UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf
Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica

Universitat Politècnica de València

Desenvolupament de Constructors de Grafs de Fonemes

Projecte Final de Carrera

Enginyeria Informàtica

Autor: Raül Fabra Boluda

Director: Jon Ander Gómez Adrián

28 de juny de 2013

A mon pare

Agraïments

Gràcies a Dr. Jon Ander Gómez Adrián per la gran dedicació, ajuda, orientació i temps dedicat com a director per a la realització d'aquest projecte.

També vull agrair a tots els meus amics els ànims que m'han donat al llarg de la carrera. En especial a Adrián Palacios Corella, que ha estat ací des que tinc memòria i sense el qual la carrera no haguera sigut el mateix.

Finalment, vull agrair tota la meua família el fet que hagen estat sempre donant-me suport i confiança incondicionalment. I especialment a mon pare, que sempre m'ha encoratjat a estudiar el que realment m'agrada i sense el qual no m'haguera pogut convertir en la persona que sóc.

Resum

El problema a resoldre en aquest projecte pertany a l'àmbit del Reconeixement Automàtic de la Parla. Donada una seqüència d'àudio degudament preprocessada i descodificada a nivell acústic-fonètic, el mòdul a desenvolupar ha de reconèixer els distints fonemes que un locutor ha pronunciat al llarg del seu discurs. A més, també ha de determinar en quin instant de temps comença a pronunciar-se cada fonema i en quin instant de temps acaba la seua pronunciació.

El senyal d'àudio inicial necessita ser preprocessat amb la finalitat d'eliminar la informació que no resulte útil o discriminant per a la tasca de reconeixement. El mòdul encarregat d'aquesta feina és el Parametritzador i produeix com eixida una seqüència de vectors o *frames* acústics que serveixen com entrada al següent mòdul, l'Estimador de Probabilitats Fonètiques.

Rebent com entrada els vectors acústics generats pel Parametritzador, l'Estimador de Probabilitats Fonètiques s'encarrega, com el seu nom indica, de generar una sèrie de vectors de probabilitats fonètiques. Aquestos vectors són de dimensió igual al nombre d'unitats fonètiques que s'estan contemplant en la tasca de reconeixement, i cada element representa la probabilitat de què un fonema haja sigut pronunciat.

La seqüència de vectors de probabilitats generada per l'Estimador de Probabilitats Fonètiques serveix com entrada a distints mòduls del sistema i és enviada mitjançant cues FIFO controlades per semàfors. Un dels mòduls que rep aquesta seqüència de vectors és el Constructor de Grafs de Fonemes. Aquest mòdul és la part central d'aquest projecte.

El Constructor de Grafs de Fonemes va obtenint de la cua FIFO els vectors de probabilitats fonètiques generats per l'Estimador de Probabilitats Fonètiques. A mesura que són obtinguts, es van inserint en un *buffer* circular capaç de emmagatzemar les últimes 10 *frames*. D'aquesta manera, es conserva una finestra temporal capaç d'emmagatzemar 100 ms d'àudio codificats com vectors de probabilitats fonètiques, assumint que s'obté una *frame* cada 10 ms.

Amb la finalitat d'esbrinar quina unitat fonètica es manifesta en cada moment, s'utilitza un criteri basat en llindars de probabilitat que manté un conjunt d'hipòtesis actives. Concretament es treballa amb dos llindars: un per a la detecció i un per a l'extensió. El llindar de detecció serveix per a detectar si un fonema ha sigut pronunciat, mentre que el llindar d'extensió serveix per a determinar en quin moment ha començat a ser pronunciat i en quin moment deixa de ser-ho.

Els grafs generats per aquest mòdul tenen una topologia *Left-To-Right*, són dirigits i acíclics. A més, tots els nodes tenen únicament arcs cap al següent node. Els nodes representen en quin moment comença o acaba la detecció de la pronunciació de qualsevol fonema, i cada transició indica de quin fonema es tracta.

També cal tenir en compte problemes de caràcter més pràctic, com ara evitar considerar que s'ha pronunciat un fonema quan només ha sigut detectat durant una o dues *frames*, i més encara si es tracta de fonemes habitualment llargs, com són les vocals.

També hi haurà fonemes que requeriran un tractament especial, com és el cas dels fonemes africats i oclusius sords, on s'ha de distingir la oclusió prèvia a l'explosió posterior.

Finalment, hi ha que realitzar una sèrie d'experiments per tal d'avaluar la qualitat dels grafs obtinguts: d'una banda, cal ajustar el valor dels llindars per tal que es generen millors grafs. D'altra banda, sobre aquestos darrers grafs s'aplicarà un algorisme de descodificació acústic-fonètica emprant distintes tècniques per tal d'obtenir la millor seqüència continguda en aquestos.

Paraules clau: reconeixement automàtic de la parla, àudio, acústic, fonètic, mòdul, fonema, locutor, discurs, pronunciació, senyal, preprocés, vector, *frame*, descodificació, unitat fonètica, graf, probabilitat, llindar, hipòtesi, detecció, extensió, node, arc

Índex

Capítol 1

Introducció

1.1 Plantejament del Problema

El problema del Reconeixement Automàtic de la Parla o *Automatic Speech Recognition* (ASR) consisteix en prendre com a entrada un senyal d'àudio que continga la pronunciació d'una locució (paraula, frase, discurs, etc.) per part d'un locutor i reconèixer o determinar, de la forma més precisa possible, allò que el locutor ha pronunciat al llarg del seu discurs.

Per un costat podem parlar de reconeixement de paraules aïllades, on el locutor pronuncia paraules separades per silencis i el sistema ha de determinar quines han sigut pronunciades d'entre un conjunt de possibles paraules. Per un altre costat, també es pot dissenyar un sistema per al reconeixement de la parla contínua, on un locutor pronuncia un discurs i el sistema ha de reconèixer allò que s'ha pronunciat. Aquesta última tasca és notablement més complexa que la primera.

La manera d'abordar el problema dependrà entre altres de la finalitat amb què el sistema és creat, les exigències que aquest ha de satisfer o la informació disponible per a la resolució de la tasca. Per exemple, si per a la resolució del problema es compta únicament amb l'àudio que conté el discurs, serà molt més complicat obtenir resultats satisfactoris que si es compta amb altres recursos, com ara transcripcions de l'àudio, vocabularis, models de pronunciació, etc.

Altres aspectes clau a determinar són la unitat de parla mínima a reconèixer i la independència que el sistema ha de tindre. Pel que fa a la unitat mínima de parla, s'ha de decidir si el sistema ha de ser capaç de reconèixer frases, paraules, síl·labes o fonemes. Depenent de quin tipus d'unitat s'elegisca, les tècniques per al reconeixement varien significativament. Pel que fa a la independència del sistema, se li pot exigir que siga independent o

no del locutor, independent o no del llenguatge, etc. Lògicament, quan més independent es pretén que siga el sistema més complexe resulta aportar una solució factible, ja que el corpus de dades requerit per a la tasca ha de ser més variat per tal d'oferir millors resultats.

Per al present projecte, el problema a resoldre consisteix en la implementació d'un mòdul que reconega els distints fonemes que un locutor pronuncia al llarg del seu discurs. A més, el mòdul també ha de determinar en quin instant de temps comença i acaba la pronunciació de cada fonema. Per a la resolució d'aquest problema, en aquest projecte es proposa la utilització d'una estructura en forma de graf com a representació intermèdia d'allò que s'ha pronunciat.

Altres autors han proposat solucions en què s'utilitzen estructures similars com a representació intermèdia del discurs. Dites estructures són emprades amb diverses finalitats, algunes directament relacionades amb el reconeixement[1, 2] i altres que busquen l'execució de certes tasques de forma més eficient, com ara l'alineament de text amb l'àudio o la indexació i recuperació de la informació[3, 4]. Aquestes últimes permeten, per exemple, realitzar la cerca de paraules o termes (donats en forma de locució) directament sobre els fitxers d'àudio indexats. Dites tasques es coneixen com a *Word Spotting* si allò que es cerca són paraules o *Spoken Term Detection* quan allò que es cerca són termes. En alguns treballs, aquesta tasca s'aconsegueix amb independència del vocabulari emprat[5, 6].

1.2 Objectius

El principal objectiu d'aquest projecte és crear un mòdul per a un sistema d'ASR genere grafs de fonemes com a representació intermèdia del discurs i que puguen ser utilitzats per a les següents tasques:

1. Indexació dels discursos i la seua recuperació mitjançant la cerca de termes o paraules, donats també com a àudio.

2. Alineació de l'àudio amb el text que conté el discurs pronunciat.

3. Utilització dels grafs com a etapa intermèdia del reconeixement.

El primer objectiu podrà ser assolit si els grafs de fonemes resulten ser una estructura adequada per a la indexació i la cerca. Pel que fa al segon objectiu, els grafs de fonemes poden servir per a subtitular un vídeo de forma automàtica, donat el seu àudio i un text que continga allò que es pronuncia al llarg del vídeo. En quant a l'últim objectiu, es pretén utilitzar els grafs

de fonemes per a la descodificació acústic-fonètica amb la finalitat d'esbrinar amb la major exactitud possible allò que s'ha pronunciat al llarg del discurs.

1.3 Solució Adoptada

El mòdul a implementar en el present projecte és el Constructor de Grafs de Fonemes o CGF, encarregat d'obtenir grafs de fonemes com representació intermèdia d'una locució (frase, discurs, etc.). Aquestos han de servir per a posteriors processos, ja siguen de reconeixement, d'alineament o de cerca de paraules clau.

Amb els grafs de fonemes els posteriors processos no cal que facen ús de models acústics, treballen directament amb allò que aquest mòdul ha detectat. El CGF és un complement als models acústics, de manera que es pot obtenir una seqüència fonètica sense que calga implementar algorismes de descodificació utilitzats per al reconeixement. No obstant això, els grafs de fonemes també poden ser usats com a entrada d'un algorisme de descodificació per a fer reconeixement de la parla.

El CGF rep vectors de probabilitats fonètiques que tenen dimensió igual al nombre d'unitats fonètiques utilitzades per a la tasca de reconeixement. En cadascun d'aquestos vectors, cada element representa la probabilitat de què un fonema haja sigut pronunciat.

A mesura que el CGF rep els vectors de probabilitats, els insereix en un *buffer* circular que consta de 10 elements, sent així capaç de mantindre les 10 últimes *frames*. Assumint que s'emet una *frame* cada 10 ms, el *buffer* tindrà capacitat per a emmagatzemar 100 ms.

Per esbrinar quines unitats fonètiques és més probable que s'hagen manifestat en cada moment, el CGF utilitzarà un criteri basat en llindars de probabilitat que mantenen un conjunt d'hipòtesis actives. El CGF treballa amb dos llindars: el llindar de detecció i el llindar d'extensió. El primer llindar serveix per a detectar si algun fonema ha sigut pronunciat, mentre que el segon s'utilitza per a determinar en quin instant de temps comença o acaba la seua pronunciació.

La informació obtinguda d'aquesta manera resulta suficient per a construir un graf de fonemes. Una vegada construït el graf associat a una seqüència d'àudio donada, les transicions indiquen les unitats fonètiques detectades al llarg del discurs i la seua probabilitat. Els nodes actuen com a marques temporals entre les deteccions de les unitats fonètiques. Normalment, l'interval de temps en què es detecta la pronunciació d'un fonema comprèn diversos arcs consecutius.

Capítol 2

Descripció del Sistema

2.1 Visió General

Els principals mòduls que formen part d'aquest sistema són el **Parametritzador**, utilitzat per al preprocés del senyal d'àudio; l'**Estimador de Probabilitats Fonètiques** (EPF), encarregat de l'estimació de les probabilitats fonètiques *frame* a *frame*, i el **Constructor de Grafs de Fonemes** (CGF).

El CGF és el mòdul que s'encarrega de la detecció de fonemes dins del sistema d'ASR. Per a aquest mòdul són molt importants les probabilitats fonètiques que s'obtenen de l'EPF, el qual fa ús dels models acústics. L'arquitectura d'aquest sistema segueix l'esquema típic utilitzat per a la resolució de problemes de Reconeixement de Formes.

El sistema també compta amb altres mòduls. Pel que fa a la part de descodificació, hi ha implementades diverses versions d'un descodificador basat en l'algorisme de Viterbi aplicat sobre models de Markov. També hi ha un mòdul que s'encarrega de construir grafs de paraules a partir de les seqüències de vectors amb probabilitats fonètiques, sense passar per grafs de fonemes. Aquest mòdul fa ús del lèxic en arbre. Un altre mòdul s'encarrega de construir grafs de paraules a partir dels grafs de fonemes. Altres mòduls inacabats s'encarreguen de reconèixer a partir dels grafs de paraules.

Al diagrama de la Figura 2.1 es resumeix l'arquitectura del sistema d'ASR (almenys les parts més rellevants per al projecte) en el qual s'integra el CGF.

Com es pot observar, cada mòdul llegeix la seua entrada d'una cua FIFO i insereix la seua eixida en una altra cua FIFO si aquesta informació és usada per un altre mòdul. El sistema està dissenyat d'aquesta manera perquè cada mòdul és executat per un fil diferent, amb la finalitat d'aconseguir una major eficiència paral·lelitzant aquelles tasques que ho permeten. Per a sincronitzar

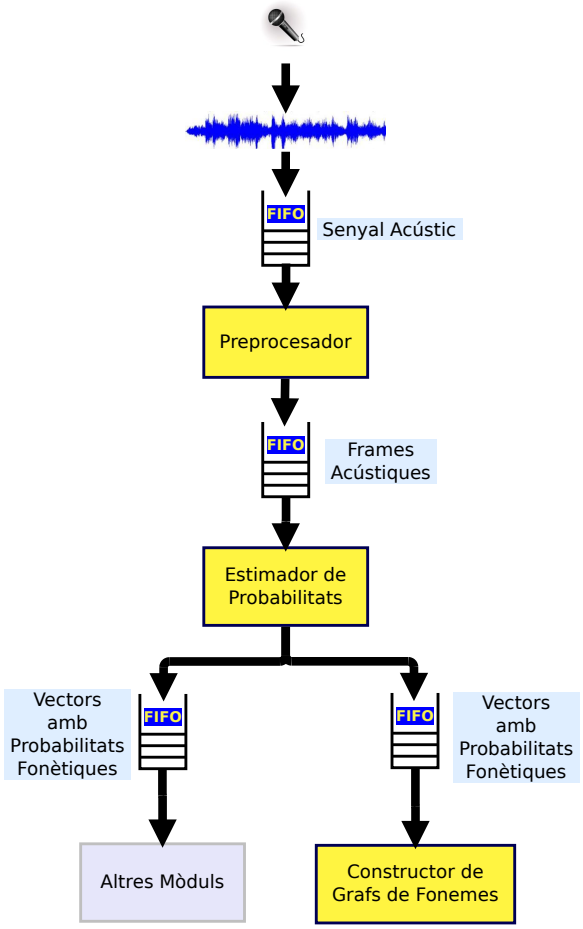


Figura 2.1: Esquema general del sistema.

els distints fils les cues FIFO estan controlades per semàfors.

El senyal d'àudio provinent del micròfon és inserit en una cua FIFO que és llegida pel Parametritzador, que s'encarrega del preprocés. Aquest mòdul segmenta el senyal d'entrada i emet un vector acústic cada cert interval de temps, típicament cada 10 ms, i els insereix en la següent cua FIFO a mesura que va generant-los. Aquestos vectors acústics es coneixen com a *frames* en el context de l'ASR i contenen la informació discriminant que s'utilitzarà en la tasca de classificació. Els vectors tenen dimensió 39, on el primer element és l'energia i els 12 següents són els 12 primers coeficients cepstrals. Els 26 elements restants són les primeres i segones derivades dels 13 primers ja descrits.

El mòdul que obté les *frames* acústiques inserides en la cua FIFO és l'EPF.

Partint d'aquestos *frames*, l'EPF genera com a eixida una sèrie de vectors de probabilitats de dimensió igual al nombre d'unitats fonètiques utilitzades per a la tasca de reconeixement. Donat un d'aquests vectors, cadascun dels seus components representa la probabilitat de què s'haja pronunciat un fonema concret. Els vectors de probabilitats fonètiques són inserits en diverses cues FIFO corresponents a distints mòduls del sistema. Un d'aquestos mòduls és el Constructor de Grafs de Fonemes o CGF. Encara que no es veu el detall al diagrama de la Figura 2.1, l'EPF emet dos tipus de vectors, uns amb probabilitats fonètiques (que són probabilitats *a posteriori*), i uns altres amb densitats de probabilitat. Els primers són la normalització dels segons. Aquesta informació dual arriba a tots els mòduls que utilitzen l'eixida de l'EPF, incloent el CGF, que només utilitzarà les probabilitats *a posteriori*. Pot semblar redundant, però a partir de l'EPF ja no cal normalitzar, el programador de cada mòdul connectat a l'eixida d'aquest podrà decidir si fer ús de les probabilitats *a posteriori* o de les densitats.

2.2 Grafs de Fonemes

El CGF utilitza els vectors de probabilitats fonètiques, obtinguts de la cua associada a l'eixida de l'EPF. A la vegada, cadascun de dits vectors correspon amb les *frames* acústiques generades pel Parametritzador, el qual realitza una segmentació del senyal d'àudio inicial, tal i com es mostra a la Figura 2.2.

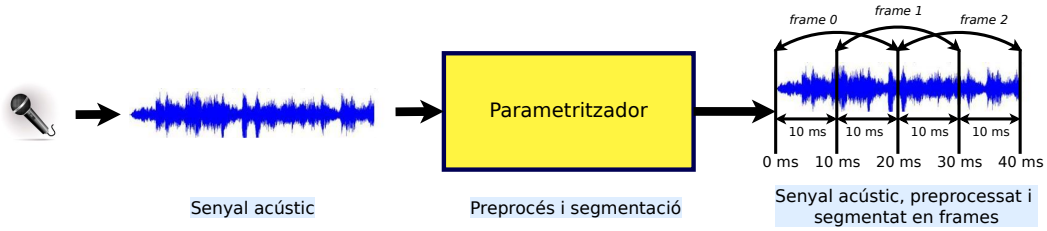


Figura 2.2: Segmentació del senyal d'àudio en *frames* acústics.

Com que el Parametritzador genera una *frame* acústica cada 10 ms, cadascun dels vectors de probabilitats fonètiques representa la probabilitat de què cada fonema s'haja pronunciat en la seua *frame* associada. No obstant, la pronunciació d'un fonema té una durada superior a 10 ms, i per tant requereix més d'una *frame*.

