UNIVERSITAT POLITÈCNICA DE VALÈNCIA
ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA INFORMÀTICA
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

# Adaptació massiva de models acústics en transcripció de vídeo-xarrades en valencià.

Projecte Final de Carrera - Enginyeria Informàtica

Juan Eudaldo Mataix Sempere

Supervisat per:
Dr. Alfons Juan Císcar
Dr. Jorge Civera Saiz

30 de setembre de 2013

*Als companys de València.*
*Als companys del laboratori, en especial a Joan Albert.*
*A Pepe i Jorge, per haver compartit aquestos 5 anys.*
*I sobretot a ma mare i al meu oncle Mere.*
*Gràcies!*

# Índex

Capítol 1

# Introducció

## 1.1 Motivació

En la societat actual la quantitat d'informació en la qual es traballa cada vegada és major. Tradicionalment aquesta informació ha estat emmagatzemada per mitjans escrits, però la tecnologia digital actual ens permet un accés i un tractament de la informació més ràpid. A més a més, amb una societat que cada vegada es recolza més amb Internet es fa necessària la creació de mitjans per poder difondre tota aquesta informació a la màxima quantitat de gent possible. Aquesta informació en un gran nombre de casos és acústica o audiovisual, i en aquest context els sistemes de reconeixement de la parla automàtics prenen una gran rellevància quan parlem de difussió. El fet que aquesta informació siga oïble afegeix una dificultat tecnològica. Un exemple pot ser la cerca d'un segment d'una xarrada educativa on es parla d'una temàtica desitjada. Aquesta tasca, coneguda com cerca semàntica, requereix d'una tecnologia avançada de reconeixement de la parla, doncs cal obtindre primer la transcripció de la parla del senyal acústic per poder realitzar posteriorment la cerca textual que posicione la reproducció de la xarrada en el lloc desitjat per l'usuari.

Amb tot, la subtitolació automàtica de videoxarrades representa una millora ben aparent de l'accessibilitat i la difussió de la informació que pretén transmetre. De l'aplicació d'aquesta tecnologia es poden beneficiar persones amb discapacitats auditives, o les audiències no natives (amb dificultats per entendre el llenguatge), a banda d'habilitar altres tasques relacionades com la cerca semàntica. Cal remarcar a més, en comparació amb la generació manual de subtítols, els beneficis que ofereixen aquestes tecnologies en quant a la reducció d'esforç humà, i en conseqüència, del cost econòmic.

Malgrat els grans avanços tecnològics la transcripció automàtica no està exempta d'errors, especialment en llengües minoritàries, com és el cas del valencià, on l'obtenció d'un volum gran de dades per a l'entrenament de models estadístics pot ser complicat.

L'objectiu d'aquest projecte és fer un estudi de com millorar les prestacions d'aquests sistemes, prenent el valencià com a cas d'estudi. Per tant aquest projecte se centrarà en la transcripció de vídeo-xarrades generades a la Universitat Politècnica de València, concretament al repositori anomenat poliMèdia, tasca que està emmarcada

dintre del projecte europeu transLectures.

## 1.2 Antecedents històrics del reconeixement de la parla

Els primers intents per a reconéixer la parla automàticament van començar a partir del 1950 als Laboratoris Bell. Al 1952 els laboratoris Bell van construir un sistema per al reconeixement de dígits aïllats per a un locutor [DBB52]. Aquest sistema es basava en mesurar la ressonància durant la pronúncia de les vocals. Uns anys més tard (1956) als Laboratoris RCA es va intentar el reconeixement de deu síl·labes amb la mateixa tècnica [OB56]. Tres anys més tard a la University College d'Anglaterra es va construir un reconeixedor de quatre vocals i nou consonants [Fry59]. El més novedós que va aportar aquest estudi va ser l'ús de certa informació estadística respecte a les seqüències vàlides de fonemes en anglés, que es pot entendre com una definició de sintaxi rudimentària.

En els anys seixanta es van desenvolupar tecnologies que van ser fonamentals per als actuals sistemes. Aquestes tecnologies van anar adreçades a donar una solució al problema de distingir quan està parlant el locutor i quan no, i al problema de la no uniformitat temporal de la parla. El problema consisteix a que no sempre una paraula dura les mateixes unitats de temps, podem parlar més ràpid o més lent. T.B. Martin junt als seus companys als Laboratoris RCA van ser els responsables d'aquestos avanços [MaZ64].

Una dècada després i gràcies a l'esforç dels Laboratoris Bell, el reconeixement de paraules aïllades va passar a ser una tecnologia viable i usable. Havent aconseguit aquest èxit, els Laboratoris Bell van seguir més enllà i van dirigir els seus esforços en el reconeixement de parla contínua i en que els reconeixedors foren independents del locutor.

Així com als anys setanta l'objectiu va ser el d'aconseguir el reconeixement de paraules aïllades, als huitanta l'objectiu principal va ser aconseguir crear un sistema capaç de reconéixer un encadenament de paraules pronunciades de forma fluida. Al principi de la dècada l'aproximació va ser cap a models basats en plantilles, però després es van desenvolupar mètodes estadístics que van donar millors resultats. Aquestos mètodes estadístics es continuen utilitzant a l'actualitat i s'anomenen Models Ocults de Markov [Rab89]. Altra tecnologia ja existent, les xarxes neurals, que es van desenvolupar teòricament amb independència del reconeixement de la parla, van resultar útils i es van introduir en els sistemes a finals de la dècada [WHH+89].

Des dels anys noranta fins a avui els esforços se centren en aconseguir reconeixedors independents del locutor amb grans vocabularis de paraules, fent servir les tecnologies creades als anys huitanta i millorant-les amb noves tecnologies o creant tecnologies mixtes. Un exemple actual de l'ús del reconeixement de la parla el podem veure als serveis de telefonia on l'usuari descriu el seu problema i el sistema automàticament redirigeix l'usuari al departament de l'empresa escaient.

## 1.3 Reconeixement de la parla contínua

Un sistema reconeixedor de la parla és aquell que a partir de senyals acústics produeix la corresponent transcripció acústica. El sistema treballa amb un vocabulari finit de paraules les quals podrà reconéixer i les quals apareixeran en el text transcrit. Com estem tractant el reconeixement de parla contínua el vocabulari haurà de ser suficientment gran com per a contindre les paraules més comuns. Amb un cop d'ull es pot veure que el vocabulari haurà de contenir diversos milers de paraules. Si pensem d'afegir un verb, també haurem d'afegir la seua flexió completa. Açò ens dóna una idea de la quantitat de dades que el sistema requereix.

