.org/iel4/89/14168/00650310.pdf?isnumber=14168&prod=JNL&arnumber=650310&arSt=49&ared=60&arAuthor=Lee%2C+L.%3B+Rose%2C+R.

Lippman, R. P. (1997): "Speech recognition by machines and humans", *Speech Communication* 22/1, 1-15.

Manning, C. D. & H. Schütze (1999): *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Nguyen, L.; S. Matsoukas; J. Davenport; F. Kubala; R. Schwartz & J. Makhoul (2002): "Progress in transcription of broadcast news using Byblos", *Speech Communication* 38, 1-2, 213-230.

Picone, J (1990): "Continuous speech recognition using Hidden Markov Models", *IEEE ASSP Magazine*, 26-41. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00054527.

Rabiner, L. R. (1989): "A tutorial on Hidden Markov Models and selected applications in speech recognition", in *Proceedings of. the IEEE*, 77/2, 257-286. http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf.

Rabiner, L. R & B. H. Juang (1986): "An introduction to hidden Markov models", *IEEE ASSP Magazine*, 3/1, 4-16.

Shriberg, E. (2005): "Spontaneous speech: how people really talk and why engineers should care", *Interspeech 2005*, 1781-1784. http://www.isca-speech.org/archive/interspeech_2005/i05_1781.html.

Welch, L. R. (2003): "Hidden Markov Models and the Baum-Welch algorithm", *IEEE Information Theory Society Newsletter*, 53/4, 1;10-13.

Zolnay, A.; R. Schluter & H. Ney (2005): "Acoustic feature combination for robust speech recognition", in *Proceedings of ICASSP '05,* vol. 1, 457-460. http://ieeexplore.ieee.org/iel5/9711/30650/01415149.pdf?isnumber=30650&prod=CNF&arnumber=1415149&arSt=+457&ared=+460&arAuthor=+Zolnay%2C+A.%3B++Schlueter%2C+R.%3B++Ney%2C+H.