Cal recordar que el mòdul ha de retornar una estructura de dades en la qual ha d'aparèixer especificat en quin instant de temps comença la pronunciació de cada fonema, en quin instant de temps acaba i de quin fonema es tracta. L'estructura idònia per a guardar aquest tipus de dades és un graf.

La informació mantinguda pels nodes i els arcs apareix resumida a la Taula 2.1:

Taula 2.1: Informació gestionada pels nodes i els arcs.

Nodes	***Arcs***
• Llista d'arcs entrants	• Node origen
• Llista d'arcs ixents	• Node destí
• Número de *frame*	• Unitat fonètica
• Instant de temps (segons)	• Probabilitat fonètica (*a posteriori*)

Els nodes actuen com a marcadors de temps. Cada node té associat el número de *frame*, l'instant de temps associat a dita *frame* i dues llistes d'arcs, una per als arcs entrants i l'altra per als ixents. Cada vegada que es detecte l'inici o el final de la pronunciació d'una unitat fonètica, es crea un nou node. No obstant, l'interval de temps en què una unitat fonètica és detectada sol comprendre més d'un arc. És a dir, donada una unitat fonètica que comença en un node i acaba en un altre posterior, el CGF pot haver inserit altres nodes intermedis entre aquestos dos.

Pel que fa als arcs, la informació més rellevant és la unitat fonètica associada i la seua probabilitat. A més, els arcs també guarden un node origen i un node destí. Com que cada node guarda un instant de temps, amb aquesta informació és possible calcular la durada de la pronunciació de la seua unitat fonètica, ja que es sap quan comença pel node origen i quan acaba pel node destí.

En la Figura 2.3 es mostra la topologia que tenen els grafs de fonemes obtinguts pel CGF. Com es pot observar, és una topologia *Left-To-Right* i per tant els grafs són dirigits i acíclics. Tots els arcs que ixen de cada node transiten únicament al següent.

Un node no pot tenir un arc amb el qual transite a si mateix. Si tal cosa ocorreguera, significaria que la pronunciació d'un fonema comença en un instant de temps donat i acaba en eixe mateix instant. Aquest raonament està faltat de sentit, ja que un fonema no pot ser pronunciat en un interval de temps nul.

Els grafs de fonemes són una estructura de dades que resulta adequada per a aquest tipus de tasca, ja que és possiblement l'estructura mínima necessària

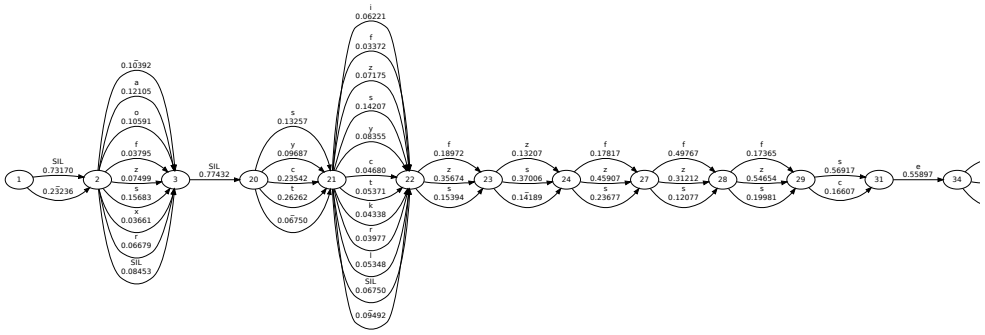


Figura 2.3: Primers 340 ms d'un graf de fonemes, extrets d'un exemple real.

per a representar aquest tipus d'informació. Apart de resultar molt còmoda de llegir, al programador li serà més senzill treballar amb grafs que amb vectors de probabilitats.

En la següent secció s'explica l'algorisme de construcció de grafs de fonemes. D'ara endavant i per simplicitat, cada vegada que s'utilitze el terme *frame* no ens referirem a les *frames* acústiques generades pel Parametritzador, sinó als vectors de probabilitats generats pel l'EPF que estan associades a dites *frames*.

2.3 Algorisme de Construcció de Grafs

2.3.1 Introducció a l'Algorisme de Construcció

Com s'ha comentat anteriorment, el Constructor de Grafs de Fonemes o CGF utilitza una tècnica basada en llindars de probabilitat que mantenen actives un conjunt d'hipòtesis. Més concretament utilitza dos llindars: el llindar de detecció, utilitzat per a la detecció dels fonemes, i el llindar d'extensió, usat per a determinar l'instant en què comença o acaba la pronunciació d'un fonema.

La informació més rellevant gestionada pel CGF és la següent:

- Llindar de detecció.
- Llindar d'extensió.
- Conjunt d'hipòtesis.
- *Buffer* de *frames*.

- *Buffer* de segments detectats.

Pel que fa al conjunt d'hipòtesis, hi ha tantes hipòtesis com unitats fonètiques considerades en la tasca de reconeixement. Cada hipòtesi s'activa quan es detecta l'inici de la pronunciació del seu fonema, i es desactiva al final de la mateixa. Gràcies a les hipòtesis, el CGF és capaç de crear els nodes a mesura que s'analitzen les *frames* entrants. Els nodes representen l'instant en què comença o acaba la detecció de la pronunciació d'un fonema, i aquesta és precisament la informació proporcionada per les hipòtesis.

Cal tenir present que tot aquest procés es realitza de forma *on-line*, a mesura que el CGF va rebent les *frames* proporcionades per l'Estimador de Probabilitats Fonètiques (EPF). Per tant, no comptem amb totes les *frames* des del principi, sinó que són processades a mesura que van arribant al mòdul.

El *buffer* de *frames* és un *buffer* circular que té com a finalitat guardar les 10 últimes *frames* que arriben al CGF. Com que el Parametritzador emet una *frame* cada 10 ms, el *buffer* és capaç d'emmagatzemar els últims 100 ms d'àudio.

Pel que fa al *buffer* de segments detectats, és una llista isomorfa al *buffer* de *frames*. També serà capaç d'emmagatzemar 10 elements, on cadascun d'ells correspon amb els 10 elements del *buffer* de *frames*. Cada component d'aquest últim *buffer* és una llista de booleans, la dimensió de la qual és el nombre d'unitats fonètiques usades en el reconeixement. Aquest *buffer* s'utilitza per a indicar si s'ha detectat o no la pronunciació de cada unitat fonètica en cadascuna de les *frames* mantingudes pel *buffer* circular.

2.3.2 Funcionament de l'Algorisme

L'algorisme de construcció de grafs està dividit en tres fases, com es mostra a la figura 2.4

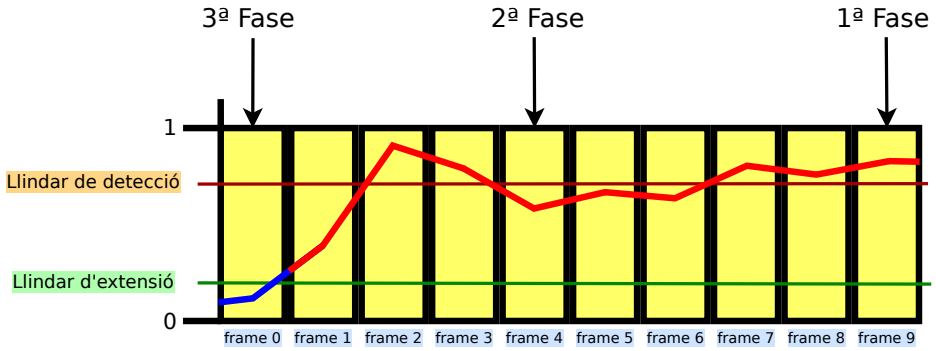


Figura 2.4: Fases de l'algorisme.

1ª Fase - Detecció i extensió enrere. Es comprova si en l'última *frame* entrant es detecta la pronunciació de qualsevol unitat fonètica i es tracta de determinar on comença i acaba dita pronunciació.

2ª Fase - Recuperació de *frames* i eliminació els aïllades. És possible que durant l'etapa de preprocés apareguen problemes, com per exemple que no es puga filtrar adequadament el soroll del senyal acústic. Açò pot portar a què en una *frame* de manera aïllada la probabilitat d'un fonema supere el llindar de detecció, però en les *frames* del voltant no es supere el llindar d'extensió. Aquesta *frame* deu ser eliminada. Anàlogament pot passar el contrari. En eixe cas, cal recuperar la *frame* en què no es supera el llindar d'extensió.

3ª Fase - Activació de les hipòtesis i actualització del graf. Una vegada corregits els errors i restaurades les *frames*, cal analitzar el contingut del *buffer* per a activar o desactivar les hipòtesis pertinents, les quals serviran per a construir el graf de fonemes.

Abans de començar a explicar l'algorisme en profunditat, és necessari conèixer la naturalesa dels fonemes africats i oclusius sords. Aquestos són els corresponents a les grafies '*p*', '*k*', '*t*' i '*ch*' de la llengua espanyola. Per a la pronunciació d'aquestos fonemes, el locutor acumula aire durant uns mil·lisegons en algun dels seus òrgans fonadors. Acte seguit solta l'aire de colp, alliberant tota l'energia acumulada i donant lloc a la part audible de la pronunciació del fonema.

El temps en què el locutor està acumulant aire es veu reflectit en els vectors de probabilitats com un silenci, al qual anomenarem silenci preoclusiu. Aquestos fonemes van a rebre un tractament una mica diferent a la

resta, doncs els silenci preoclusiu no és realment un silenci, sinó part de la pronunciació del fonema i per tant ha de ser tractat com a tal.

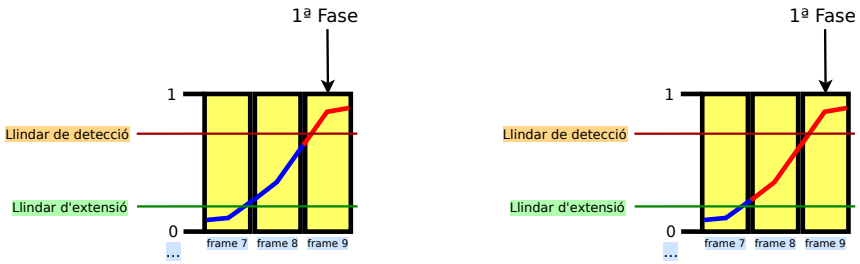
A continuació s'explica amb més detall cadascuna de les tres fases de l'algorisme de construcció.

1ª Fase - Detecció i extensió enrere

Quan arriba una nova *frame*, s'insereix al *buffer* de *frames* i s'analitzen les probabilitats de cadascuna de les unitats fonètiques contingudes en aquesta. En cas que alguna de les probabilitats supere el llindar de detecció, es marca la posició corresponent al *buffer* de segments com a detectada.

El fet que s'haja detectat la pronunciació d'un fonema no és suficient, doncs sabem en quin instant de temps ha superat el llindar de detecció, però no sabem en quin moment ha començat realment la seua pronunciació.

Per a resoldre aquest problema, es compta amb el llindar d'extensió. Quan es detecta un fonema, comencen a ser analitzades de nou aquelles *frames* immediatament anteriors. Si en alguna d'aquestes s'ha superat el llindar d'extensió per a la unitat fonètica detectada, s'assumeix que dita unitat ja havia començat la seua pronunciació abans. Per tant, per a eixes *frames* caldrà indicar al *buffer* de segments que la unitat fonètica ha sigut detectada.



(a) Abans de l'extensió (b) Després de l'extensió

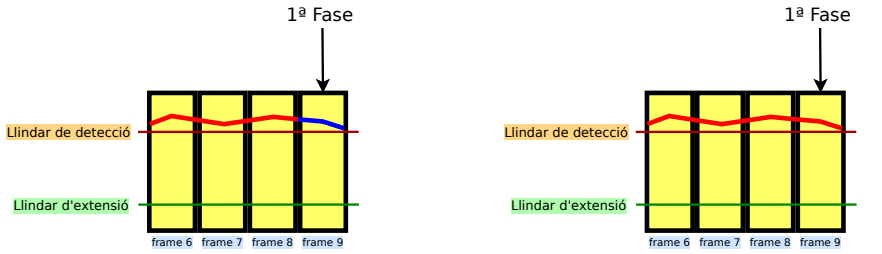
Figura 2.5: Extensió enrere.

En la Figura 2.5 es pot observar el *buffer* de *frames* per a una unitat fonètica qualsevol. Les línies blaves i roges representen la probabilitat de què eixe fonema haja sigut pronunciat en la *frame* on es troba. Les *frames* on aquesta línia apareix en roig representen aquelles on s'ha detectat la unitat fonètica, mentre que les que apareixen en blau indiquen el contrari.

En la Figura 2.5 assumim que l'última *frame* inserida al *buffer* és la novena. Com que la probabilitat fonètica supera el llindar de detecció, ha de

marcar-se com a detectada. Després s'estén la detecció cap enrere, mentre les anteriors *frames* superen el llindar d'extensió. En el cas mostrat en la Figura 2.5, després de marcar la novena *frame* com a detectada, s'analitza la vuitena. Com que aquesta supera el llindar d'extensió, també cal marcar-la com a detectada. A continuació s'analitza la sèptima, i ja que aquesta no supera el llindar d'extensió, el procés d'extensió enrere acaba.

El mòdul deixarà d'estendre la detecció quan una *frame* deixe de superar el llindar d'extensió o quan ja haja sigut estesa (Figura 2.6). Si ja ha sigut estesa amb anterioritat, significa que a la *frame* anterior a la que està sent analitzada ja havia superat el llindar de detecció, i per tant ja s'havia detectat la pronunciació del fonema. Conseqüentment, totes les *frames* anteriors on s'ha d'estendre la detecció ja han sigut marcades com a detectades, i no cal tornar a marcar-les.

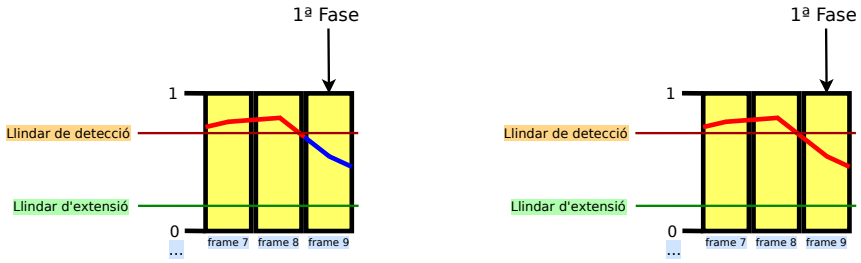


(a) Abans de l'extensió (b) Després de l'extensió

Figura 2.6: Extensió, cas en què no cal estendre enrere.

Hi ha una altra situació que cal tenir en compte, la qual apareix reflectida en la Figura 2.7. A mesura que la pronunciació d'un fonema s'allargue a través de les *frames*, arribarà un moment on es supere el llindar d'extensió, però no el de detecció. Com que la pronunciació del fonema encara no ha acabat, és necessari marcar com a detectades les *frames* que es vegen en aquesta situació.

Per a realitzar aquesta tasca, senzillament es comprova si la probabilitat de dita unitat supera el llindar d'extensió, però no el de detecció. Si es dóna aquest cas, allò que cal fer és examinar la *frame* anterior i comprovar si està marcada com a detectada. En cas que ho estiga, també caldrà marcar l'última *frame* entrant com a detectada, per a eixa unitat fonètica. Aquest cas es sol donar quan la pronunciació d'un fonema està acabant, ja que és quan la seua probabilitat comença a reduir-se de tal forma que no supera el llindar de detecció.



(a) Abans de l'extensió (b) Després de l'extensió

Figura 2.7: Extensió, cas en què no es supera el llindar de detecció.

Els fonemes amb silenci preoclusiu van a requerir un tractament diferent pel que fa a l'extensió. Quan un fonema oclusiu sord o africat supera el llindar de detecció, s'examinen les *frames* anteriors per a la l'extensió. Per a aquelles *frames* anteriors en què el silenci supere el llindar d'extensió, s'estén la detecció per a la unitat fonètica considerada, ja que el silenci preoclusiu també forma part de la pronunciació de dit fonema.

2ª Fase - Recuperació de *frames* i eliminació de les aïllades

Una vegada el *buffer* de *frames* ha sigut omplert i les unitats fonètiques corresponents han sigut detectades, cal corregir aquelles *frames* que puguen contindre errors.

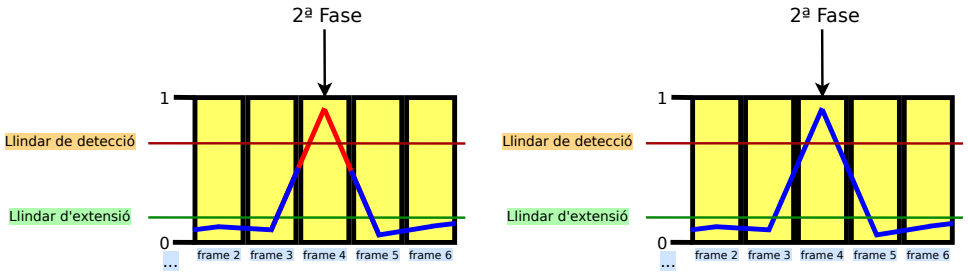
En primer lloc, considerarem que tenen errors aquelles *frames* que es vegen en alguna de les tres situacions següents:

(1) Si la *frame* ha sigut marcada com a detectada i les del voltant no ho estan.

(2) Si la *frame* no està marcada com a detectada i les del voltant sí ho estan.

(3) Si en la *frame* no s'ha detectat cap pronunciació de cap unitat, sabent que el discurs encara no ha acabat.

Per a corregir qualsevol de les tres situacions especificades, cal examinar la *frame* ubicada al mig del *buffer*. Les dues primeres situacions poden ser provocades per esdeveniments espuris amb els que no s'ha pogut tractar durant l'etapa del preprocés, mentre que la tercera pot donar-se degut a què

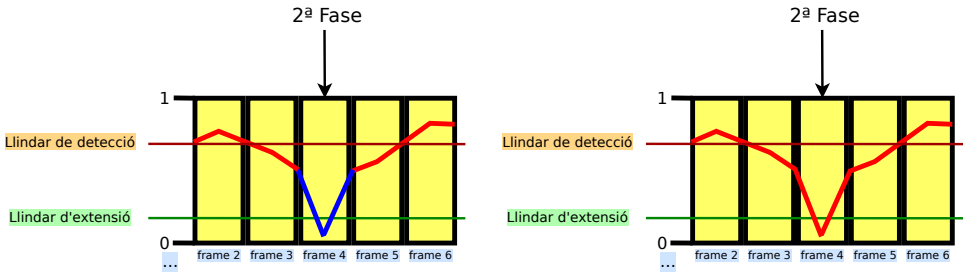


(a) Cas (1), abans de la correcció (b) Cas (1), després de la correcció

Figura 2.8: Eliminació de les *frames* aïllades.

hi haja molta confusió entre les probabilitats de cada fonema, de tal forma que no se'n detecte cap.