No obstant això, dirigir els esforços en un reconeixedor per a qualsevol tipus de temàtica pot ser una tasca molt costosa i que no aporte resultats satisfactoris. Aquest fet és degut a l'íntima relació que hi ha entre el vocabulari i la tasca. Si per exemple es pretén transcriure vídeo-xarrades d'arquitectura, s'hauran d'incloure al vocabulari paraules específiques d'arquitectura, no siguent necessàries per exemple paraules de medicina. El reconeixedor farà una cerca dintre d'un conjunt de paraules més petit, i per tant de menor cost.

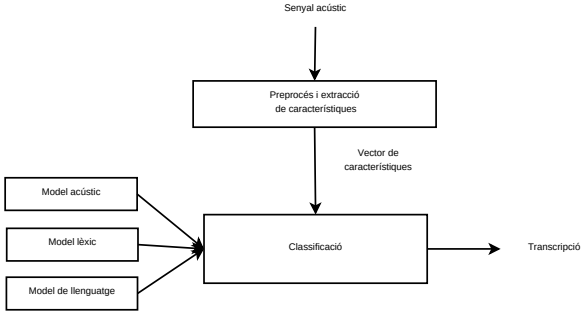


**Figura 1.1:** Sistema bàsic per al reconeixement de la parla

Un esquema bàsic d'un sistema reconeixedor de la parla és el que trobem a la Figura 1.1, que consta de dos passos bàsics.

El primer pas és el preprocés del senyal acústic, que s'encarrega en primer lloc d'enaltir les qualitats de la parla i de minimitzar l'efecte d'altres informacions acústiques no desitjades (soroll, música, etc), i en segon lloc, de generar una representació de les característiques més rellevants del senyal acústic que permeten distingir entre diferents pronunciacions fonètiques. Aquestes característiques s'agrupen en un vector, que a partir d'ara anomenarem vector de característiques.

Amb el senyal acústic preprocessat, el següent pas és classificar els vectors de característiques. En aquest procés intervenen tres models: els models acústic, lèxic i de llenguatge. En primer lloc, el model acústic s'encarrega de modelar estadísticament les diferents formes de pronunciar un determinat fonema atenent a propietats temporals (pronunciació lenta o ràpida) i acústiques (pròpies del sexe, edat, propietats fonològiques del locutor com la longitud del tracte vocal, etc.) En segon lloc, el

model lèxic conté informació sobre com agrupar fonemes per formar paraules. Finalment, el model de llenguatge s'encarrega de modelar estadísticament com es combinen les diferents paraules del vocabulari per formar frases del llenguatge semànticament correctes.

Aquests tres models intercanvien informació a diferents nivells (fonètic, lèxic, semàntic) per tal de trobar la transcripió més probable del senyal acústic. De forma intuïtiva, es pot veure com que el model acústic genera la seqüència de fonemes que millor explica la senyal acústica d'entrada (nivell fonètic), tot baix la supervisió, d'una banda, del model lèxic per tal que les seqüències de fonemes siguen vàlides (nivell lèxic), i d'altra banda, del model de llenguatge perquè les seqüències de paraules generades formen frases amb sentit (nivell semàntic).

El models s'obtenen aplicant tècniques d'aprenentatge estadístic prenent un conjunt de dades d'entrenament. En el cas del model acústic parlem d'exemples de pronunciació de fonemes, o en el cas del model de llenguatge, d'exemples de frases del llenguatge.

## 1.4 Formulació matemàtica

Tenim un conjunt de mostres a etiquetar $E$ i un conjunt d'etiquetes de classe $\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_c\}$ a assignar a cada mostra $x \in E$. Volem esbrinar la probabilitat a posteriori de que $x$ pertany la classe $\omega_i$, és a dir, $P(\omega_i \mid x)$. Aplicant la Regla de Bayes, obtenim:

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)} \tag{1.1}$$

Des del punt de vista del reconeixement de la parla podem definir aquest problema utilitzant la Regla de Bayes [RJ93]. Donat una seqüència de vectors de característiques $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, ..., \boldsymbol{x_n}\}$ volem trobar la seqüència de paraules més probable $\hat{W} = \{w_1, w_2, w_3, ..., w_m\}$:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W \mid X) \tag{1.2}$$

Donat que la probabilitat a posteriori $P(W \mid X)$ és difícil d'estimar, s'aplica la regla de Bayes (Eq. 1.1), obtenint:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W \mid X) = \underset{W}{\operatorname{argmax}} \frac{P(W)P(X|W)}{P(X)} \tag{1.3}$$

Com que la maximització no depén del conjunt de vectors de característiques $X$, podem prescindir del terme $P(X)$, simplificant l'equació:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W)P(X \mid W) \tag{1.4}$$

L'equació 1.4 és la fòrmula fonamental del reconeixement automàtic de la parla, sent el terme $P(X \mid W)$ modelat pel model acústic, mentre que $P(W)$ és modelat pel model de llenguatge.

## 1.5 Extracció de característiques

Les característiques que s'extrauren del senyal acústic són uns coeficients anomenats MFCC(Mel Frequency Cepstral Coefficients) [SS12], coeficients cepstrals en les freqüències de Mel. Aquestos coeficients són una aproximació que modela la forma en la que els humans percebem els sons. Es calcula de la següent forma:

1. El senyal acústic es discretitza, es defineix una finestra que recorrerà el senyal i s'aplica la Transformada de Fourier a cada extracte de la finestra.

2. Aplicar una correspondència entre l'energia de l'espectre obtinguda abans a l'escala de Mel utilitzant una finestra triangular.

3. Càlcul del logaritme de l'energia per a cada una de les freqüències de Mel.

4. Aplicar la transformada del cosinus discreta a la llista de log energia com si es tractara d'un senyal.

5. Els MFCC són les amplituts de l'espectre resultant.

Els MFCC s'utilitzen perquè fan que el tractament dels senyals acústics siga més eficient. Tanmateix, els MFCC tenen una debilitat, no són robustos front al soroll en el senyal. Existeixen algunes tècniques per a suavitzar aquest problema, entre altres l'adaptació de models acústics.

## 1.6 Models acústics

Representem una paraula com a una concatenació de les unitats fonètiques que la componen. Típicament una unitat acústica és un fonema de la llengua. Podem trobar diferents tipus d'unitats depenent de si tenen en compte el context o no. Les unitats que són independents del context s'anomenen monofonemes i les que depenen del context polifonemes. Per a tasques de poca complexitat solen usar-se els monofonemes, però en tasques de més alta complexitat i amb grans vocabularis, com és el reconeixement de la parla, s'utilitzen els polifonemes.

Podem tindre diversos tipus de polifonemes depenent de la quantitat de context que recullen. El més utilitzat és el trifonema, el qual recull la informació del fonema anterior i del posterior.