Pel que fa a la situació (1), mostrada a la Figura 2.8, la considerem errònia perquè no és possible que la pronunciació d'un fonema comprenga només una *frame*. És a dir, si una unitat fonètica es detecta per a una *frame*, per a poder considerar que dita unitat ha sigut pronunciada és necessari que alguna de les dues *frames* contigües supere el llindar d'extensió. En conseqüència, quan es detecta una *frame* aïllada amb aquestes característiques cal corregir-la marcant-la com a no detectada al *buffer* de segments.



(a) Cas (2), abans de la correcció (b) Cas (2), després de la correcció

Figura 2.9: Recuperació de *frames*.

La situació (2), mostrada a la Figura 2.9, és la contraria a la primera. Si s'està pronunciant un fonema, no té sentit que aquest fonema deixe de ser pronunciat durant una *frame* i que de seguida reprenga la seua pronunciació.

Si es dóna aquesta situació, per a corregir-la cal marcar la *frame* afectada com a detectada al *buffer* de segments.

En quant a la situació (3), tampoc té sentit que no es detecte cap pronunciació en una *frame* quan el discurs encara no ha acabat[1]. Si aquest és el cas, per a corregir la *frame* afectada es marquen com a detectades les mateixes unitats que s'han detectat en la *frame* anterior.

Com que tot aquest procés s'aplica sempre respecte a l'última *frame* inserida, la que està ubicada al mig no serà la mateixa en cada iteració de l'algorisme, sinó que va canviant a mesura que es van inserint noves *frames*. Per tant, es garanteix que totes les *frames* van a ser comprovades i corregides, si ho requereixen.

3ª Fase - Activació de les hipòtesis i actualització del graf

Una vegada detectades les unitats fonètiques i corregits els errors associats a la detecció, ja es disposa de la informació necessària per a procedir amb l'activació de les hipòtesis i la construcció del graf.

Per a activar les hipòtesis el primer que fem és examinar únicament la *frame* que es troba al final del *buffer*, respecte a l'última *frame* entrant. Quan analitzem aquesta *frame*, poden donar-se tres casos:

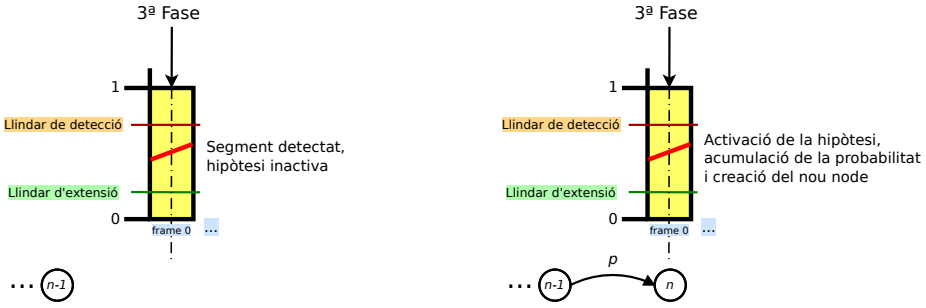
Taula 2.2: Possibles casos donat un punt d'anàlisi.

Cas	***Unitat fonètica (detectada)***	***Hipòtesi (activa)***	***Significat***
(a)	Sí	Inactiva	Inici d'una pronunciació
(b)	No	Activa	Final d'una pronunciació
(c)	Sí	Activa	Pronunciació en curs

A continuació analitzem per separat els tres possibles casos:

(a) En cas que s'haja detectat alguna unitat fonètica en la *frame*, si la seua hipòtesis no estava ja marcada com a activa, significa que acaba de començar la pronunciació del fonema associat. Si la pronunciació del fonema hagués començat abans, la hipòtesis ja estaria activada. Com que en aquest punt s'ha detectat l'inici d'una pronunciació, cal activar la seua hipòtesis, crear un nou node i connectar-lo amb l'anterior. En la Figura 2.10 es mostra aquesta situació.

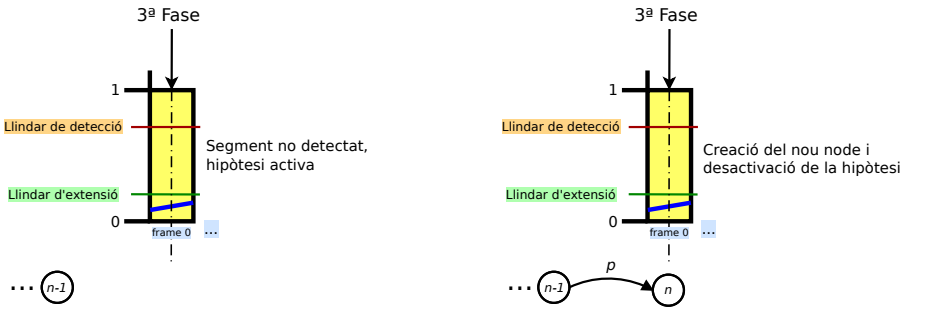
[1]El silenci i les pauses es consideren unitats fonètiques en aquest sistema.



(a) Abans de crear el nou node (b) Després de crear el nou node

Figura 2.10: Inici de la pronunciació.

(b) Aquest cas és el contrari a l'anterior. Si no es detecta la pronunciació d'un fonema però la seua hipòtesis està activa, significa que prèviament estava sent pronunciat i ha deixat de ser-ho. Per tant, ens trobem al final de la pronunciació d'un fonema. Quan es dóna aquesta situació, cal desactivar la hipòtesi corresponent i crear un arc i un node. L'arc tindrà com a node destinació l'últim creat i com a node origen l'anterior, el qual estarà associat a la *frame* en què s'ha detectat l'inici de la pronunciació. En la *frame* que està sent analitzada ja no es detecta la pronunciació, tal i com es mostra en la Figura 2.11. Per tant, l'últim node creat (el node destinació) no estarà associat a eixa *frame* sinó a l'anterior, que és realment fins on arriba la detecció.



(a) Abans de crear el nou node (b) Després de crear el nou node

Figura 2.11: Final de la pronunciació.

(c) Aquesta situació indica que s'ha detectat la pronunciació d'un fonema, però la seua hipòtesis ja estava activada. Per tant, no és ni l'inici ni el final d'una pronunciació, sinó que aquesta està en curs. En aquest cas no hi ha que crear cap node o arc. L'únic que cal fer és acumular les probabilitats, tal i com s'explica posteriorment.

Fins ací s'ha explicat com funciona la creació de nodes i arcs en funció de les hipòtesis. No obstant, no s'han explicat altres aspectes importants, com ara l'acumulació de les probabilitats als arcs o com acaba l'algorisme.

Pel que fa a la acumulació de les probabilitats, quan es crea una nova hipòtesi també se li passa la probabilitat del fonema associat. A mesura que es van analitzant les següents *frames* i aquesta hipòtesi continua activa, la probabilitat de les noves *frames* és acumulada en la hipòtesi. Aquesta probabilitat acumulada serà la que se li associe a l'arc quan siga creat al final de pronunciació.

De nou, els fonemes oclusius sords i africats van a tindre un tractament una mica diferent. En lloc d'acumular directament la probabilitat del fonema, es decideix acumular la màxima entre la probabilitat d'eixe fonema i la del silenci. Quan s'acumula la probabilitat d'un d'aquestos fonemes en la part que conté el silenci preoclusiu, la probabilitat de dit fonema serà baixa, mentre que la del silenci serà alta. Com que el silenci també forma part de la pronunciació, realment cal acumular eixa probabilitat i considerar que forma part de la unitat fonètica que està sent pronunciada.

Pel que fa a la terminació de l'algorisme, quan per part de l'EPF no arriben més *frames*, al CGF se li passen vectors de probabilitats nuls amb la finalitat d'indicar que no s'ha pronunciat cap unitat fonètica i que per tant ha acabat la pronunciació. Caldrà passar-li aquestes *frames* fins que el CGF acabe d'analitzar les que li queden pendents en el *buffer*. Quan això ocorre, el CGF comença a processar les *frames* on no s'ha pronunciat res, i com que no hi ha cap nova pronunciació, l'algorisme acaba de construir el graf de fonemes i finalitza la seua execució.

2.3.3 Variant de l'Algorisme: Llindars Relatius

Tal i com s'ha explicat l'algorisme fins ara, s'entén que el valor dels llindars és el mateix per a totes les *frames*. No obstant, aquesta aproximació pot resultar una mica pobra en determinades situacions.

Per exemple, quan totes les probabilitats fonètiques són molt baixes, és possible que una unitat fonètica no supere el llindar de detecció (o fins i tot el d'extensió) quan en realitat sí que deuria detectar-se. Aquesta situació

tendeix a donar-se entre les pronunciacions de dos fonemes, és a dir, quan la pronunciació d'un fonema està acabant o quan la del següent està començant.

Per tal d'afinar els resultats quan es done aquesta situació, s'ha optat per modificar el valor dels llindars de tal forma que no siguen constants, sinó relatius a la probabilitat màxima de la *frame* que està sent processada. És a dir, donada una *frame* i sent P_{max} la probabilitat més alta d'entre totes les unitats fonètiques per a eixa *frame*, els valors dels llindars es calculen de la següent manera:

Llindar de detecció:

$$Llindar_{detecció} = P_{max} \cdot LD$$

Llindar d'extensió:

$$Llindar_{extensió} = P_{max} \cdot LE$$

On LD i LE són dos coeficients per a la detecció i l'extensió, respectivament, que caldrà ajustar de forma empírica. Amb aquesta tècnica s'espera poder tractar de forma més eficient amb el problema descrit prèviament, ja que amb els nous valors dels llindars es seguiran detectant aquelles unitats fonètiques que ho necessiten, encara que la seua probabilitat siga baixa.

Capítol 3

Experimentació i Resultats

3.1 Descripció de l'Experimentació

L'experimentació portada a terme té dues finalitats: d'una banda l'avaluació dels grafs de fonemes obtinguts, i d'altra la utilització d'aquestos per a la Descodificació Acústic-Fonètica (DAF).

Amb l'objectiu d'ajustar els valors dels llindars per a que es generen els grafs amb la millor qualitat, es realitza en primer lloc una sèrie d'experiments en els quals es tracta de determinar quins són els millors valors per als llindars. Hi ha dos criteris a tenir en compte a l'hora d'analitzar els resultats obtinguts: en quina mesura els grafs generats contenen allò que s'ha pronunciat i la densitat d'aquestos.

Una vegada fixats els valors dels llindars i generats els grafs de millor qualitat, es pot aplicar sobre aquestos un algorisme per a DAF. L'algorisme empra dos tipus de models: Models de Llenguatge i Models de Durada Fonètica. Per tal de millorar els resultats, aquestos models seran utilitzats conjuntament amb la Penalització per Inserció de Fonemes. Les tècniques nomenades van a requerir l'ajustament empíric d'una sèrie de factors que tenen com a finalitat ponderar la importància de la informació que proporcionen aquestes.

Tota l'experimentació es duu a terme utilitzant els dos tipus de llindars descrits fins al moment: aquells que es mantenen constants per a totes les *frames*, denominats **llindars absoluts**, i aquells que varien en funció de la probabilitat màxima continguda en cada *frame*, els **llindars relatius**, descrits en la secció 2.3.3.

En quant a la mesura de l'error, s'empraran les mètriques descrites en la secció 3.3.

3.1.1 Ajustament dels Llindars

Els primers experiments que cal realitzar tenen com a finalitat determinar quins són els millors valors per als llindars de detecció i extensió. Aquestos valors seran (en principi) aquells amb què es puguen generar uns grafs en els quals estiga continguda, en la major mesura possible, la seqüència fonètica original o seqüència de referencia. Per tal d'obtenir la similitud entre les seqüències contingudes al graf i la de referència, es fa ús d'un algorisme tipus *Dynamic Time Warping* o DTW.

Quan més xicotets siguen els llindars, menys restrictiu serà l'algorisme a l'hora de detectar unitats, activar hipòtesis i generar arcs i nodes. En altres paraules, els grafs resultants seran més densos, amb un major nombre de nodes i arcs. Si els grafs resulten ser massa densos, els posteriors mòduls del sistema que hagen de treballar amb ells realitzaran les seues tasques emprant un major temps de còmput. És per això que no resulta suficient elegir aquells valors que presenten l'error més reduït per a l'experimentació posterior, sinó que a més cal tenir en compte la densitat dels grafs resultants.

3.1.2 Models de Llenguatge

Una vegada s'han determinat els valors dels llindars que generen els grafs de millor qualitat, es poden utilitzar Models de Llenguatge (ML) per a millorar els resultats de DAF.

Normalment els ML s'utilitzen a nivell de paraula, i s'entrenen per tal que siguen capaços d'estimar la probabilitat de que una seqüència de n paraules es manifeste. Dites seqüències s'anomenen n-grames i típicament es solen utilitzar seqüències de dues o tres paraules (bi-grames o tri-grames).

En el present projecte, els n-grames estimats per l'ML no són seqüències de paraules, com en els sistemes d'ASR convencionals, sinó de fonemes. És a dir, cada fonema utilitzat en la tasca de reconeixement es considera com una paraula a nivell de model de llenguatge, i la probabilitat que els n-grames tenen associada indica la probabilitat de què dita seqüència fonètica es manifeste.

El model de llenguatge s'utilitza en l'algorisme de DAF sobre els grafs per combinar les probabilitats dels n-grames amb les probabilitats acústic-fonètiques que hi ha als arcs del graf. D'aquesta manera la probabilitat d'una hipòtesi en un moment determinat és fruit de la combinació de dues fonts de informació en forma de probabilitat: la proporcionada per l'ML en relació a les seqüències fonètiques i la proporcionada pels arcs del graf en relació a quan probable és que cada fonema haja sigut pronunciat donat un segment de veu.

Quin pes se li dóna a cada font d'informació es decideix amb experimentació per tal d'ajustar el que es coneix com *Grammar Scale Factor* (GSF). És habitual que aquest factor s'aplique a les probabilitats de l'ML per tal de minvar (si < 1.0) o ressaltar (si > 1.0) el seu pes front a les probabilitats acústic-fonètiques.

3.1.3 Models de Durada Fonètica

A més d'utilitzar ML, també es tractarà de millorar els resultats amb l'ús de Models de Durada Fonètica o MDF.

La construcció dels MDF es realitza mitjançant tècniques estadístiques. La idea bàsica consisteix en la implementació d'un model que, donada una unitat fonètica i la durada que presenta, calculen la probabilitat de que dita unitat fonètica tinga realment eixa durada.

Si resulta que una unitat fonètica es manifesta durant massa temps (o massa poc), l'MDF indicarà una baixa probabilitat per a dita unitat. Això permet penalitzar les probabilitats d'aquelles unitats fonètiques detectades que no presenten una durada adequada d'acord amb l'MDF, per tal que s'ajusten més a la realitat.

Aquests tipus de models es construeixen com s'explica a continuació. Donada una sèrie de discursos on es coneixen tots els fonemes pronunciats i la seua durada, es crea un histograma per cada unitat fonètica. En aquest es registra la distribució de la freqüència amb què dita unitat fonètica presenta cadascuna de les duracions trobades en els discursos.

No obstant aquesta informació no resulta del tot útil, ja que interessa obtenir una probabilitat que es calcule per a una unitat fonètica en funció de la seua durada. Per tal d'aconseguir-ho, es divideix cadascuna de les freqüències registrades entre la suma de totes elles per a un mateix fonema. Amb aquest càlcul, els valors en l'histograma per a cada durada estaran compresos entre 0 i 1, i poden ser utilitzats com a probabilitats.

Una vegada construït aquest model, la manera d'aplicar-lo als grafs de fonemes resulta prou senzilla. Les probabilitats proporcionades per l'MDF són emprades en l'algorisme de DAF d'una manera semblant a com s'apliquen les probabilitats de l'ML. En aquest algorisme estes probabilitats tenen com a finalitat, tal qual s'ha explicat abans, penalitzar que en una hipòtesi s'assigne un segment de veu massa curt o massa llarg a un fonema.

De manera anàloga al GSF, també s'ajustarà un factor, anomenat MDF_{factor}, per donar-li més o menys pes a les probabilitats de l'MDF.

3.1.4 Penalització per Inserció de Fonemes

Conjuntament amb els ML i els MDF, s'utilitzarà la Penalització d'Inserció de Fonemes o *Phoneme Insertion Penalty* ($PhIP$).

El $PhIP$ és una penalització (o reducció de la probabilitat d'una hipòtesi) que s'aplica cada vegada que es canvia d'unitat fonètica. Aquest és anàleg al *Word Insertion Penalty* (WIP) que s'aplica en els algorismes de descodificació en reconeixement a nivell de paraula.

Per tal d'entendre com pot el $PhIP$ millorar els resultats, considerem el graf mostrat en la Figura 3.1.

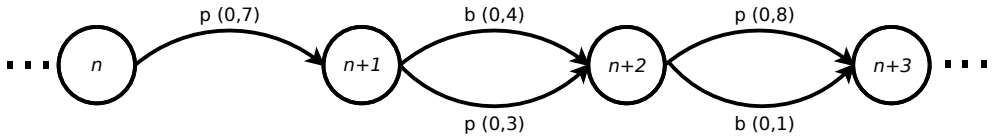


Figura 3.1: Exemple artificial d'un fragment de grafs de fonemes.

S'observa que el graf té quatre nodes connectats per diversos arcs. En aquestos apareix la unitat fonètica associada ('*p*' i '*b*') i la probabilitat acumulada per a cada arc.

Si es realitza la descodificació obtenint simplement la seqüència de major probabilitat i considerant únicament les probabilitats acústic-fonètiques acumulades als arcs, obtenim que dita seqüència és '*pbp*', que resulta impossible de pronunciar. No obstant, si apliquem el $PhIP$ la probabilitat de l'arc amb unitat fonètica b ubicat entre els nodes $n+1$ i $n+2$ queda penalitzada, sent així la '*p*' la més probable per a eixe arc. Per tant, la seqüència més probable passa a ser '*ppp*', transcrita com a '*p*'[1] , la qual segurament serà més encertada.

En conclusió, amb aquesta aproximació es pretén no canviar d'hipòtesi de forma prematura durant l'aplicació de l'algorisme de DAF, amb la qual cosa s'espera obtenir millors resultats.

3.2 Algorisme de DAF

L'algorisme emprat per a la descodificació tracta d'obtenir la millor seqüència fonètica mitjançant l'exploració dels grafs. Dita exploració es realitza de forma iterativa generant i mantenint una sèrie d'hipòtesis, entenent com a

[1]Cal recordar que la detecció d'un fonema pot aparèixer en diversos arcs consecutius.

hipòtesi una seqüència fonètica que pot haver sigut pronunciada fins el punt d'anàlisi en què es troba l'exploració.

Per a la creació de cada hipòtesi nova, es té en compte la probabilitat acústic-fonètica de l'arc del graf que va a ser processat, la probabilitat proporcionada pel Model de Llenguatge (ML), la proporcionada pel Model de Durada Fonètica (MDF), i la de la hipòtesi a partir de la qual es va a crear la nova.

3.2.1 Sistema de Puntuació

Com que cercar entre totes les possibles seqüències contingudes al graf resulta un problema intractable, cal establir un sistema de puntuació que servisca per a comparar les hipòtesis entre si i ser així capaços d'explorar només aquelles que puguen oferir un millor resultat.

Com s'ha dit abans, l'algorisme té en compte fins a tres fonts d'informació per a la descodificació: les probabilitats contingudes als arcs, les probabilitats de l'ML i les probabilitats de l'MDF. El GSF i l'MDF_{factor} ponderen l'LM i l'MDF respectivament. Cal recordar també que, quan es canvia d'unitat fonètica, s'aplica el $PhIP$ per tal de penalitzar aquest canvi, evitant així canviar prematurament de hipòtesi.

El sistema de puntuació emprat per tal d'establir un criteri de qualitat a l'hora de comparar les hipòtesis està establert sobre aquestos tres factors i les probabilitats fonètiques contingudes als arcs. Atenent a les següents definicions:

- uf_{nova} (uf_i): unitat fonètica continguda en l'arc que s'està explorant en l'algorisme.
- uf_{previa} (uf_{i-1}): última unitat fonètica pronunciada en la hipòtesi anterior.
- $P_r(uf_i|arc)$: probabilitat acústic-fonètica per a la unitat uf_i continguda en l'arc que està sent explorat.
- $P_r(uf_i|uf_{i-k}...uf_{i-1})$: probabilitat de que es pronuncie la unitat fonètica uf_i, sabent que les anteriors k unitats pronunciades han sigut $uf_{i-k}...uf_{i-1}$. Al llarg de l'experimentació, s'utilitza $k = 2$. És a dir, l'ML funciona amb tri-grames: té en compte la unitat fonètica actual i les dues anteriors.
- $P_r(MDF(uf_{previa}, t))$: probabilitat de que uf_{previa} haja acabat la seua pronunciació quan aquesta ha durat un temps $\leq t$.

La puntuació de la hipòtesi actual ($S(H_{nova})$) es calcula com segueix:

$$\begin{aligned} S(H_{nova}) = S(H_{previa}) &+ \log Pr(uf_i|arc) + PhIP \\ &+ GSF \times \log P_r(uf_i|uf_{i-k} \dots uf_{i-1}) \\ &+ MDF_{factor} \times \log P_r(MDF(uf_{i-1}, t)) \end{aligned} \tag{3.1}$$

on $S(H_{previa})$ és la puntuació de la hipòtesi predecessora.

3.2.2 Definició de les Estructures de Dades

Hipòtesis

Una de les estructures de dades o objectes que van a ser utilitzats són les hipòtesis. La informació gestionada per les hipòtesis i les operacions que es poden realitzar sobre elles estan especificades en la definició de la classe mostrada en l'Algorisme 1.

Algorisme 1 Definició de la classe hipòtesi.

class Hipòtesi
public:
Arc **getArc()** ▷ Torna l'arc de la hipòtesi
double **getPuntuacio()** ▷ Torna la puntuació
string **getSequencia()** ▷ Torna la seqüència fonètica associada
int **getDuracio()** ▷ Torna la Durada de l'últim fonema
ML_Node **getMLNode()** ▷ Torna el node en l' ML
int **getHashCode()** ▷ Torna el *hash code* associat

private:
Arc **arc** ▷ Arc de la hipòtesi
double **puntuacio** ▷ Puntuació de la hipòtesi
string **sequencia** ▷ Seqüència fonètica pronunciada
int **duracio_ultima_unitat** ▷ Durada de l'últim fonema
ML_Node **ml_node** ▷ Node en l'ML
int **hash_code** ▷ *Hash code*, obtingut a partir de la sequencia

end class

Tota hipòtesi té associat un **arc** del graf que es correspon amb una unitat fonètica. A més, cada arc conté els nodes origen i destinació, més la probabilitat de que la unitat fonètica haja sigut pronunciada al segment de veu que representa l'arc, és a dir, des de la *frame* del node origen fins a la *frame* anterior al node destinació.

L'arc de cada hipòtesi és l'últim arc del graf que ha sigut examinat per l'algorisme de DAF, per tal d'arribar al node al qual la hipòtesi quedarà fixada. L'arc, per tant, conté la unitat fonètica que en la hipòtesi es considera ha sigut l'última en ser pronunciada. La puntuació de cada hipòtesi es calcula com s'ha especificat en l'apartat anterior.

L'atribut **sequencia** en la definició de la classe és la seqüència fonètica que es considera pronunciada fins al moment de la creació de la hipòtesi. Pel que fa a l'atribut **duracio_ultima_unitat**, aquest representa el temps que persisteix la pronunciació de l'última unitat fonètica de la hipòtesi.

Un altre atribut és el **ml_node**, de manera que junt amb l'atribut "node del graf" referit pel node destinació de l'arc, es té que cada hipòtesi està ubicada en un node del ML i un node del graf. No es permetrà mai dues hipòtesis amb aquests dos atributs iguals. Així, a l'hora d'expandir una hipòtesi, s'utilitzaran els arcs del graf que ixen del node destinació de l'arc associat a la hipòtesi (**arc**), i els nodes de l'ML que són nodes successors al node actual del ML (**ml_node**). Per cada parella possible es crearà una nova hipòtesi, que en afegir-se a la llista d'hipòtesis podrà sobreviure o no segons la seua puntuació.

Per últim falta l'atribut **hash_code**, que és un codi que es calcula a partir de la seqüència junt amb la unitat fonètica de l'arc. Amb aquest atribut es localitza de manera eficient si ja existeix una hipòtesi o no en una taula *hash* amb la mateixa seqüència fonètica. Pot ocórrer que s'obtinga el mateix *hash code* a partir de diferents seqüències fonètiques, en aquest cas s'aplicarà la tècnica corresponent per a resoldre col·lisions dins de la taula *hash*.

Model de Llenguatge, Model de Durada Fonètica, Cua de Prioritat i Taula Hash

En l'algorisme que es descriu en la secció 3.2.3, l'ML i l'MDF fan ús dels següents mètodes:

- MDF:

  `double getLogProbCambiar(Unitat Fonetica, Duracio):` Donada una unitat fonètica i la seua durada, torna la probabilitat de que dita unitat fonètica acabe la seua pronunciació. Quan major siga la durada que presenta la unitat fonètica, major serà la probabilitat de canviar a la següent.

- ML:

  `double getProbNodeSeguent(ML_Node, Unitat Fonetica):` Donat un node dels ML i una unitat fonètica, indica la probabilitat de transitar des del node donat al següent amb dita unitat fonètica.

  `ML_Node seguentNode(ML_Node, Unitat Fonetica):` Donat un node i una unitat fonètica, torna el node al que es transita amb dita unitat fonètica des del node donat.

L'algorisme requereix mantenir dues llistes d'hipòtesis i accedir primer a aquelles que oferisquen la millor puntuació. La cua de prioritat és una bona estructura per a emmagatzemar-les gràcies a que permet mantenir una llista d'hipòtesis ordenada segons la seua puntuació, i a més de forma eficient.

Basat en el *hash code* de cada hipòtesi, l'algorisme fa ús d'una taula *hash* per tal d'evitar que hi haja dues hipòtesis amb la mateixa seqüència fonètica. Quan es crea una nova hipòtesi, aquesta es buscada en la taula *hash*. En el cas de què ja hi haja una la mateixa seqüència fonètica, només sobreviurà aquella que oferisca la millor puntuació. Així s'espera reduir la quantitat d'hipòtesis expandides i reduir el temps de còmput i la memòria emprada per l'algorisme.

3.2.3 Descripció de l'Algorisme

L'algorisme de DAF està descrit l'Algorisme 2.

Algorisme 2 Descodificació Acústic-Fonètica (1)

ENTRADA:
Graf de Fonemes gf ▷ Graf al que s'aplica la descodificació.
Model de Llenguatge ml
Model de Durada Fonètica mdf
int MDF_{factor} ▷ Pes assignat als MDF.
int GSF ▷ Pes assignat als ML.
int PhIP ▷ Penalització per Inserció de Fonemes.

EIXIDA: string millor_sequencia ▷ Seqüència més probable.
Cua_Prioritat hips_actuals ← {}
Cua_Prioritat hips_noves ← {}

Taula_hash taula_hash
hips_actuals.push(creaHipotesi(ml.iniciFrase()))

Algorisme: Descodificació Acústic-Fonètica (2)

```
13: max_hipotesis ← 1000
14: for all node_actual ∈ gf.getNodes() do          ▷ Recorregut dels nodes
15:                                                  ▷ en ordre temporal.
16:     num_hipotesis ← 0
17:     while hips_actuals ≠ {} do
18:         hip_actual ← hips_actuals.top()
19:         hips_actuals.pop()
20:         if num_hipotesis < max_hipotesis then
21:             num_hipotesis ← num_hipotesis + 1
22:             for all arc ∈ node_actual.getArcsEntrants() do
23:                 uf_nova ← arc.getUnitatFonetica()
24:                 uf_previa ← hip_actual.getArc().getUnitatFonetica()
25:                 puntuacio ← hip_actual.getPuntuacio() +
26:                                     log(arc.getProbAcumulada())
27:                 if uf_previa ≠ uf_nova then
28:                     puntuacio + = PhIP
29:                     puntuacio + = MDF_factor ×
30:                             log(mdf.getProbCambiar(uf_previa,
31:                             hip_actual.getDurada()))
32:                     sequencia ← hip_actual.getSequencia() + uf_nova
33:                     prob_ML ← 0.0
34:                     if uf_nova ≠ SIL then
35:                         node_previ ← hip_actual.getMLNode()
36:                         prob_ML ← GSF ×
37:                                     log(ml.getProbSeguentNode(node_previ,
38:                                                                 uf_nova))
39:                         seguent_ml_node ← ml.seguentNode(node_previ,
40:                                                                 uf_nova)
41:                         puntuacio + = prob_ML
42:                     else
43:                         seguent_lm_node ← node_previ
44:                     end if
45:                 end if
```

Algorisme: Descodificació Acústic-Fonètica (3)

```
46:                 nova_hipotesi ← creaHipotesi(hip_actual,
47:                                              arc,
48:                                              puntuacio,
49:                                              sequencia,
50:                                              seguent_ml_node)
51:                 hip_hash = taula_hash.busca(nova_hipotesi)
52:                 if hip_hash = NULL then
53:                     hips_noves.push(nova_hipotesi)
54:                     taula_hash.inserir(nova_hipotesi)
55:                 else if puntuacio > hip_hash.getPuntuacio() then
56:                     taula_hash.substitueix(nova_hipotesi)
57:                     hips_noves.substitueix(nova_hipotesi)
58:                 end if
59:             end for
60:         end if
61:     end while
62:     swap(hips_actuals, hips_noves)
63: end for
64:
65: millor_sequencia ← hips_actuals.top().getSequencia()
66: return millor_sequencia
```

L'algorisme rep com a entrada els següents arguments:

- **gf**, graf de fonemes
- **ml**, Model de Llenguatge
- **mdf**, Model de Durada Fonètica
- MDF_{factor}, per a ponderar l'MDF
- GSF, per a ponderar l'ML
- **PhIP**, Penalització per Inserció de Fonemes

L'eixida d'aquest algorisme és la millor seqüència fonètica trobada, donada per la variable **millor_sequencia**.

En primer lloc es creen les estructures necessàries per a l'algorisme: dues cues de prioritat i una taula *hash*. Les cues de prioritat són **hips_actuals**, que manté les hipòtesis que estan pendents d'explorar, i **hips_noves**, que

guarda les hipòtesis que es generen a partir de les anteriors per a la seua posterior exploració.

Abans de començar l'exploració del graf, s'insereix una hipòtesi que es considera que està ubicada en el node inicial del graf (línia 12) i s'inicialitza a 1000 el nombre màxim d'hipòtesis (**max_hipotesis**). Aquest és el nombre màxim d'hipòtesis que es poden explorar en cada iteració, encara que hi haja més hipòtesis en la cua **hips_actuals**. Açò ajuda a evitar que les cues cresquen indefinidament.

Per a cada node del graf, excepte per a l'inicial, s'obtenen els seus arcs entrants i per a cada hipòtesi continguda en **hips_actuals**, es creen noves hipòtesis amb les unitats fonètiques contingudes als arcs que estan sent explorats.

Primerament, s'extrau la millor hipòtesi de la cua **hips_actuals** i es genera la nova informació que ha de tenir la hipòtesi que es crearà pròximament. La unitat fonètica nova (uf_{nova}) és la que està continguda a l'arc que s'està explorant, i la unitat fonètica anterior (uf_{previa}) és la que s'ha generat i guardat anteriorment en la hipòtesi extreta de la cua, és a dir, **hip_actual**. Pel que fa a la puntuació, en primer lloc s'inicialitza amb la puntuació de la hipòtesi actual i se li suma la probabilitat de l'arc que està sent explorat (línies 25-26).

La resta del codi té com a finalitat modificar aquesta puntuació d'acord amb la informació proporcionada per l'MDF i l'ML. Per a poder aplicar els dos tipus de models és necessari que acabe la pronunciació de la unitat fonètica continguda en la hipòtesi. Si es canvia d'unitat fonètica, aleshores s'aplica la penalització per inserció de fonemes (**PhIP**, línia 28) i seguidament s'utilitza l'MDF (**mdf**, línies 29-31). Aquest rep la unitat fonètica anterior i la durada que ha presentat. Amb aquesta informació, l'MDF indica la probabilitat de que la unitat fonètica anterior acabe la seua pronunciació, probabilitat que és ponderada per l'MDF_{factor} i s'acumula el seu logaritme a la puntuació calculada prèviament.

A continuació i de forma similar, s'aplica l'ML. Com que el silenci no està considerat com una unitat fonètica en l'ML, només avançarem al següent estat de l'ML quan la unitat fonètica nova no siga un silenci. En cas que fóra un silenci, el node de l'ML de la nova hipòtesi és el mateix que en l'anterior. Aleshores, quan no és silenci, es calcula la probabilitat de l'ML (**prob_ML**) i es pondera amb el GSF (línies 36-38). Seguidament s'acumula el logaritme de la probabilitat a la puntuació obtinguda prèviament i es transita d'estat a l'ML.

Arribats a aquest punt, ja es disposa de tota la informació necessària per a crear la nova hipòtesi. Quan ja ha sigut creada es comprova amb la taula *hash* si ja existeix una hipòtesi amb la mateixa seqüència fonètica (línia 52).

Si no és així, s'insereix en la taula *hash* i en la cua de noves hipòtesis. Si pel contrari ja hi ha una hipòtesi amb la mateixa seqüència fonètica, la nova substituirà a l'existent només si millora la seua puntuació (línies 55-57).

Finalment, la cua de les hipòtesis actuals estarà buida i la de les hipòtesis noves contindrà les expandides durant la iteració. Per a que l'algorisme continue, s'intercanvia el contingut de les dos cues.

Quan la part iterativa de l'algorisme ha finalitzat, només queda tornar la seqüència obtinguda per la millor de les hipòtesis que han arribat al final.

3.3 Mesura de l'Error

Les mètriques per a mesurar l'error seran les mateixes per a tots els experiments. Com que es té la seqüència de fonemes correcta de cada pronunciació, la mesura consisteix en comparar dita seqüència amb la millor que s'obté dels grafs.

Aquesta comparació entre les seqüències es duu a terme mitjançant la distància de Levenshtein. Dita distància és la suma dels costos de les operacions d'edició que hi ha que realitzar sobre la seqüència obtinguda per a que siga igual a l'original. Pel que fa a aquestes operacions, la nomenclatura emprada serà la següent:

- N = Nombre d'etiquetes o símbols en les transcripcions de referència.
- I = Nombre d'insercions realitzades.
- B = Nombre d'esborrats realitzats.
- S = Nombre de substitucions realitzades.

Aplicant l'algorisme que calcula la distància de Levenshtein obtenim tots els paràmetres anteriors.

A continuació s'expliquen els coeficients que van a ser utilitzats per a mesurar l'error. Considerant el nombre de substitucions (S) i esborrats (B), la primera mètrica emprada és el *Percent Correct* (PC).

$$PC = \frac{N-B-S}{N} \times 100$$

El PC és un percentatge que indica com de correcta és la seqüència trobada. Aquest es mesura com la proporció entre la diferència de la longitud de la transcripció de referència i el nombre d'esborrats i substitucions requerides respecte a dita longitud.

Una altra mètrica més precisa és el *Percent Accuracy* (PA), definit de la següent manera:

$$PA = \frac{N-B-S-I}{N} \times 100$$

El PC i el PA són quasi equivalents. La única diferència és que l'últim considera també les operacions d'inserció (I), mentre que el primer les ignora.

Ambdues mètriques estan convertint-se en mesures estàndard per a l'avaluació de l'error en el camp de l'ASR. Aquestes són utilitzades (per exemple) per HTK, un *software* àmpliament utilitzat per a la recerca no només en el camp de l'ASR, sinó també en reconeixement de caràcters o seqüenciament de l'ADN.

Per a presentar els resultats de l'experimentació, la mesura emprada serà el PA, per tenir major precisió que el PC.

3.4 Corpus de Veu

El corpus de veu utilitzat és ALBAYZIN [CGL+91, CGL+92], que comprèn tres blocs ben diferenciats, cadascun dels quals respon a un objectiu específic:

a) Un **corpus fonètic** genèric [MPB+93], sense restriccions semàntiques, la finalitat del qual és construir un marc de referència el més ampli possible de la llengua castellana.

b) Un **corpus específic** que depèn d'una aplicació [DRP+93], en la que s'establisquen restriccions semàntiques i sintàctiques a les frases contingudes en ell, amb la finalitat d'aplicar aquestes restriccions en un model realista.