L'aproximació més acceptada per a les unitats acústiques són els Models Ocults de Markov (Hidden Markov Models). Els HMMs s'utilitzen per a modelar la probabilitat d'observar una seqüència. Els HMMs assumixen que la seqüència ha sigut generada per un màquina d'estats finits en la qual cada estat genera una observació d'acord amb una certa distribució de probabilitat. La seqüència d'estats responsable de la generació de la seqüència roman desconeguda, és per això que s'anomenen models ocults[Jel97]. Formalment els HMMs estan definits per la tupla $(Q, I, F, X, a, b)$ on:

- Conjunt finit $Q$ d'estats que inclou l'estat inicial i l'estat final.

- $I$ és l'estat inicial del HMM.
- $F$ és l'estat final del HMM.
- $X$ és un espai de nombres reals multidimensionals de vectors $x$
- $a$ és una funció de distribució de probabilitat de transició entre els estats $q_i$ i $q_j$.
- b és una funció de densitat de probabilitat d'emetre un vector $x \in X$ en un estat $q_i \in Q$.

Per a modelar la probabilitat d'emissió d'un vector $x \in X$ en un estat $q_i \in Q$ del HMM s'utilitza una mixtura de Gaussianes. La grandària d'aquesta mixtura s'haurà de determinar experimentalment per tal de trobar l'òptim, ja que un nombre massa gran de mixtures resultaria en un model que no seria capaç de generalitzar a partir de les mostres donades; i un nombre massa menut de mixtures donaria un model excessivament general.

La topologia dels HMMs sol ser l'anomenada d'esquerra a dreta. En aquesta topologia no es permet una transició cap a un estat anterior però es pot fer una transició al mateix estat. Açò té com a objectiu que el model done compte de l'evolució temporal del senyal acústic. Les transicions al mateix estat modelen el fet que al pronunciar una paraula no sempre triguem el mateix temps.

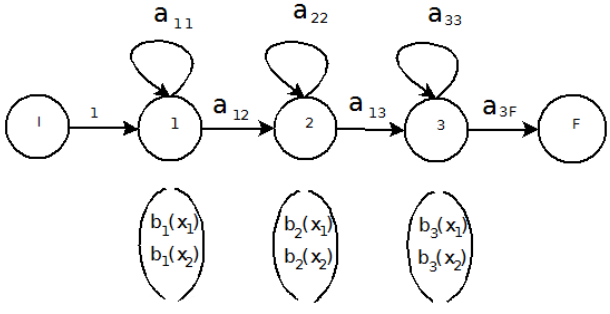


**Figura 1.2:** Model de Markov: $a_{ij}$ és la probabilitat d'anar de l'estat $i$ a l'estat $j$, $b_k(x_t)$ és la probabilitat d'emetre $x_t$ a l'estat $b_k$

.

L'estimació dels paràmetres dels HMMs es du a terme mitjançant l'algoritme de Baum-Welch que utilitza tècniques de màxima versemblança. Aquest algoritme també és conegut com l'algoritme Forward-Backward.

Si utilitzem polifonemes el nombre d'estats del HMM pot arribar a créixer excessívament, és per això que algunes unitats fonètiques s'agrupen sota la mateixa distribució de probabilitat i d'aquesta forma es redueix el nombre d'estats. Aquestes tècniques s'anomenen CART (Classification and Regression Trees). Un exemple pot ser agrupar les unitats fonètiques que continguen la $a$ seguida d'una consonant bilabial, així agruparíem els estats de la $a$ seguida de $b$, la $a$ seguida de $p$ i la $a$ seguida

de $m$. No cal oblidar que les tècniques CART són automàtiques i poden crear altre tipus d'agrupacions, no necessàriament del tipus abans esmentat.

## 1.7 Models lèxics

El model lèxic s'encarrega de modelar com es concatenen fonemes de forma vàlida per formar paraules. Aquest model és una espècie de diccionari en el que com a entrades apareixen totes les paraules del vocabulari del sistema, definint per a cadascuna d'elles les seves corresponents transcripcions fonètiques, que poden ser més d'una atenent a les diferents pronunciacions d'una mateixa paraula. A la Taula 1.1 es mostren uns exemples de paraules i les seues transcripcions fonètiques corresponents:

**Taula 1.1:** Transcripció fonètica del vocabulari

| **Paraula** | **Transcripció fonètica** |
|---|---|
| bresquilla | b r e s k i L a |
| casa | k a z a |
| espardenya | e s p a r d e N a |
| gat | g a t |
| Jerusalem | G e r u z a l e m |
| metgessa | m e t G e s a |
| socarrat | s o c a R a t |

## 1.8 Models de llenguatge

El model de llenguatge s'encarrega de modelar de forma estocàstica com formar construccions sintàctiques correctes [CG98]. La tecnologia més utilitzada per als models de llenguatge són els $n$-grames.

Un $n$-grama és una sub-seqüència d'$n$ elements extrets d'una seqüència de text. En el cas que ens ocupa, un grama es correspon a una paraula. Depenent de l'ordre de l'$n$-grama rep el nom d'unigrama (1-grama), bigrama (2-grama), trigrama (3-grama) i així successívament. A la Taula 1.2 trobem la divisió en $n$-grames de la frase *setze jutges mengen fetge*.

Des del punt de vista estadístic un model d'$n$-grames calcula la probabilitat que una frase $w$ siga formada per les paraules $w_1, w_2, ..., w_I$ amb la següent fórmula:

$$p(w) = \prod_{i=1}^{I} p(w_i|w_1, ..., w_{i-1}) \tag{1.5}$$

La probabilitat de la frase $w$ es calcula com el producte de la probabilitat a posteriori de cadascuna de les paraules que conformen la frase donades les paraules

**Taula 1.2:** Agrupament en $n$-grames

| 1-grames | 2-grames | 3-grames | 4-grames |
|---|---|---|---|
| setze | setze jutges | setze jutges mengen | setze jutges mengen fetge |
| jutges | jutges mengen | jutges mengen fetge | |
| mengen | mengen fetge | | |
| fetge | | | |

immediatament anteriors (història). No obstant, no és factible emmagatzemar tota la història de paraules vistes, doncs això implicaria una explosió del nombre de paràmetres del model, que a més no serien estimats correctament per falta de dades d'entrenament. És per això que els models d'$n$-grames són limitats típicament a ordres de 2, 3 o 4, depenent de la quantitat de dades d'entrenament disponibles. Si consideràrem 3-grames, l'equació seria:

$$p(w) = \prod_{i=1}^{I} p(w_i|w_{i-2}, w_{i-1}) \tag{1.6}$$

Utilitzant l'exemple anterior la probabilitat de la frase *setze jutges mengen fetge* es calcula de la següent forma:

$$\begin{aligned} p(setze\ jutges\ mengen\ fetge) = \\ p(setze) * p(jutges|setze) \\ * p(mengen|setze, jutges) \\ * p(fetge|jutges, mengen) \end{aligned}$$

Com a contrapartida, els models d'$n$-grames tenen el problema que poden trobar-se amb una seqüència de paraules lingüísticament vàlida no vista en les dades d'entrenament, a la qual se li assignaria incorrectament una probabilitat de zero. Aquesta falta de generalització es soluciona parcialment aplicant tècniques de suavitzat, les quals resten massa de probabilitat a les observacions més freqüents per a després repartir-la entre les menys freqüents i esdeveniments no vists en l'entrenament del model[KN95b].