   Aquest corpus correspon a una tasca de consulta a una base de dades sobre geografia espanyola.

c) Un **corpus sorollós**, l'objectiu del qual és verificar el procés de reconeixement en ambients hostils (alt nivell de soroll front a la senyal vocal).

   Aquest corpus està format per gravacions de frases dels altres corpus produïdes amb efecte Lombard.

Els tres corpus foren dissenyats pels sis components del consorci, i les gravacions les dugué a terme l'empresa Page Ibérica, S.A. (PISA) de Madrid.

Característiques dels corpora

A continuació es descriu el contingut concret del corpus fonètic i del geogràfic.

Corpus fonètic

Els disseny del corpus fonètic es realitzà tenint en compte dos criteris bàsics: 1) l'obtenció d'una cobertura dels al·lòfons en una proporció el més pròxima a la de la parla normal, i 2) l'obtenció d'un nombre mínim d'elocucions de cada al·lòfon en cada context. Per això es defineixen els al·lòfons existents en castellà i es consideraren els distints factors contextuals de variació fonètica (entorn fonètic immediat, posició en la síl·laba, grau d'accentuació, etc.), es tingué en compte no únicament la importància quantitativa de cada al·lòfon o factor, sinó també la seua rellevància des del punt de vista fonètic. Com a referència per a la parla normal s'utilitzà la transcripció de tres entrevistes a diferents locutors sobre temes variats i d'una hora de durada cadascuna. Així doncs, encara que els corpus contenen material llegit, el punt de partida del disseny ha sigut la parla espontània.

El corpus ha sigut dividit en dos subcorpus, un d'aprenentatge i un altre de prova, que han sigut dissenyats de forma independent. El primer, que consta de 4.800 frases, servirà per a entrenar els sistemes de reconeixement a nivell acústic-fonètic. El segon, de menor envergadura (2.000 frases), ha sigut concebut principalment per a avaluar dits sistemes a nivell fonètic. Pot utilitzar-se també per a validar criteris de disseny, i per a refinar models acústics entrenats amb el subcorpus d'aprenentatge.

El subcorpus d'aprenentatge es genera mitjançant l'elocució repetida de 200 frases per diversos locutors. El subcorpus de prova s'obté a partir de 500 frases repetides 4 vegades, cadascuna per un locutor. Aquestes 500 frases van ser escollides d'entre 1.000 seleccionades de textos de Manuel Vázquez Montalbán ("*El laberinto griego*" i "*Los mares del sur*") i d'un text de Miguel Delibes ("*Señora de rojo sobre fondo gris*").

Corpus geogràfic

El segon corpus és més específic i és dependent de l'aplicació, estant format per frases corresponents a una tasca de consulta a una base de dades geogràfica. Les frases són sotmeses a una forta restricció semàntica amb la finalitat d'incloure informació relativa a aquest aspecte en el reconeixement o comprensió de la parla. Les construccions sintàctiques reflecteixen la forma natural de la parla en llengua castellana. Per a extraure-les s'han analitzat 14.918 frases obtingudes mitjançant 408 persones en què aquestes intenten obtenir informació sobre la geografia espanyola. Dites persones procedien principalment de les ciutats on estan ubicats dos grups que dugueren a terme el projecte ALBAYZIN, i eren majoritàriament estudiants universitaris. L'anàlisi ha consistit en la revisió ortogràfica de les frases, la quantificació

de quantitats i la normalització de les unitats de mesura, i la classificació de les frases. Algunes frases han sigut rebutjades degut a les incorreccions gramaticals, y la resta han sigut classificades segons criteris lingüístics (tipus de interrogació), semàntics (pertinença a una tasca simplificada -subtasca- o tasca completa), i de complexitat estructural (d'acord amb el nombre de verbs i entitats semàntiques presents, el qual ha requerit la classificació del vocabulari).

Dins d'aquest corpus geogràfic s'ha dissenyat una subtasca de domini semàntic restringit, a la qual pertanyen un total de 500 frases diferents, 300 de les quals estan incloses en el subcorpus d'aprenentatge, les altres 200 pertanyen al subcorpus de prova. Cada una ha sigut pronunciada dues vegades per locutors diferents, disposant de 1.000 pronunciacions (600 d'aprenentatge i 400 de prova) per a la subtasca. El domini semàntic d'aquesta subtasca engloba frases en les que únicament intervenen tres entitats base (mars, rius i comunitats autònomes) amb els seus respectius atributs (nom, extensió, cabal i longitud), i amb les relacions de dependència (naixement, desembocadura, banyada per, etc.).

El corpus està dividit en dos subcorpus, un d'aprenentatge i un de prova, tenint cadascun d'ells una part que correspon a la subtasca. Les frases que formen ambdós subcorpus s'han elegit amb un criteri de màxima distància relativa entre elles. El conjunt de frases conté 1.170 paraules diferents.

Contingut quantitatiu dels corpus

Corpus fonètic

Aprenentatge:	4 locutors pronuncien 200 frases cada un (FA1) 160 locutors pronuncien distints grups de frases fins un total de 4.000 (FA2) En total 4.800 frases
Prova:	40 locutors pronuncien 50 frases cada un d'entre 500 diferents (FP) (2.000 frases)
Total:	6.800 frases, 204 locutors, 6 hores de veu

Corpus geogràfic

Aprenentatge:	88 locutors pronuncien 50 frases (GA) (4.400 frases)
Prova:	48 (els 4 de FA1, més 32 de FA2 més 12 nous) pronuncien 50 frases (GP) (2.400 frases)
Total:	6.800 frases, 136 locutors, 6 hores i 40 minuts de veu

El nombre total de locutors distints és 304, 152 de cada sexe. Tots ells van

ser elegits de les comunitats de Castella la Manxa, Castella Lleó i Cantàbria, per a poder obtenir una representació dialectal del castellà en la seua varietat central. El nombre total de frases és 15.600, i la durada en terme mitjà de cada frase és de 4 segons.

Característiques de les gravacions

En tots els aspectes del projecte ALBAYZIN es seguiren els estàndards proposats en el projecte europeu *Speech Assessment Methods* (SAM) del programa ESPRIT.

L'equipament utilitzat fou:

- Càmera condicionada acústicament
- Micròfon Shure SM 10A amb fixació al cap
- Placa d'adquisició OROS AU-22, compatible amb l'estàndard SAM
- Dispositiu compatible amb disc òptic multifunció de SONY
- DAT per a la gravació acústica de l'evolució de cadascuna de les sessions de gravació
- Software EUROPEC de ESPRIT-SAM

La freqüència de mostreig és 16 kHz, el senyal és quantificat amb un convertidor A/D de 16 bits. Tots els senyals contenen 200 ms de so ambiental al principi i al final. Els senyals foren filtrats pas alt per eliminar el soroll a baixes freqüències induït en el procés de gravació. El filtre es dissenyà per a cancel·lar 40 dB d'una component freqüencial entorn als 16 Hz, i per a deixar passar, quasi sense atenuació (0,1 dB), les components superiors a 150 Hz.

A més de les gravacions, per a cada frase es disposa de la seua transcripció ortogràfica. Es disposa d'un subconjunt de 1.000 frases del corpus fonètic etiquetat i segmentat manualment.

Les taules 3.1 i 3.2 mostren les freqüències d'aparició de cada un dels fonemes utilitzats en el sistema per al corpus fonètic i el corpus geogràfic de la base de dades ALBAYZIN.

Taula 3.1: *Freqüència relativa d'aparició (f) dels fonemes en castellà [Roj91], i en el corpus **fonètic** de la base de dades ALBAYZIN. Per al corpus fonètic es mostren les freqüències relatives i absolutes (F) en tot el corpus, en el subcorpus d'aprenentatge i en el subcorpus de proves.*

Fonema	Castellà	Total		Aprenentatge		Proves	
	f	*f*	*F*	*f*	*F*	*f*	*F*
p	2,66	2,35	6288	2,28	4056	2,49	2232
t	4,48	4.29	11504	4,34	7728	4,21	3776
k	3,98	3,82	10224	3,87	6888	3,72	3336
b	2,66	3,14	8412	3,18	5664	3,06	2748
d	4,79	4,35	11656	4,11	7320	4,83	4336
g	0,95	1,33	3556	1,41	2520	1,16	1036
m	3,09	3,55	9504	3,66	6528	3,32	2976
n	6,99	7,32	19596	7,38	13152	7,18	6444
h	0,19	0,47	1268	0,57	1008	0,29	260
f	0,68	0,55	1472	0,54	960	0,57	512
z	1,68	1,56	4192	1,52	2712	1,65	1480
s	7,58	7,26	19460	7,48	13320	6,85	6140
x	0,73	0,66	1764	0,61	1080	0,76	684
y	0,22	0,50	1336	0,59	1056	0,31	280
c	0,28	0,52	1380	0,61	1080	0,33	300
l	5,08	4,62	12372	4,50	8016	4,86	4356
L	0,38	0,56	1488	0,59	1056	0,48	432
r	5,67	5,25	14072	5,21	9288	5,33	4784
@	0,79	0,59	1584	0,50	888	0,78	696
i	7,50	6,88	18416	6,78	12072	7,07	6344
e	13,51	12,93	34644	12,70	22632	13,39	12012
a	13,40	13,90	37240	13,93	24816	13,85	12424
o	9,57	9,81	26280	9,81	17472	9,82	8808
u	3,16	3,79	10156	3,85	6864	3,67	3292

Taula 3.2: *Freqüència relativa d'aparició (f) dels fonemes en castellà [Roj91], i en el corpus* ***geogràfic*** *de la base de dades ALBAYZIN. Per al corpus geogràfic es mostren les freqüències relatives i absolutes (F) en tot el corpus, en el subcorpus d'aprenentatge i en el subcorpus de proves.*

Fonema	Castellà	Total		Aprenentatge		Proves	
	f	*f*	*F*	*f*	*F*	*f*	*F*
p	2,66	1,87	5453	1,86	3515	1,87	1938
t	4,48	3,95	11548	3,96	7468	3,95	4080
k	3,98	5,33	15570	5,33	10056	5,33	5514
b	2,66	1,95	5684	1,93	3648	1,97	2036
d	4,79	6,64	19399	6,70	12643	6,53	6756
g	0,95	0,69	2005	0,66	1247	0,73	758
m	3,09	5,07	14794	5,05	9526	5,09	5268
n	6,99	7,38	21549	7,36	13881	7,42	7668
h	0,19	0,47	1363	0,46	859	0,49	504
f	0,68	0,13	389	0,14	255	0,13	134
z	1,68	1,27	3707	1,28	2421	1,24	1286
s	7,58	7,68	22436	7,70	14518	7,66	7918
x	0,73	0,26	755	0,26	481	0,26	274
y	0,22	0,28	807	0,27	511	0,29	296
c	0,28	0,07	206	0,06	120	0,08	86
l	5,08	5,85	17085	5,84	11015	5,87	6070
L	0,38	0,21	617	0,21	405	0,21	212
r	5,67	3,62	10562	3,62	6822	3,62	3740
@	0,79	1,25	3656	1,23	2328	1,28	1328
i	7,50	6,95	20296	6,95	13106	6,95	7190
e	13,51	12,52	36553	12,49	23557	12,57	12996
a	13,40	14,05	41030	14,12	26636	13,92	14394
o	9,57	8,71	25435	8,66	16327	8,81	9108
u	3,16	3,80	11106	3,85	7258	3,72	3848

3.5 Disseny dels Experiments

Algorisme de DAF sobre Vectors de Probabilitats

En els sistemes d'ASR convencionals és habitual aplicar un algorisme per a DAF directament sobre els vectors de probabilitats generats per l'Estimador de Probabilitats Fonètiques (EPF), el qual calcula la probabilitat de que cada unitat fonètica haja sigut pronunciada en cada *frame* acústica. En canvi, en aquest projecte dits vectors són l'entrada d'un nou mòdul, el Constructor de Grafs de Fonemes o CGF, que genera els grafs de fonemes sobre els quals s'aplica la descodificació.

Amb l'objectiu de comparar les tècniques convencionals amb les desenvolupades al llarg d'aquest projecte, un dels experiments consisteix en aplicar un algorisme de DAF directament sobre els vectors de probabilitats generats per l'EPF. Els resultats obtinguts amb aquesta tècnica es comparen amb els que s'obtenen mitjançant l'algorisme de DAF aplicat sobre els grafs de fonemes.

Ajustament dels Llindars

El primer conjunt d'experiments té com a finalitat ajustar els valors dels llindars per tal que el CGF genere els millors grafs possibles. En quant als llindars, cal recordar que l'algorisme pot emprar dues tècniques: llindars absoluts i llindars relatius. Per tal de determinar quin dels dos ofereix un millor resultat, es duu a terme l'experimentació tant per a llindars absoluts com per a llindars relatius. Els valors assignats als llindars per a l'experimentació estan dins dels rangs mostrats en la Taula 3.3.

Taula 3.3: Rangs dels valors per als distints tipus de llindar.

Tipus de llindar	***Llindar de detecció***	***Llindar d'extensió***
Llindars Absoluts	$[0.05, 0.15]$	$[0.005, 0.05]$
Llindars Relatius	$[0.05, 0.6]$	$[0.003, 0.6]$

Per a comprovar la qualitat dels grafs es tenen en compte dos aspectes. En primer lloc, en quina mesura els grafs generats contenen allò que s'ha pronunciat. Per tal de fer aquesta mesura, es compara la seqüència de referència amb el contingut dels grafs aplicant un algorisme de tipus DTW, utilitzant les mètriques explicades en la secció 3.3. L'altre aspecte a considerar, com s'ha explicat anteriorment, és la densitat que presenten els grafs

resultants. Per a avaluar aquesta densitat, es tenen en compte els quatre factors explicats a continuació.

Si N és el nombre de nodes que hi ha al graf, A és el nombre total d'arcs continguts al graf, T és la durada del discurs en segons i U és el nombre d'unitats fonètiques considerades, definim els següents factors:

- Nombre mitjà de nodes per segon (NPS):

$$NPS = \frac{N}{T}$$

- Nombre mitjà d'arcs que ixen de cada node, conegut com a factor de ramificació o *Branching Factor* (BF):

$$BF = \frac{A}{N}$$

- Proporció de Densitat Màxima (PDM), definit com la relació entre producte dels nodes per segon i el factor de ramificació, i el nombre d'unitats fonètiques per 100. El nombre màxim de nodes que podria haver és d'un per *frame* (100). En quant als arcs, el nombre màxim que se'n poden generar per node és d'un per cada unitat fonètica (U). Aquesta fórmula és utilitzada per a avaluar la densitat del graf obtingut respecte la densitat màxima que podria presentar.

$$PDM = \frac{NPS \cdot BF}{100 \cdot U}$$

Després d'analitzar els grafs resultants amb aquestes tècniques, cal elegir els valors dels llindars que oferisquen un millor resultat, sense perdre de vista la densitat dels grafs obtinguts.

Seguidament, es tornen a generar els grafs de fonemes amb els valors escollits (tant per a llindars absoluts com relatius) i sobre aquests últims es duu a terme la resta de l'experimentació utilitzant les tècniques explicades en les seccions anteriors.

Algorisme de DAF sobre els Grafs

Sobre els darrers grafs generats, s'aplica l'algorisme de DAF ja explicat. Des del punt de vista de l'experimentació, l'algorisme rep tres paràmetres importants el valor dels quals varia en cada experiment. Aquestos paràmetres són el GSF, l'MDF_{factor} i el $PhIP$. Com ja s'ha explicat anteriorment, els dos primers són els factors encarregats de ponderar la informació proporcionada

per l'ML i l'MDF respectivament. El tercer, que serveix per a evitar canviar prematurament d'hipòtesi, s'aplica conjuntament amb els altres dos.

Aquesta part de l'experimentació té com a finalitat l'estudi de l'efecte que tenen els tres factors ja explicats en el procés de descodificació, tant conjuntament com per separat.

Els valors assignats als tres factors són els següents:

- $GSF \in [0, 4]$
- $MDF_{factor} \in [0, 8]$
- $PhIP \in [0, -7]$

Quan el $PhIP$ pren valor 0, no s'aplica cap penalització per inserció de fonemes. Si qualsevol dels altres dos pren valor 0, significa que la informació proporcionada pel model que ponderen no té cap efecte en la descodificació.

De cara a la experimentació, en primer lloc cal determinar quins són els millors valors per als tres factors. Amb aquesta finalitat s'aplica l'algorisme de DAF per a 200 frases del corpus d'entrenament. Una vegada determinada la millor combinació per als tres factors, es repeteix de nou l'experimentació per a distints valors dels llindars, tant absoluts com relatius i amb el corpus d'entrenament complet.

La Figura 3.2 resumeix el procés d'experimentació.

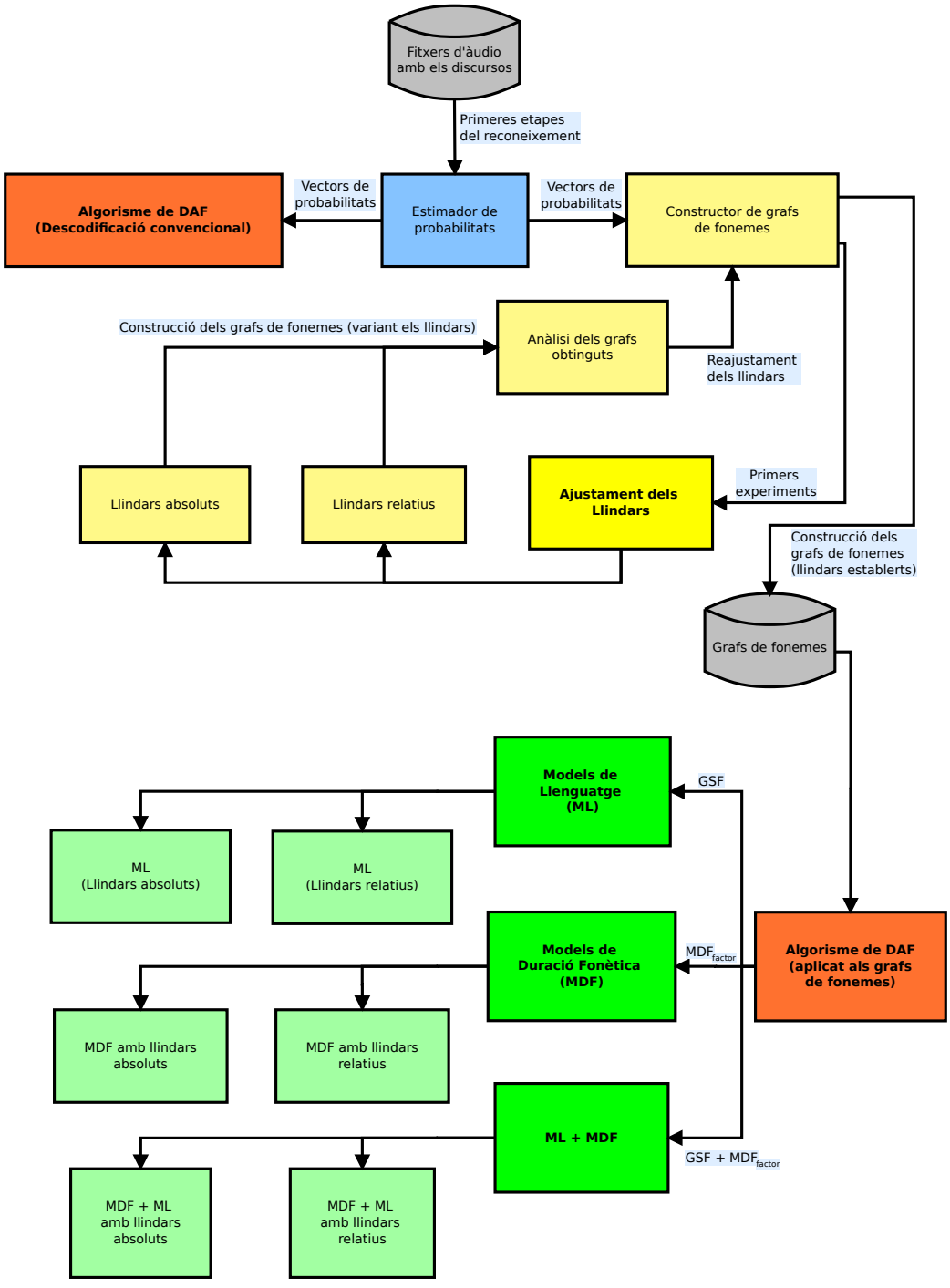


Figura 3.2: Disseny de l'experimentació.

3.6 Resultats

3.6.1 Ajustament dels Llindars

Les gràfiques de les Figures 3.3 i 3.4 mostren com varia el PA en funció dels llindars de detecció i extensió, tant per a llindars absoluts com relatius. Recordem que els llindars absoluts són aquells en què els valors dels llindars són fixats i es mantenen constants per a totes les frames, mentre els llindars relatius varien per a cada *frame* en funció de la seua probabilitat màxima.

Cal tenir en compte que l'objectiu d'aquest experiment és comprovar en quina mesura els grafs de fonemes contenen la seqüència fonètica de referència. Amb aquesta finalitat, s'aplicarà l'algorisme *Dynamic Time Warping* (DTW) per tal de buscar la seqüència continguda en els grafs que més es parega a la de referència.

De les Figures 3.3 i 3.4 es pot deduir que el PA és molt més sensible a les variacions del llindar d'extensió que a les del de detecció. Aquest comportament es manifesta tant per als llindars absoluts com per als relatius. A més, també s'observa que el valor màxim del PA s'obté quan més reduïts són els llindars, arribant a valer un 95.46% per als llindars absoluts i un 97.51% per als llindars relatius. És a dir, quan els llindars prenen aquestos valors els grafs contenen el 95,46% i 97.51% de la seqüència de referència, respectivament.

Això és deu a què quan menys restrictius són els llindars més elevada és la quantitat d'unitats fonètiques detectades. Això dóna lloc a que l'algorisme cree una major quantitat d'arcs i nodes, i per aquest fet la possibilitat de què la seqüència fonètica de referència estiga continguda als grafs és molt més elevada. De fet, si els llindars prengueren els valors més xicotets possibles (lleugerament superiors a 0), de cada node eixiria un arc per cada unitat fonètica i el PA seria del 100%. No obstant, aquest cas no seria gens convenient perquè totes les unitats sempre serien detectades, i conseqüentment els grafs no eliminarien informació inútil per a la Descodificació Acústic-Fonètica (DAF).

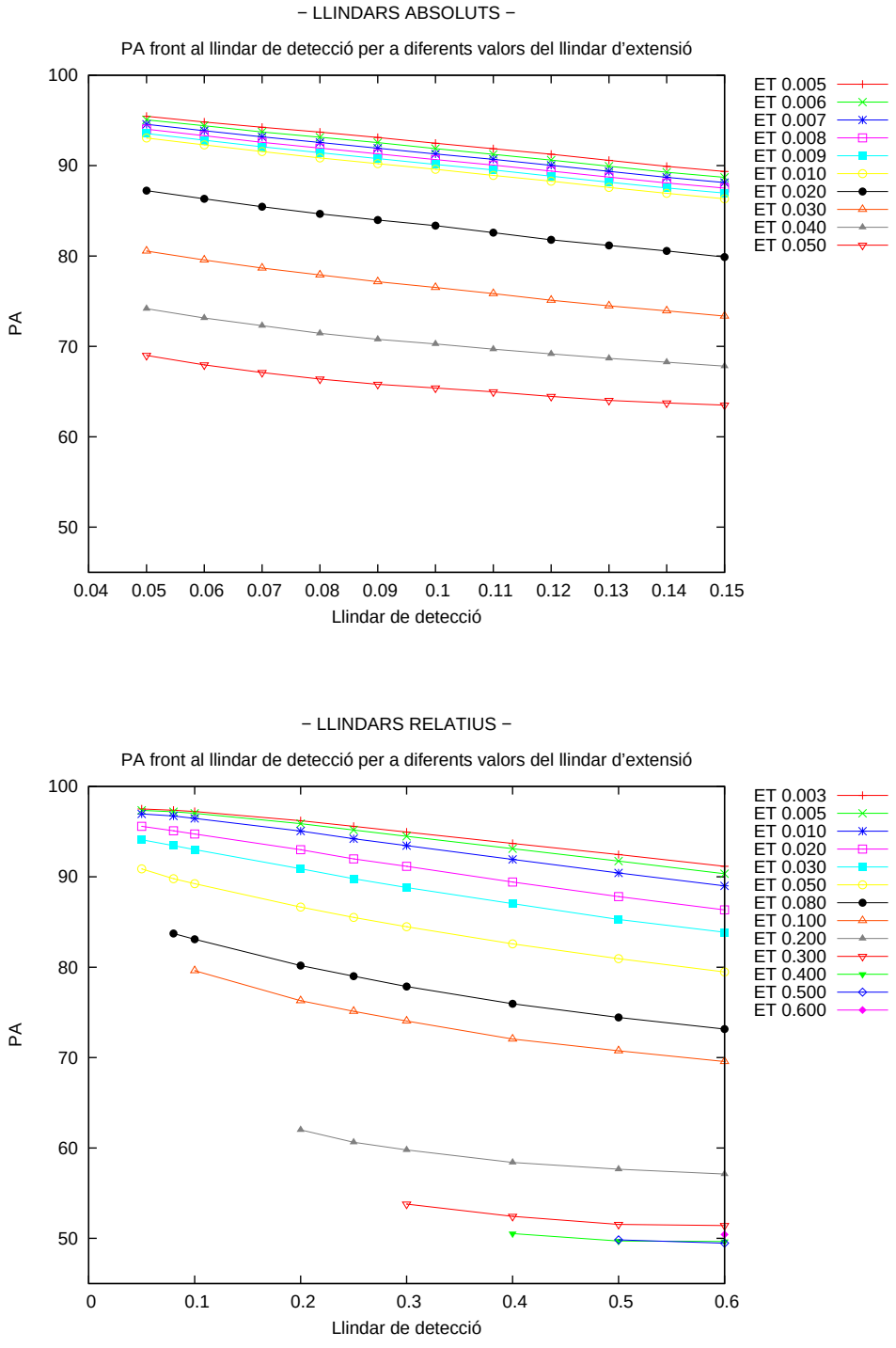


Figura 3.3: PA en funció del llindar de detecció, per als dos tipus de llindars.

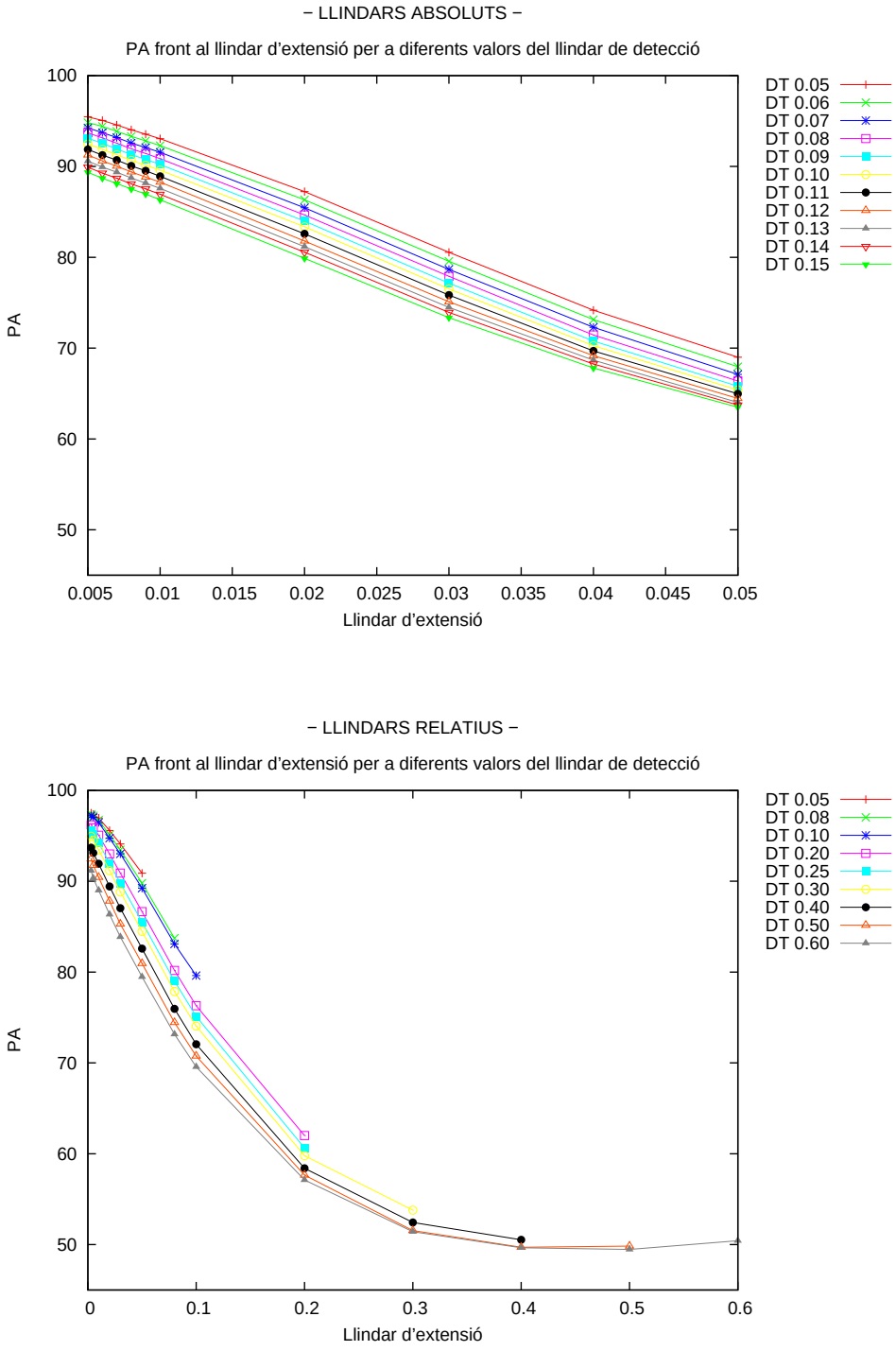


Figura 3.4: PA en funció del llindar d'extensió, per als dos tipus de llindars.

Per tant, resulta convenient dur a terme un estudi que pose de manifest la densitat dels grafs en funció dels llindars. Per a la realització d'aquest estudi, es consideraran els tres factors ja explicats: la quantitat mitjana de Nodes Per Segon (NPS), el Factor de Ramificació (BF) i la Proporció de Densitat Màxima (PDM). Els resultats d'aquest estudi estan recollits a les Figures 3.5, 3.6, 3.7, 3.8, 3.9 i 3.10.

A les Figures 3.5 i 3.6 es pot observar que l'NPS decreix a mesura que augmenten els valors dels llindars, i pareix ser que la seua variació és molt més sensible a les modificacions del llindar d'extensió que a les del de detecció. En aquest sentit, s'observa una evolució similar en el BF, tal i com mostren les Figures 3.7 i 3.8.

Per tal d'entendre aquest comportament, en primer lloc cal recordar que l'algorisme crea un nou node cada vegada que es detecta l'inici o el final d'una pronunciació. Com que la detecció de cada unitat fonètica ha de tindre un inici i un final, quantes més unitats fonètiques es detecten major serà la quantitat de nodes creats per l'algorisme. Però el llindar de detecció no té tant de pes a l'hora de generar arcs i nodes com el llindar d'extensió, tal i com mostren les Figures 3.5, 3.6, 3.7 i 3.8. Si es redueix el llindar d'extensió, la detecció de les unitats fonètiques tendeix a començar en un instant de temps anterior i a acabar en un posterior. Si entre aquestos instants de temps existeixen altres nodes, la detecció també perdurarà durant eixes *frames* i en conseqüència també es crearan més arcs que connecten aquestos altres nodes intermedis. I si a aquest efecte li sumem el fet de què la quantitat de nodes també augmenta a mesura que es redueix el llindar, el resultat és un augment més pronunciat de la quantitat d'arcs creats.

Pel que fa al PDM (Figures 3.9 i 3.10), es pot vore que són idèntiques a les gràfiques que mostren l'evolució del BF amb la diferència de què el PDM està comprés entre 0 i 1. També s'observa que els grafs més densos generats amb llindars absoluts presenten un PDM lleugerament superior al 22%. És a dir, l'algorisme és capaç de generar uns grafs que contenen el 95.46% de la seqüència de referència i sense afegir un 78% de la informació (arcs i nodes) que resulta innecessària. En el cas dels llindars relatius els grafs contenen el 97.51% de la seqüència de referència, amb un PDM del 36% aproximadament.

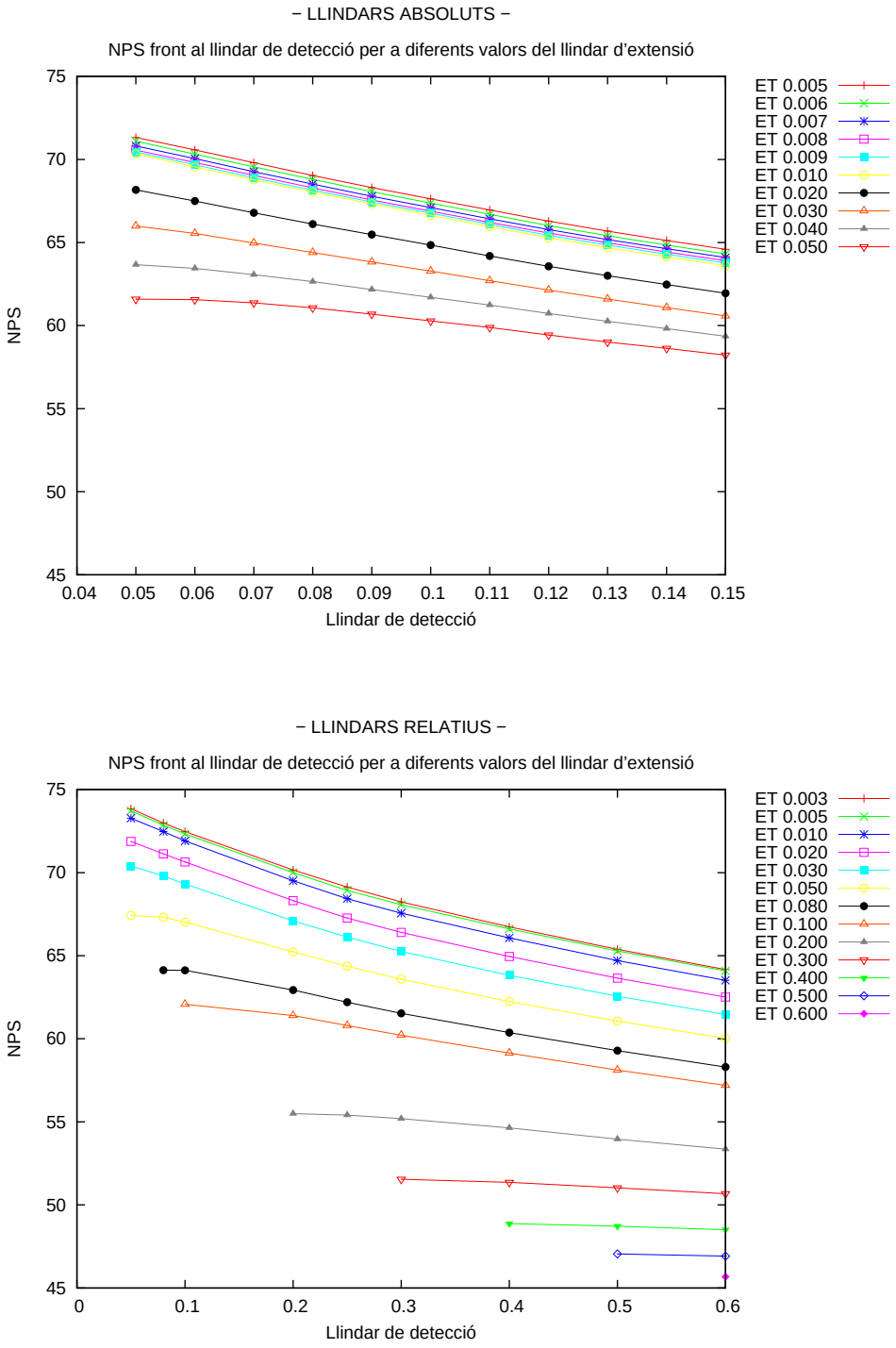


Figura 3.5: NPS en funció del llindar de detecció, per als dos tipus de llindar.