## 1.9 Generació de mostres d'entrenament

A l'hora de crear exemples per a l'entrenament dels models s'han seguit una sèrie de regles amb l'objectiu d'aconseguir millors resultats. Les mostres s'han generat amb l'aplicació Transcriber, la qual genera un fitxer de tipus XML en el que anota l'interval de temps i el que s'ha dit. A continuació descric les normes utilitzades:

- Cal crear segments que siguen frases amb sentit propi.
- Si el locutor fa una pausa llarga, com pot ser la pausa entre frases, s'haurà de crear un segment amb l'etiqueta *[sonido de fondo]* exclusivament. De forma que anotem el que diu quan parla i quan no ho fa anotem que no està parlant.
- Si el locutor fa una pausa curta dintre d'una frase utilitzarem l'etiqueta */SF//*.
- Si el locutor repeteix paraules o pronuncia paraules malament s'anotarà amb barres de forma que s'escriurà */el que diu/el que hauria d'haver dit/*. D'aquesta forma el model acústic s'entrenarà amb la part esquerra i el de llenguatge amb la part dreta. En el cas de la repetició l'etiqueta serà */paraula repetida//*.
- En el cas de dubtes es considerarà com el cas de paraules repetides. Per exemple:

  *les característiques deeee (silenci) de la façana*

  *les característiques /deeee// /SF// de la façana*
- Per a les paraules en altres idiomes es seguirà la mateixa regla: *Polsem el botó /estart/start/*
- Els nombres, les lletres i les fórmules s'escriuran sempre amb lletra. */efa/f/ sub u al quadrat menys tres mil dos-cents*

## 1.10 Mètodes d'Avaluació

Per tal d'avaluar la qualitat d'un sistema automàtic, i poder comparar-lo amb altres sistemes, sorgeix la necessitat de disposar d'un mètode d'avaluació. Existeixen dues alternatives diferents: supervisió humana o avaluació automàtica. L'avaluació automàtica presenta una sèrie d'avantatges sobre la humana, com són la preservació de l'objectivitat i la rapidesa del mètode: mentre que un humà pot tardar hores en avaluar un sistema de transcripció de veu, un mètode automàtic pot realitzar-ho en uns pocs segons. No obstant, com a contrapartida, els mètodes automàtics requereixen de la definició d'un conjunt d'avaluació que permeta comparar l'eixida del reconeixement del sistema amb la transcripció correcta, a la que anomenarem referència.

En reconeixement de la parla, la mètrica d'avaluació automàtica més comunament emprada és el WER (Word Error Rate), que representa el nombre mínim d'insercions, esborrats o substitucions a nivell de paraula necessàries per tal de transformar (corregir) la transcripció automàtica en la referència. El WER es calcula sumant el nombre d'operacions d'inserció ($I$), esborrat ($E$) i substitució ($S$) i dividint pel nombre de paraules de la referència ($N$):

$$\text{WER} = \frac{I + E + S}{N} \tag{1.7}$$

Aquesta mesura és pot entendre de forma intuïtiva com el percentatge de paraules incorrectament transcrites pel sistema automàtic i que caldria corregir per obtindre una transcripció perfecta.

Capítol 2

# Reconeixement de la Parla

És objectiu d'aquest capítol l'aprofundiment i l'enteniment dels sistemes reconeixedors de la parla contínua. Aquest capítol se centrarà en què és el model acústic, com s'entrena, què és una adaptació del model acústic i com es realitza aquesta. Finalment es veurà com s'aconsegueix reconéixer la parla a partir dels models generats.

## 2.1 Adaptació de models acústics

Malgrat els bons resultats dels reconeixedors de la parla amb grans vocabularis encara existeix una gran variabilitat dels resultats quan tenim en compte a diversos parlants. Els resultats poden arribar a empitjorar molt si un nou parlant és molt diferent dels parlants que apareixen en les dades d'entrenament. Aquestes diferències poden vindre per característiques anatòmiques, com la gola i les cavitats bucals i nasals; i també per hàbits del parlant, com per exemple el seu accent o el dialecte que parla. Una tècnica molt extesa que pot millorar els resultats és la d'adaptar el model acústic a les característiques del locutor, al canal i a la tasca.

### 2.1.1 Vocal Tract Length Normalization

Vocal Tract Length Normalization (VTLN) o Normalització de la Longitud del Tracte Vocal tracta de reduir la variabilitat acústica entre els parlants degut a característiques anatòmiques. Açò es pot modelar amb una deformació lineal de l'espectre de freqüències del locutor. Típicament s'estima una mitja global de l'espectre amb les dades d'entrenament i es calcula un factor d'escala a aquesta mitja per a cada locutor. En reconeixement quan arribe un nou locutor es calcula el seu factor d'escala i s'aplica. Aplicant aquestos factors les dades queden més homogènies. L'estimació d'aquests factors es du a terme mitjançant tècniques de cerca en un espai de paràmetres (grid search) que maximitzen la versemblança del locutor en respecte a un model independent del locutor[GGB04].

### 2.1.2 Constrained Maximum Likelihood Linear Regresion

La tècnica Constrained Maximum Likelihood Linear Regresion (CMLLR) té com a objectiu adaptar el model acústic mitjançant una transformació afí dels vectors acústics. La mitja i la variància d'una densitat Gaussiana associada a un HMM s'adapta amb la següent equació [Gal97]:

$$\hat{\mu} = A\mu + b, \quad \hat{\sigma} = A\sigma A^t \tag{2.1}$$

On $A$ i $b$ són la matriu i el vector de la transformació i $A^t$ la matriu resultant de transposar $A$. La mitja es representa per $\mu$ i la variància per $\sigma$. Aquesta equació afegeix una restricció, que la transformació ha de ser la mateixa per a la mitja i per a la variància. En el cas que no existisca esta restricció la tècnica s'anomena MLLR [Gal97].