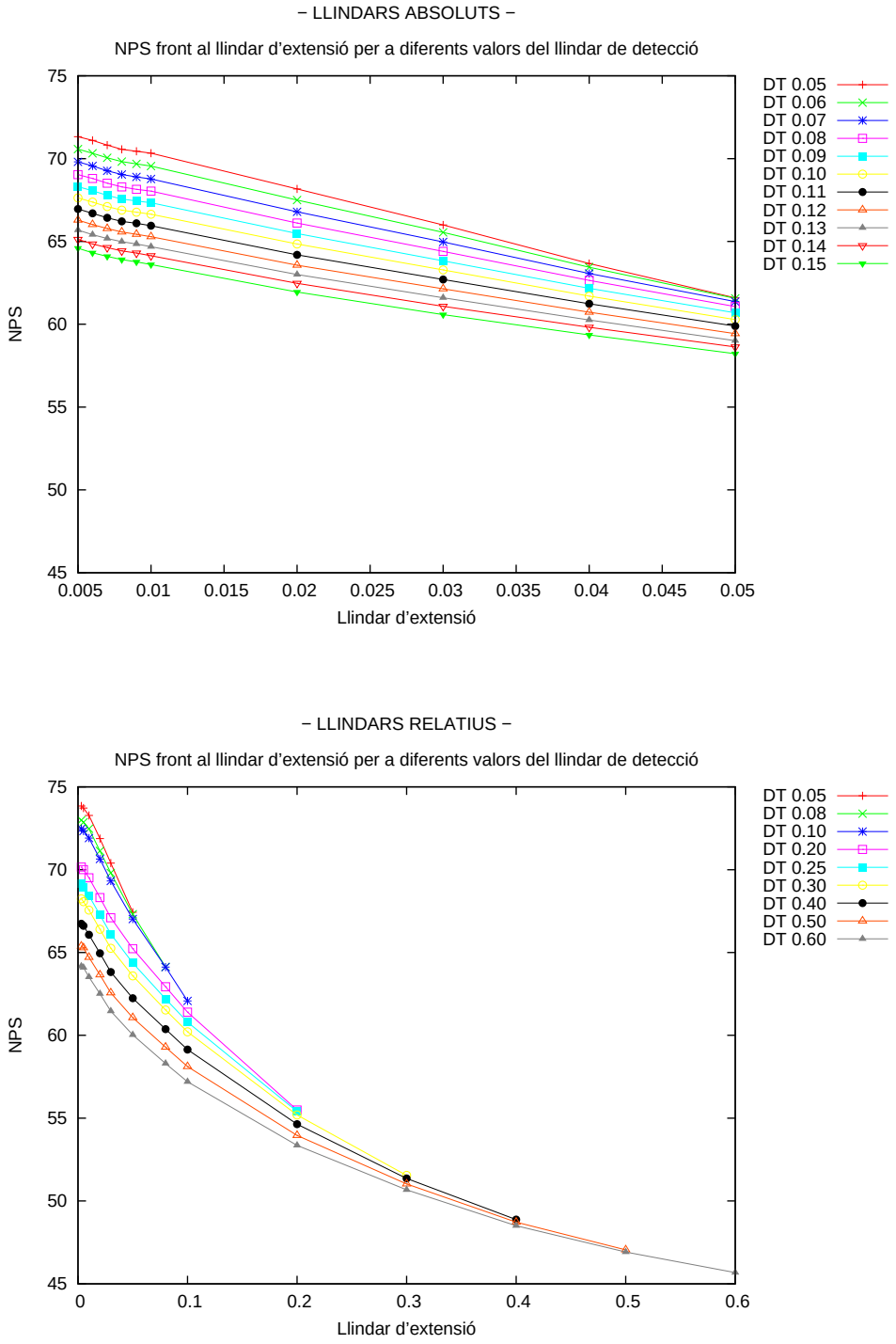


Figura 3.6: NPS en funció del llindar d'extensió, per als dos tipus de llindar.

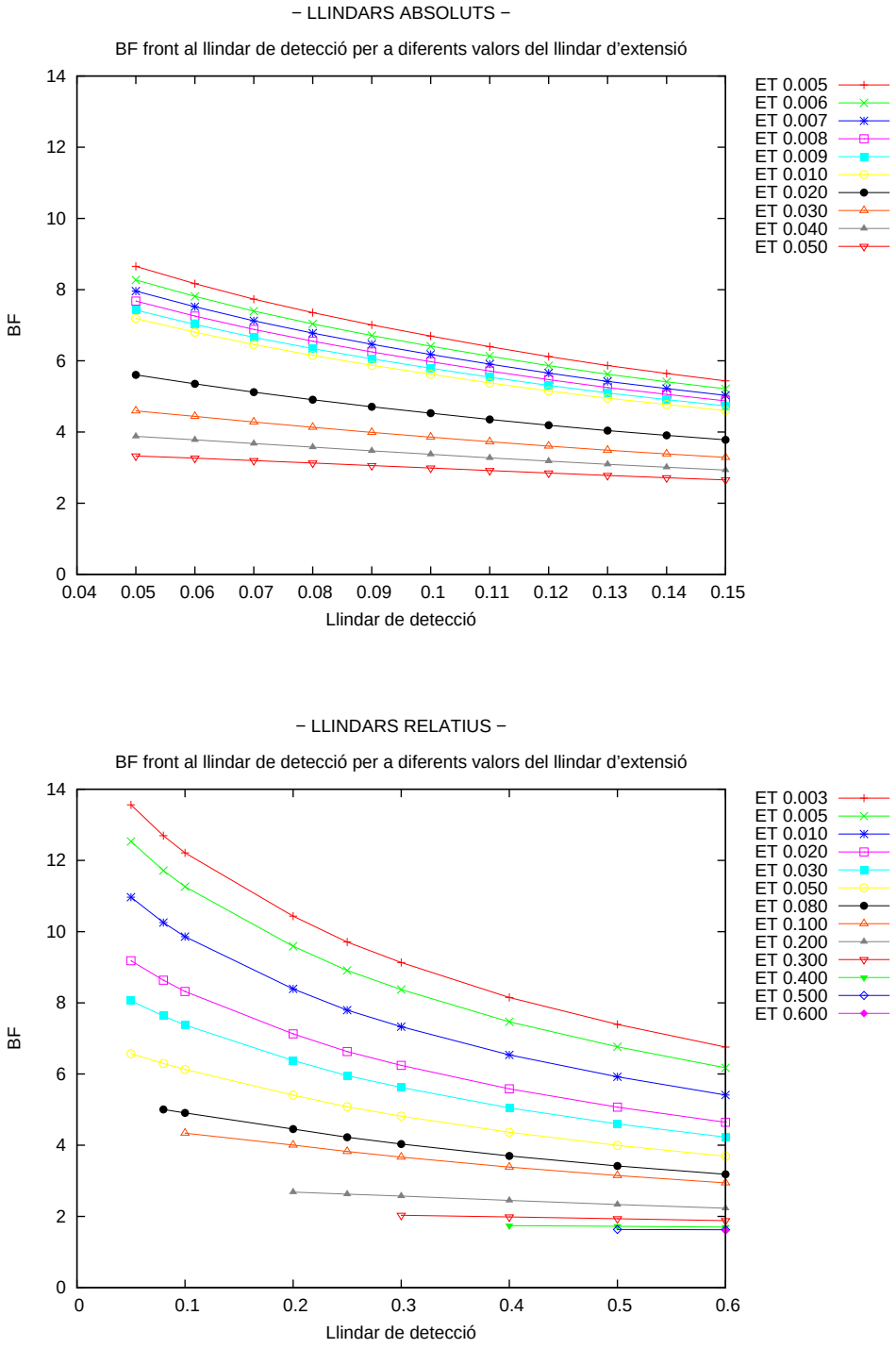


Figura 3.7: BF en funció del llindar de detecció, per als dos tipus de llindar.

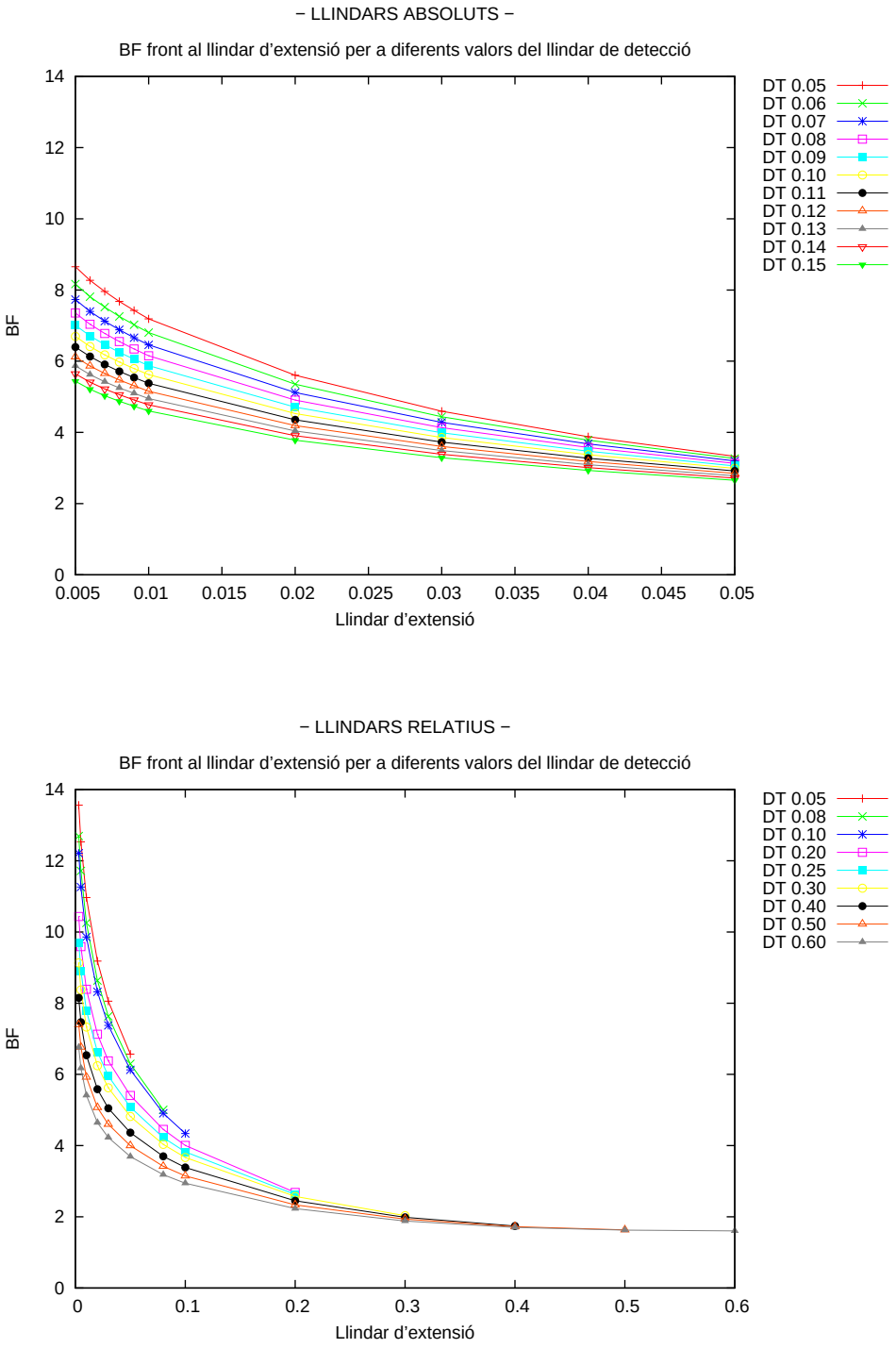


Figura 3.8: BF en funció del llindar d'extensió, per als dos tipus de llindar.

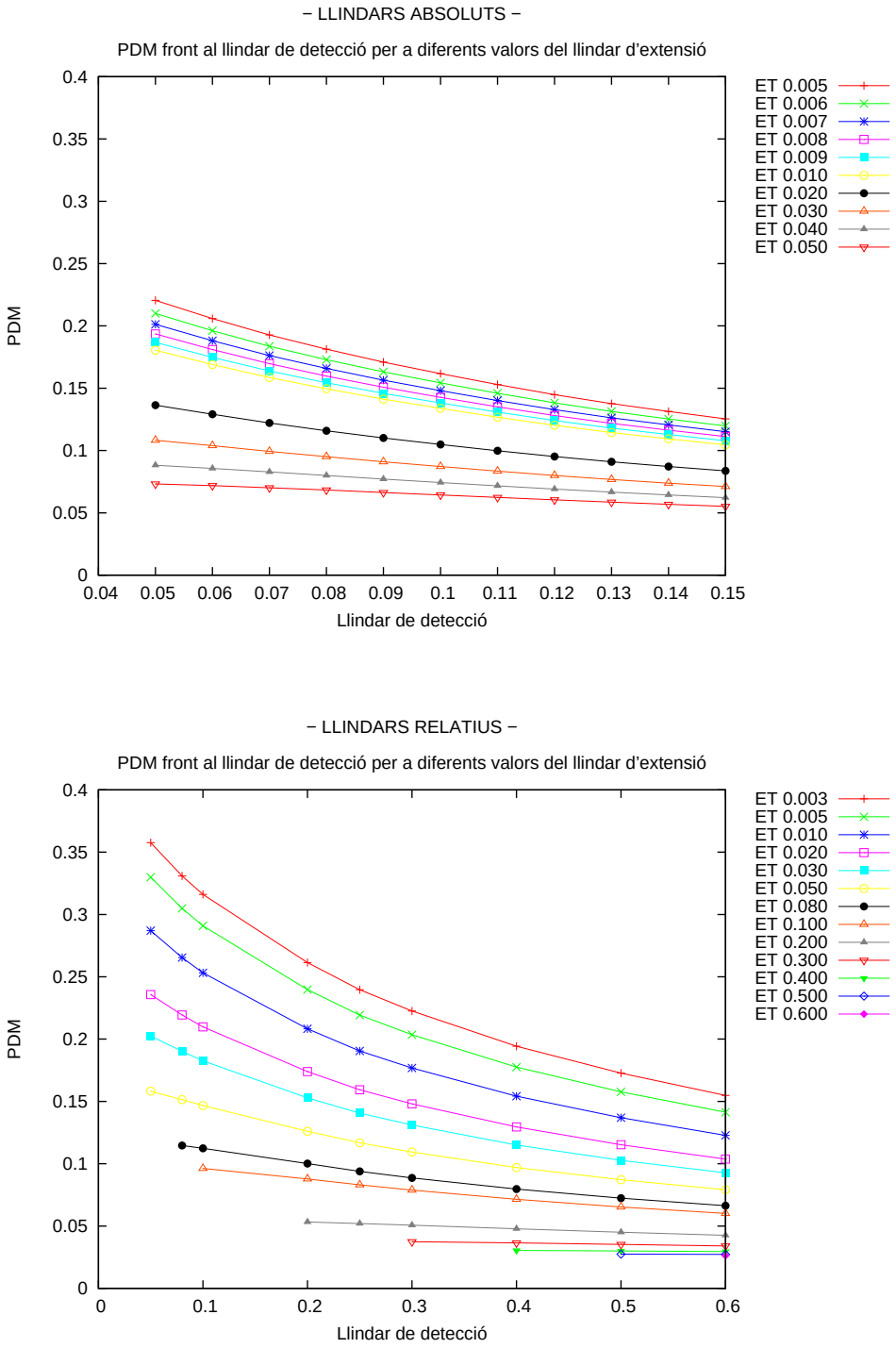


Figura 3.9: *PDM* en funció del llindar de detecció, per als dos tipus de llindar.

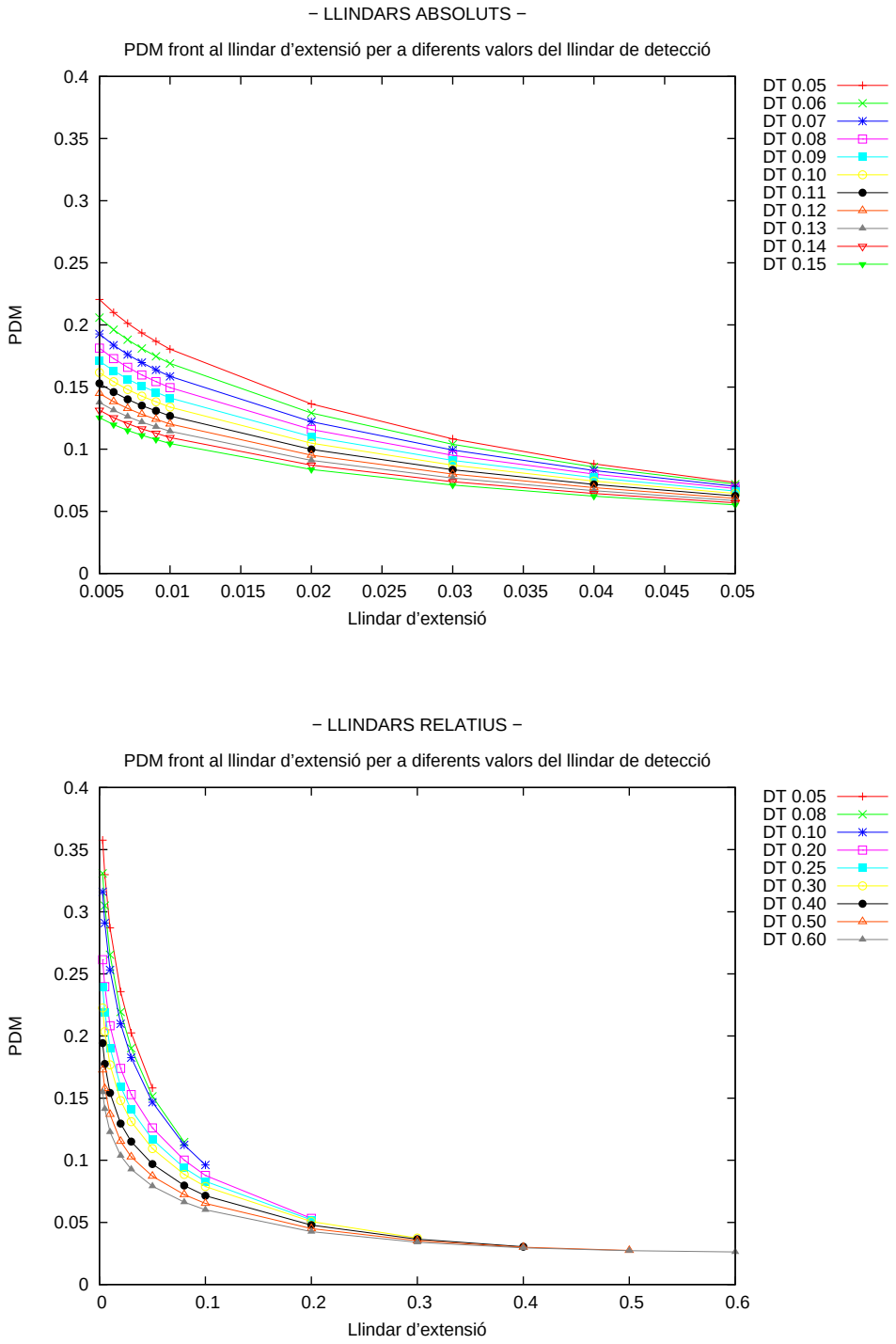


Figura 3.10: *PDM* en funció del llindar d'extensió, per als dos tipus de llindar.