La matriu $A$ pot ser bloc-diagonal o diagonal. A [Gal97] s'observa que utilitzar una matriu bloc-diagonal dóna millors resultats però el cost computacional és significativament major, i segons els resultats no s'obté una millora que justifique el cost computacional. L'algoritme utilitzat per a l'estimació de $\mu$ i $\sigma$ és el conegut Expectation-Maximization (EM) [DLR77].

## 2.2 Entrenament de models acústics

En aquesta secció es descriuen el passos duts a terme per entrenar un model acústic adaptat mitjançant la tècnica CMLLR.

Tant per a entrenament com per a reconeixement el preprocés consisteix en extraure l'àudio de la xarrada i transformar-lo en format WAV amb una freqüència de $16KHz$ i un sol canal. A continuació s'aplica la transformació MFCC (veure Secció 1.5), que genera els vectors de característiques. Després, depenent de si es vol entrenar un model o reconéixer una xarrada es seguirà un procés diferent.

Una vegada ja s'han extret els vectors de característiques es segueixen tres passos per a l'entrenament del model acústic, que més avant es descriuen amb més atenció:

1. S'entrena un model no normalitzat que s'empra per realitzar un primer reconeixement, al que anomenarem model estàndard. Una versió simplificada d'eixe model serà emprat posteriorment en l'adaptació CMLLR, al qual anomenarem model target.

2. S'adapten totes les mostres d'entrenament utilitzant la tècnica CMLLR sobre el model target.

3. S'entrena el model acústic adaptat amb les mostres adaptades.

### 2.2.1 Obtenció dels models estàndard i target

En primer lloc les transcripcions d'entrenament s'han de transformar. Es parteix d'una transcripció on hi ha frases i s'ha d'arribar a una on hi haja fonemes. Aquesta

tasca la durà a terme un transliterador fonètic. Donat que el valencià té una relació grafema-fonema molt alta aquestos transliteradors no tenen gran complexitat. A la Taula 2.1 es mostren uns exemples de frases transliterades. Cal fixar-se que s'afigen unes etiquetes ($SP$) amb l'objectiu de marcar els petits silencis entre paraules.

**Taula 2.1:** Transliteració fonètica

| Frase | Transcripció fonètica |
|---|---|
| Hola bon dia | SP o l a SP b o n SP d i a SP |
| Vull un gelat de torró | SP v u L SP u n SP g e l a t SP d e SP t o R o SP |

En segon lloc, s'agafen aquestes mostres transformades i s'entrena un model acústic. Aquest model tindrà un HMM de tres estats [GGB04] per a cada fonema i a més hi haurà un HMM especial (per al silenci) amb només un estat i que tindrà tindrà transicions directament de l'estat inicial i a l'estat final. A la Figura 2.1 es mostra un exemple amb la paraula "hola".

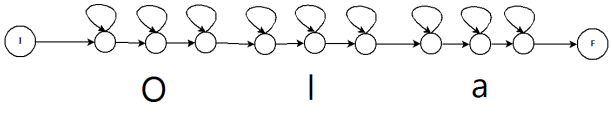


**Figura 2.1:** Exemple de HMM per a la paraula "Hola".

Les probabilitats d'emissió venen governades per una única Gaussiana que s'inicialitzarà al valor de la mitja i la variància de totes les mostres d'entrenament (per a tots els estats els mateixos valors). A continuació es realitzen vuit iteracions de l'algoritme EM.

En tercer lloc es tornaran a transliterar les transcripcions. Aquesta vegada l'objectiu és passar els fonemes a trifonemes. Un trifonema emmagatzema informació respecte al fonema anterior i posterior. A la Taula 2.2 es representa la transliteració de frases en trifonemes on el símbol $-$ fa referència al fonema anterior i el símbol $+$ al posterior.

**Taula 2.2:** Exemple de transliteració en trifonemes.

| Frase | Transcripció fonètica |
|---|---|
| Hola bon dia | SP o+l o-l+a l-a SP b+o b-o+n o-n SP d+i d-i+a i-a SP |
| Tinc fam | SP t+i t-i+n i-n+c n-c SP f+a f-a+m a-m SP |

Si ens fixem, en el model de monofonemes, el fonema $i$ només tenia un HMM associat. Ara tenim diversos tipus de $i$: la $i$ entre $t$ i $n$ $(t-i+n)$ i la $i$ entre $d$ i $a$

$(d - i + a)$.

A continuació generem un model amb trifonemes, que s'emprarà posteriorment per a inicialitzar el model estàndard. El procés és el següent. Per a cada trifonema associem un HMM que inicialitzarem amb el model monofonema. Per exemple, tots els HMM associats als trifonemes de la $i$ tindran inicialment els mateixos paràmetres que el HMM del monofonema $i$. A més, i seguint amb l'exemple de la $i$, per qüestions de sobreespecialització tots els HMM del trifonema $i$ mantindran les mateixes probabilitats de transició [YEK09]. A continuació s'entrena el model de trifonemes amb quatre iteracions de l'algoritme EM.

Generar tots els possibles trifonemes dóna com a resultat un nombre d'estats massa gran, aleshores s'apliquen tècniques CART per a fer agrupacions de trifonemes [YEK09] i reduir el nombre d'estats. Després d'aplicar la tècnica CART aplicarem quatre iteracions de l'algoritme EM. A quest model per una banda l'utilitzarem per a inicialitzar el model estàndard final, i per altra banda com a model target.

Fins ara només s'han utilitzat mixtures d'una component. Ara es segueix un procés per a augmentar el nombre de components de les mixtures i obtindre el model estàndard final.

1. Dividim les components de mixtura. Per a cada component la divisió es fa així:
   (a) 'observa l'ocupació de la component (nombre de mostres que han passat per la component).
   (b) Si l'ocupació és major que un llindar es divideix en dues. Si és menor no es divideix.
   (c) Una de les noves components rep la mitja més un factor i l'altra la mitja menys el mateix factor. La variància roman igual.
2. Es realitzen quatre iteracions de l'algoritme EM amb el nou model.
3. Tornem al pas 1 fins que s'arribe al nombre de components desitjat.

### 2.2.2 Normalització CMLLR

Una vegada s'ha obtingut el model target, s'aplica una adaptació CMLLR. Per a cada vídeo-xarrada es calcula una matriu de transformació CMLLR sobre el model target, però en lloc de crear nous models adaptats, el que es fa és aplicar la transformació inversa a les mostres de la vídeo-xarrada [GGB04], ja que resulta una transformació equivalent. Així es consegueix tindre totes les mostres normalitzades i adaptades. L'adaptació es fa a nivell de xarrada perquè a poliMèdia només hi ha un parlant per xarrada. Si hi haguera més parlants s'hauria de calcular una transformació per a cada parlant.

### 2.2.3 Model adaptat

Per a l'obtenció del model acústic final (l'adaptat) es repeteix tot el procés descrit a la Secció 2.2.1 però amb les mostres normalitzades.

## 2.3 Reconeixement

Una vegada entés el procés d'entrenament del model acústic, el reconeixement és molt senzill. S'utilitzen els tres models generats a l'entrenament: El model estàndard, el model target, i el model adaptat. A continuació es descriuen els passos per a obtindre la transcripció automàtica d'una vídeo-xarrada:

- Es preprocessa la video-xarrada (extracció dels vectors de característiques a partir del senyal acústic).
- Amb el model estàndard es genera una transcripció inicial amb errors.
- La transcripció inicial s'empra juntament amb el model target per estimar una transformació CMLLR que s'aplica a les mostres de la vídeo-xarrada.
- Per últim, es reconeixen les mostres adaptades amb el model acústic adaptat i es genera la transcripció final de la vídeo-xarrada.

Capítol 3

# Còrpora

En aquest capítol es descriuen els còrpora emprats per entrenar els models acústics, lèxics i de llenguatge per a la generació d'un sistema de reconeixement de la parla en valencià (variant català occidental).

## 3.1 Còrpora acústic

Per entrenar models acústics es requereixen conjunts de dades amb exemples de pronunciacions de frases, paraules, o fonemes, i les seues respectives transcripcions. En aquest projecte hem pres com a còrpora de referència el còrpora poliMèdia en català (occidental), sobre el qual s'ha definit una partició experimental per tal d'avaluar les prestacions del sistema. A més a més, dos còrpora addicionals han sigut emprats per augmentar el conjunt d'entrenament dels models acústics: Àgora i Glissando. La Taula 3.1 mostra la duració en hores de cadascun d'aquests còrpora.

**Taula 3.1:** Duració del còrpora acústic

| **Còrpora** | **Duració** |
|---|---|
| poliMèdia | 23.9h |
| Glissando | 6h |
| Àgora | 45h |

### 3.1.1 poliMèdia

La plataforma poliMèdia és un servei recentment creat per la UPV per a la creació i distribució de contingut multimèdia educacional. Va començar el 2008 i des d'aleshores s'han enregistrat 9222 xarrades per més de 1300 locutors, el que suma aproximadament 2100 hores de vídeo. Ha sigut dissenyada per a la producció de continguts

d'alta qualitat. La UPV disposa de sales de gravació on el professor o alumne té tota classe d'equipament per a enregistrar la xarrada. A la Figura 3.1 es veu un exemple d'aquestes sales.

**Figura 3.1:** Sala de gravació de poliMèdia

Al vídeo final el locutor queda al marge dret i a la resta de l'espai es col·loca la transparència amb la qual s'ajuda el locutor. A més de generar un vídeo aquestes sales també obtenen la sincronització de cada transparència en quins moments del vídeo s'ha utilitzat. A la Figura 3.2 es mostra la reproducció d'un vídeo de poliMèdia on a la part de baix es poden veure les transparències que usa el locutor.

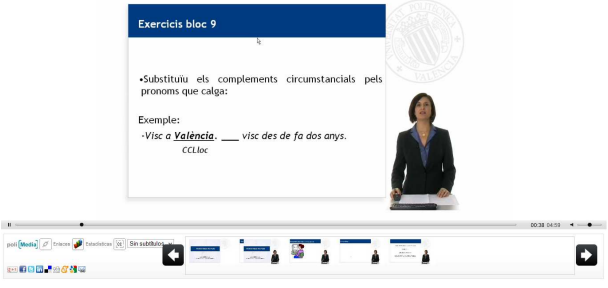


**Figura 3.2:** Exemple d'un vídeo de poliMèdia

Per tal d'entrenar un sistema capaç de transcriure vídeos del repositori poliMèdia en català, un total de 12.3h de videoxarrades en aquest idioma foren transcrites de forma manual seguint les directrius descrites a la Secció 1.9. Aquestes dades estan partides en tres conjunts: entrenament, desenvolupament, i test. El primer d'ells s'empra per entrenar models acústics, el segon per ajustar paràmetres dels models acústics i de reconeixement, i el tercer per avaluar la qualitat del sistema.

Posteriorment, es van transcriure manualment 11.6h addicionals de videoxarrades per afegir-les al conjunt d'entrenament. Així, cal distingir entre el conjunt d'entrenament original (7.9h) i el conjunt d'entrenament extés (19.5h).

La Taula 3.2 mostra estadístiques bàsiques dels quatre conjunts adés mencionats. Al realitzar la partició s'ha tractat de mantindre, en la mesura del possible, un equilibri entre locutors masculins i femenins.

**Taula 3.2:** Estadístiques bàsiques de la partició de dades del còrpora poliMèdia.

| | **Training** | **Training(extés)** | **Desenvolupament** | **Test** |
|---|---|---|---|---|
| No. xarrades | 51 | 153 | 17 | 16 |
| Duració | 7.9h | 19.5h | 2.3h | 2.1h |
| No. locutors | 12 | 40 | 6 | 6 |

### 3.1.2 Glissando

Glissando [GEA+13] és un còrpora anotat dissenyat especialment per a l'anàlisi de la prosòdia del castellà i el català oriental des de diferents perspectives (fonètica, fonologia, anàlisi del discurs, tecnologia de la parla, estudis comparatius). Les seues característiques principals són:

- Bilingüe: es tracta d'un còrpora paral·lel castellà - català.
- Gravacions de gran qualitat: tot el còrpora ha estat enregistrat en estudis professionals.
- 28 locutors per idioma, tant professionals com no professionals.
- Dos estils de parla diferents: lectura de notícies i diàleg.
- Transcrit ortogràficament i fonèticament.
- Anotat amb diferents nivells d'informació prosòdica.

Per a cada idioma hi ha disponibles més de 20 hores de parla enregistrada. No obstant, per al nostre propòsit només s'ha emprat el subconjunt de lectura de notícies (6h de dades acústiques), ja que el diàleg és un tipus de parla esporàdica i informal que s'allunya del domini del còrpora de referència, poliMèdia.

### 3.1.3 Àgora

Àgora [SF09] és un conjunt de programes televisius emesos en la Televisió de Catalunya transcrits acústicament. El còrpora acústic engloba un total de 45h d'àudio. El so és de bona qualitat i solen aparéixer diversos locutors durant el programa que debaten

un tema prèviament acordat. No obstant, el fet que la parla es trobe contaminada acústicament amb seccions de soroll de fons i música, el solapament de diferents locutors al mateix temps, i el predomini de la varietat oriental de català provoquen que Àgora siga un còrpora fora del domini de poliMèdia.

## 3.2 Corpora textual

Per a l'entrenament de models de llenguatge cal tindre a l'abast exemples de frases de l'idioma en el que es vol transcriure la parla. En aquest cas, s'ha emprat el text de les transcripcions dels còrpora acústics descrits a la secció anterior (poliMèdia, Glissando, Àgora), a més de tres recursos externs addicionals. El primer d'ells és la traducció al català del conjunt d'entrenament de poliMèdia en castellà [SCdAG+12] al català mitjançant el traductor automàtic Apertium [ape13]. El segon és el còrpora El Periódico, que recull notícies en català del diari que rep el mateix nom [elp]. Per últim, trobem el còrpora de la Viquipèdia en català [viq] que engloba tots els articles existents en català dins d'aquesta famosa enciclopèdia digital. La Taula 3.3 mostra estadístiques bàsiques dels còrpora textuals utilitzats.

**Taula 3.3:** Estadístiques bàsiques dels còrpora textuals. Nota: K = milers, M = milions.

| Nom | Paraules totals | Grandària Vocabulari |
|---|---|---|
| poliMèdia (Ca) | 146K | 12K |
| poliMèdia (Es-Ca) | 1M | 18K |
| El periódico | 34M | 254K |
| Glissando | 11K | 3K |
| Àgora | 404K | 26K |
| Viquipèdia | 85M | 1M |

Capítol 4

# Experimentació

En aquesta Secció es descriuen els experiments que s'han dut a terme per tal de construir un sistema de reconeixement de la parla en valencià el més competitiu possible, i que siga capaç de generar transcripcions automàtiques de vídeo-xarrades de poliMèdia dins del marc del projecte transLectures. Per tal d'aconseguir aquest propòsit, s'han estudiat dues vies de millora destinades exclusivament al modelat acústic: d'una banda, augmentar el conjunt de dades d'entrenament, i d'altra banda, aplicar tècniques d'adaptació al locutor, com és l'adaptació CMLLR (veure Secció 2.1.2).

## 4.1 Configuració experimental

S'han definit tres sistemes bàsics diferents atenent al conjunt de dades destinat a entrenar els models acústics (veure Secció 3.1):

- pM: conjunt d'entrenament de poliMèdia (7.9h).
- pM-Ext: conjunt d'entrenament extés de poliMèdia (19.5h).
- pM-Ext+RE: conjunt d'entrenament extés de poliMèdia, més dos còrpora externs fora de domini, Glissando i Àgora (70.5h).

Cadascun dels tres sistemes adés descrits defineixen tres sistemes base, als quals se'ls aplica la tècnica d'adaptació CMLLR, generant els respectius sistemes adaptats.

Per entrenar els models acústics s'ha emprat el transLectures-UPV Open Source Toolkit (TLK) [The], un paquet que aglutina una sèrie de ferramentes per entrenar models acústics i reconèixer senyals d'àudio, entre d'altres. Aquest software proveeix característiques similars a altres toolkits com HTK [Y+95] o RASR [R+09].

En primer lloc, s'han extret els senyals acústics de les dades d'entrenament de cadascun dels sistemes, amb l'objectiu d'obtindre els coeficients cepstrals de Mel (veure Secció 1.5) per a cada mostra. Posteriorment, per a cada sistema, s'han entrenat diferents models acústics de monofonemes i trifonemes, i s'han ajustat paràmetres d'entrenament com el nombre d'estats, components de les mixtures de Gaussianes, fulles del CART de trifonemes, etc. emprant el conjunt de desenvolupament

de poliMèdia (veure Secció 3.1.1). La Taula 4.1 mostra els paràmetres òptims de cada sistema.

**Taula 4.1:** Conjunt de paràmetres d'entrenament optimitzats per a cadascun dels sistemes emprant el conjunt de desenvolupament.

| | **Nombre d'estats** | **Components mixtures** |
|---|---|---|
| pM | 1326 | 64 |
| pM-Ext | 2482 | 64 |
| pM-Ext+RE | 4021 | 128 |

El model de llenguatge és el mateix per als tres sistemes: consisteix en un model de 4-grames fruit d'una interpolació linial [JM80] de diferents models de llenguatge, un per cada corpus d'entrenament descrit a la Secció 3.2, als que se'ls ha aplicat el descompte modificat de Kneser-Ney [KN95a]. Els pesos òptims de la interpolació s'han determinat experimentalment minimitzant la perplexitat del model de llenguatge [JM80] sobre el conjunt de desenvolupament de poliMèdia. El vocabulari del model de llenguatge s'ha definit com el vocabulari dels corpus de poliMèdia (ca) i poliMèdia (es-ca) (dades dins del domini) més les 50.000 paraules més freqüents de la resta dels corpora (dades fora del domini), donant lloc a un vocabulari d'aproximadament 60.000 paraules. Per a l'entrenament del model s'ha emprat la ferramenta SRILM [Sto02].

Pel que respecta al model lèxic, aquest s'ha entrenat generant la transcripció fonètica de totes les paraules del vocabulari del model de llenguatge. En aquest cas s'ha emprat el transliterador català-occidental Xúquer, desenvolupat pel transLectures-UPV team.

## 4.2 Resultats

L'avaluació de les prestacions dels 6 sistemes proposats s'ha realitzat sobre el conjunt de test de poliMèdia (veure Secció 3.1.1) en termes de Word Error Rate (WER) (veure Secció 1.10). La Taula 4.2 mostra les taxes d'error de cadascun d'aquests sistemes.

**Taula 4.2:** WER calculat sobre el conjunt de test de poliMèdia per a cadascun dels sistemes estudiats.

| | **Sistema base** | **Sistema adaptat (CMLLR)** |
|---|---|---|
| pM | 50.5 | 41.9 |
| pM-Ext | 41.3 | 36.2 |
| pM-Ext+RE | 41.1 | 35.3 |

D'una banda, si analitzem aïlladament l'efecte d'augmentar el conjunt de dades d'entrenament, observem que afegir 12 hores de dades del domini aporten una millora d'aproximadament un 20%, mentre que unes 50h de dades fora de domini no aporten gaire millora. Açò ens fa adonar-nos que tota ampliació del còrpora tindrà impacte en els resultats si són dades del domini.

D'altra banda, si aïllem l'efecte d'aplicar la tècnica d'adaptació CMLLR, observem que en el conjunt pM s'obté una millora relativa del 20%, quasi la mateixa millora que afegir 12 hores més a l'entrenament. L'adaptació del model acústic resulta, per tant, en una baixada de l'error significativa.

Per últim, si analitzem l'efecte combinat d'augmentar el conjunt de dades d'entrenament i aplicar adaptació CMLLR, s'observa com es produeix una davallada del 50.5% de WER del sistema base amb menys dades d'entrenament, al 35.3% de WER del sistema adaptat amb més dades d'entrenament, que implica una millora relativa aproximada de la taxa d'error del 30%.

Capítol 5

# Conclusions

## 5.1 Resum

En aquest projecte s'ha realitzat un estudi sobre com millorar les prestacions oferides per un sistema de reconeixement de la parla en valencià aplicant dues vies de millora: augmentar el conjunt de dades d'entrenament dels models acústics, i aplicar de tècniques d'adaptació al locutor de models acústics. L'experimentació s'ha dut a terme dintre del marc del projecte europeu transLectures, prenent el corpus poliMèdia com a tasca de referència per avaluar la qualitat del sistema.

D'una banda, l'augment de les dades acústiques d'entrenament es tradueix en una millora significativa de la taxa d'error (millora relativa al voltant del 20% de WER) si les dades acústiques pertanyen al mateix domini de la tasca. En canvi, si les dades acústiques pertanyen a dominis diferents, el guany és mínim (aproximadament 3% de millora relativa).

D'altra banda, l'aplicació de tècniques d'adaptació de models acústics com és CMLLR millora en tots els casos la qualitat del sistema en un 10-15%, especialment quan la taxa d'error del sistema és molt alta, observant-se una reducció de l'error d'un 20% aproximadament. Aquests resultats corroboren el que ja ens ensenya la literatura [Gal97].

Combinant ambdues millores, passem d'un sistema amb un 50% de WER a un sistema d'un 35% de WER, o el que és el mateix, una reducció relativa de l'error d'un 30%. Cal destacar que el millor sistema obtingut s'ha emprat al projecte transLectures per generar transcripcions automàtiques de vídeo-xarrades en valencià del repositori poliMèdia.

## 5.2 Treball Futur

Malgrat que en aquest projecte s'ha aconseguit millorar en un 30% la taxa d'error del sistema de reconeixement base fins arribar a un 35% de WER, creiem que encara queda un bon marge de millora, que permetrà davallar la taxa d'error per baix del

25% de WER, a partir de la qual entendríem que el sistema ofereix transcripcions acceptables. A continuació es proposen una sèrie de línies de treball orientades a assolir dita fita:

- Augmentar el conjunt de dades d'entrenament pròpies del domini (poliMèdia), com a mínim 50 hores d'entrenament, idealment 100 hores, com ocorre amb el sistema de reconeixement de la parla de poliMèdia en castellà [SCdAG+12].
- Aplicar tècniques d'adaptació de models de llenguatge emprant el text de les transparències de les video-xarrades, o recursos textuals relacionats amb la xarrada [MVdAAFJ13].
- Introduir al sistema tècniques de pre-entrenament discriminatiu basades en xarxes neuronals [DYDA12].

# Bibliografia


[ape13] Apertium. 2013.

[CG98] S.F. Chan and J. Goodman. An empirical study of smoothing techniques. *Tech. Report*, TR-10-98, 1998.

[DBB52] K.H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *Acoustic Soc. Am.*, 24, 1952.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, Series B 39 (1): 1–38, 1977.

[DYDA12] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech & Language Processing*, 20(1):30–42, 2012.

[elp] El periódico. `http://www.elperiodico.cat/ca/`. 2013.

[Fry59] D.B. Fry. Theoretical aspects of mechanical speech recognition. *British Inst. Radio Engr.*, 19, 1959.

[Gal97] M. J. F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12, 1997.

[GEA+13] J. M. GARRIDO, D. ESCUDERO, L. AGUILAR, V. CARDEÑOSO, E. RODERO, C. DE-LA-MOTA, C. GONZÁLEZ, S. RUSTULLET, O. LARREA, Y. LAPLAZA, F. VIZCAÍNO, M. CABRERA, and A. BONAFONTE. Glissando: a corpus for multidisciplinary prosodic studies in spanish and catalan. *Language Resources and Evaluation*, DOI 10.1007/s10579-012-9213-0, 2013.

[GGB04] Diego Giuliani, Matteo Gerosa, and Fabio Brugnara. Speaker normalization through constrained mllr based transforms. *International Conference on Spoken Language Processing*, INTERSPEECH 2004, 2004.

[Jel97] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1997.

[JM80] Frederick Jelinek and Robert L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *In Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, 1980.

[KN95a] R Kneser and Hermann Ney. Improved backing-off for M-gram language modeling. In *Proc. of ICASSP*, pages 181–184, 1995.

[KN95b] R. Knesser and H. Ney. Improved backing-off for m-gram language modeling. *IEEE INt. Conf. on Acoustics, Speech and Signal Processing*, II:181-184, 1995.

[MaZ64] T.B. Martin and A.L. Nelson andH.J. Zadell. Speech recognition by feature abstraction. *Technical design report*, AL-TDR-64-176, 1964.

[MVdAAFJ13] Adrià Martínez-Villaronga, Miguel A. del Agua, Jesús Andrés-Ferrer, and Alfons Juan. Language model adaptation for video lectures transcription. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8450–8454. IEEE, 2013.

[OB56] H.F Olson and H. Belar. Phonetic typewriter. *Acoustic Soc. Am.*, 28, 1956.

[R+09] D. Rybach et al. The RWTH Aachen University open source speech recognition system. In *Proc. of INTERSPEECH*, pages 2111–2114, 2009.

[Rab89] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77, 1989.

[RJ93] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice-Hall International, 1993.

[SCdAG+12] J. A. Silvestre-Cerdà, M. A. del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. translectures. In *Online Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH 2012)*, pages 345–351, Madrid (Spain), nov 2012.

[SF09] Henrik Schulz and José A. R. Fonollosa. A catalan broadcast conversational speech database. *Proceedings of the I IberianSLTech 2009*, 2009.

[SS12] M. Sahidullah and G. Saha. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. 54(4):543–565, 2012.

[Sto02] Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP*, 2002.

[The] The TransLectures-UPV team. The TransLectures UPV toolkit (TLK). `http://www.translectures.eu/tlk`.

[viq] Viquipèdia. *2013*.

[WHH+89] A. Weibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoustics Speech Signal Proc.*, 37, 1989.

[Y+95] S. Young et al. *The HTK Book.* Cambridge University Engineering Department, 1995.

[YEK09] Steve Young, Gunnar Evermann, and Dan Kershaw. *The HTK book version 3.4.* 2009.

# Índex de figures

# Índex de taules