Comparativa dels Llindars Absoluts Front als Relatius

En la Taula 3.4 es mostren alguns dels millors valors obtinguts amb els dos tipus de llindar.

Taula 3.4: Comparativa entre el PA i el PDM, per a distintes combinacions i amb els dos tipus de llindar.

Detecció Llindar	*Extensió Llindar*	*Tipus de Llindar*	PDM	PA (%)
0.050	0.005	Absoluts	0.22	95.46
0.090	0.008	Absoluts	0.15	91.31
0.090	0.010	Absoluts	0.14	90.22
0.120	0.007	Absoluts	0.13	90.03
0.050	0.003	Relatius	0.36	97.51
0.050	0.030	Relatius	0.20	94.10
0.300	0.010	Relatius	0.18	93.45
0.600	0.050	Relatius	0.08	79.48

De la taula i de les gràfiques anteriors es poden traure dues conclusions: en primer lloc, s'observa que en general el PA resulta més favorable amb l'ús de llindars relatius que no amb absoluts. En segon lloc, la densitat dels grafs resultants és notablement més elevada amb llindars relatius que amb llindars absoluts.

Ambdós fenòmens tenen la mateixa explicació. Quan la pronunciació d'un fonema està acabant o la del pròxim va a començar, les probabilitats fonètiques passen a ser molt reduïdes fins al punt de què durant unes *frames* no hi ha cap que predomine. Baix aquesta situació, l'ús de llindars absoluts provoca que la detecció tendisca a ser més pobra, ja que les probabilitats difícilment superaran els llindars. Això deriva en una reducció del PA.

La tècnica dels llindars relatius està encaminada a minimitzar els efectes d'aquest problema. Encara que les probabilitats siguen més reduïdes, com que els llindars depenen de la probabilitat màxima per a cada *frame*, es garanteix que les unitats seran detectades. Aquesta estratègia funciona, ja que el PA ofereix uns resultats més satisfactoris que amb l'ús de llindars absoluts, però a canvi d'obtenir uns grafs més densos.

Aquest fenomen té lloc perquè quan s'analitzen les *frames* que es troben en la situació descrita, hi ha moltes unitats fonètiques que tenen una probabilitat baixa i que seran detectades, creant els nodes i els arcs corresponents que amb l'ús de llindars absoluts no es crearien. Per aquest motiu, la densitat resulta prou més elevada amb l'ús de llindars relatius que amb els absoluts.

La Figura 3.11 mostra el PA en funció del PDM per als distints valors dels llindars. S'observa que el valor del PA és redueix quan menys densos són els grafs. No obstant, la caiguda del PDM és molt més accentuada que la del PA, fins al punt en què amb un PA que estiga al voltant del 90% s'obtenen grafs que són gairebé la meitat de densos.

Amb les dades obtingudes, cal elegir el tipus i els valors dels llindars. Com que l'objectiu és aconseguir el millor resultat en la tasca de DAF, és lògic elegir aquella combinació de llindars que oferisca el PA més elevat. Però a la vegada, cal tenir en compte que la densitat (PDM) dels grafs no siga excessivament elevada. Els dos millors resultats són el 95.46% amb llindars absoluts, i el 97.51% amb llindars relatius. Pel que respecta al PDM, amb llindars absoluts aquest resulta molt més reduït que amb llindars relatius (un 22.04% front a 35.77%).

Com que la diferència del PA tampoc és massa elevada, per a la posterior experimentació s'han utilitzat grafs generats amb llindars absoluts, amb un llindar de detecció del 5% (0.050) i amb un llindar d'extensió del 0.5% (0.005).

No obstant, atenent als resultats de la Figura 3.11 resulta interessant repetir la mateixa experimentació amb altres valors per als llindars, aquells que tinguen un PA al voltant del 90%. Aquesta experimentació es realitza una vegada fixats els millors valors per al GSF, l'MDF_{factor} i el $PhIP$.

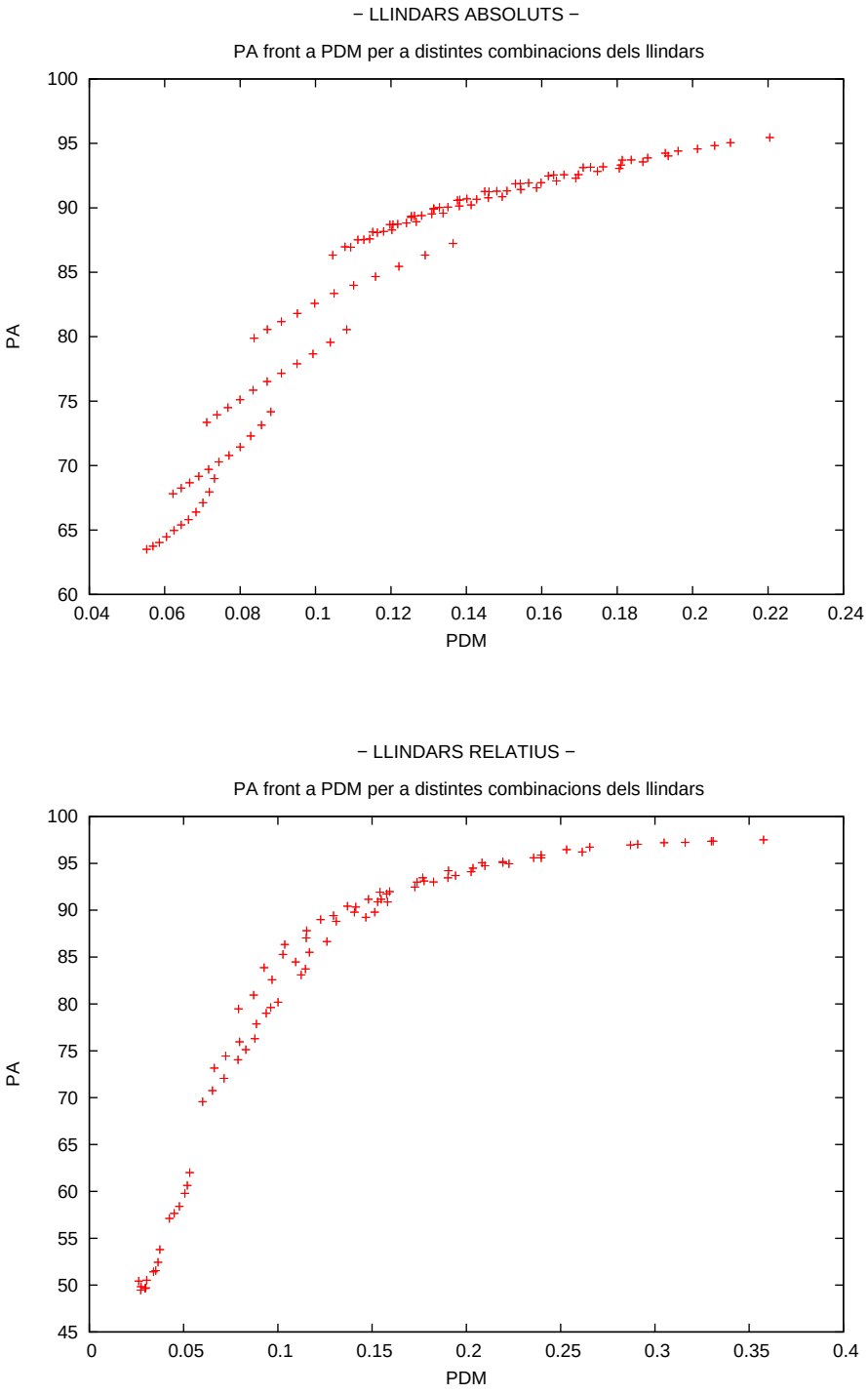


Figura 3.11: PA front a PDM, per als dos tipus de llindar.

3.6.2 Resultats de DAF

Model de Llenguatge vs PhIP

En primer lloc, analitzem la evolució del PA en funció del GSF i el $PhIP$. S'observa que quan el $PhIP$ val 0 (quan no hi ha cap penalització per inserció de fonemes) el PA creix fins que el GSF val 4. Atenent exclusivament al $PhIP$, quan el model de llenguatge no té cap efecte (quan el GSF val 0), es pot vore que el PA creix fins que el PhIP arriba a valer -5. Resulta curiós que quan el model de llenguatge no té pes, aplicant un $PhIP$ entre -4 i -6 ja s'obtenen resultats que s'aproximen prou al màxim PA.

Quan es combinen ambdós valors, el PA arriba al seu màxim quan el $PhIP$ pren valors entre -3 i -4 i quan el GSF oscil·la entre 1 i 2. El valor màxim per al PA és de 67.40%.

De la mateixa manera, per a valors del GSF superiors a 3, el $PhIP$ empitjora el resultat quan pren valors inferiors a -3.

Model de Durada Fonètica vs PhIP

L'evolució que presenta el PA pel que fa al model de durada fonètica i al $PhIP$ és la mostrada en la Figura 3.13. En aquest cas el model de llenguatge no té cap pes en la descodificació, i per tant el GSF val 0 en aquest experiment.

Es pot observar que quan el $PhIP$ val 0, el PA creix a mesura que creix l'MDF_{factor}, arribant al seu màxim quan aquest val 6, obtenint així un dels millors resultats d'aquest experiment.

Quan l'MDF_{factor} val 0, l'únic paràmetre que influeix en el procés de DAF és el $PhIP$, l'evolució del qual ja ha sigut descrit prèviament.

El màxim PA (62.50%) s'obté quan el $PhIP$ val -4 i l'MDF_{factor} val 1. Salta a la vista que aquest resultat és pitjor que el 67.40% obtingut amb l'aplicació del $PhIP$ i el model de llenguatge.

Model de Llenguatge vs Model de Durada Fonètica

En aquest apartat s'estudia l'evolució del PA en funció de l'MDF_{factor} i el GSF, quan el $PhIP$ val 0.

En la Figura 3.14 s'observa que quan l'MDF no aporta cap informació, el GSF millora els resultats passant d'un PA del 30% a un del 47%, aproximadament. Més elevat és el PA resultant d'aplicar només l'MDF, arribant al seu màxim amb un PA del 61.91% quan l'MDF_{factor} val 6.

Quan ambdós models s'apliquen conjuntament, el PA màxim al qual s'arriba és del 67.41% quan l'MDF_{factor} val 3 i el GSF val 2.

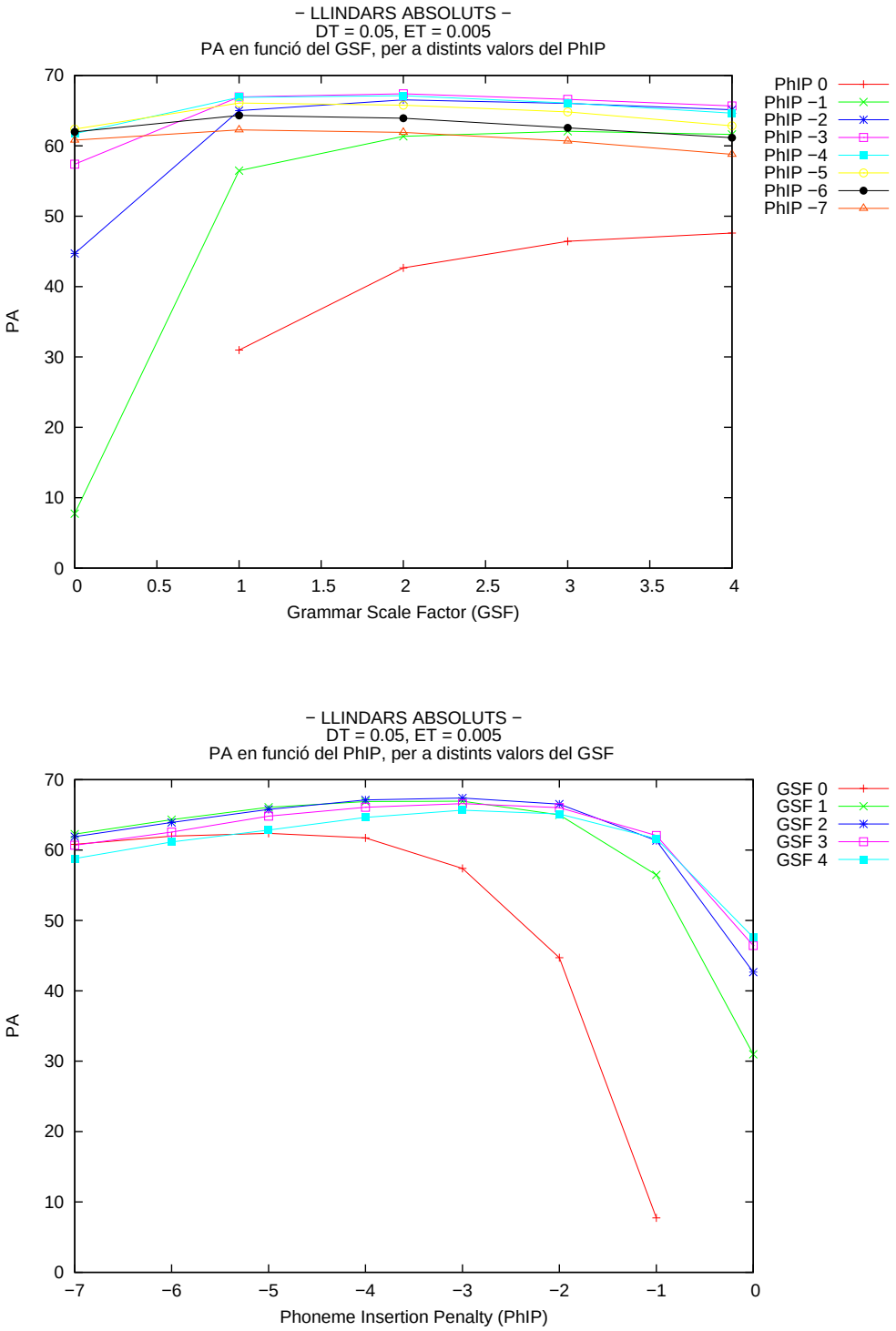


Figura 3.12: PA per a DAF, en funció del GSF i el $PhIP$.

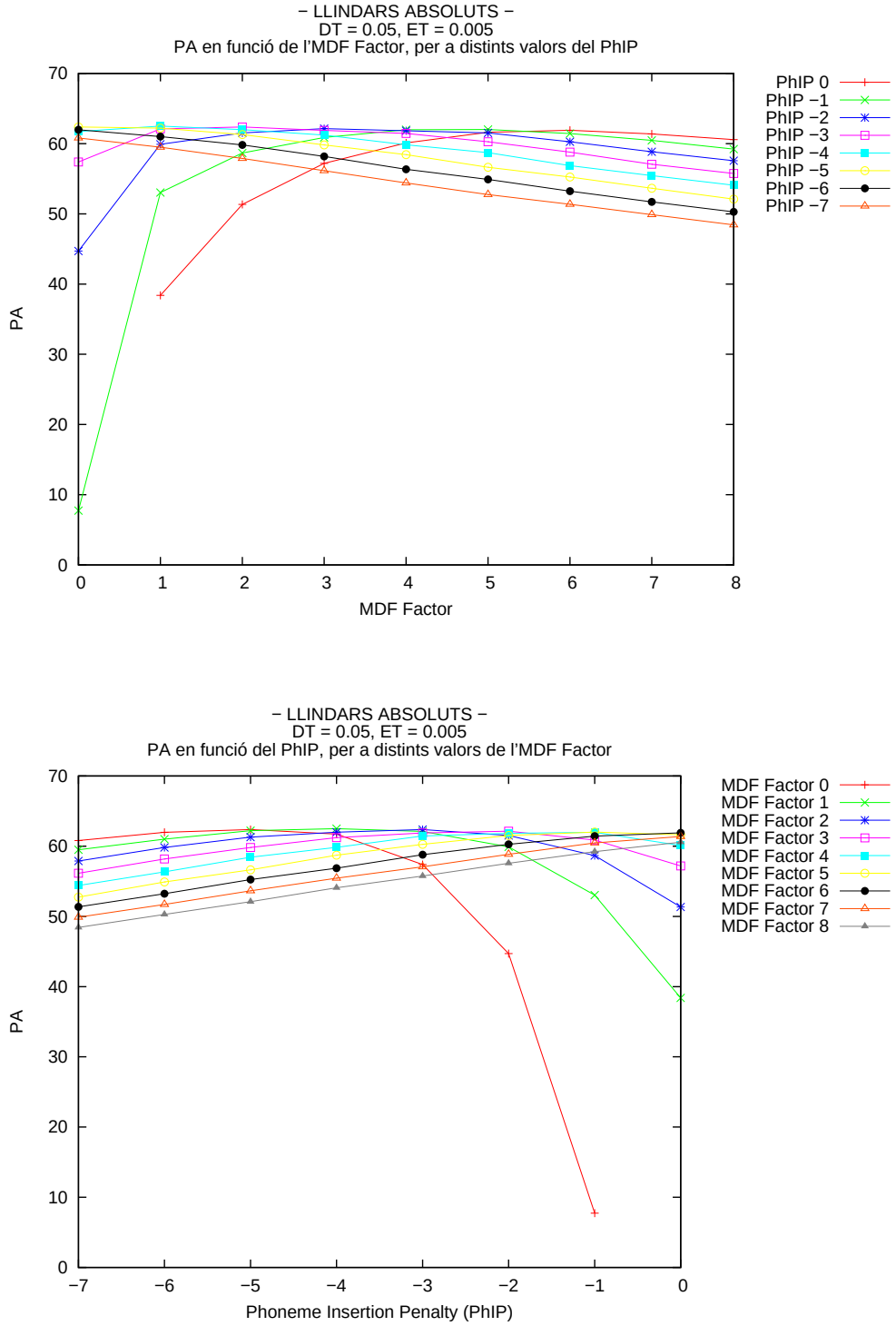


Figura 3.13: PA per a DAF, en funció de l'MDF_{factor} i el $PhIP$.

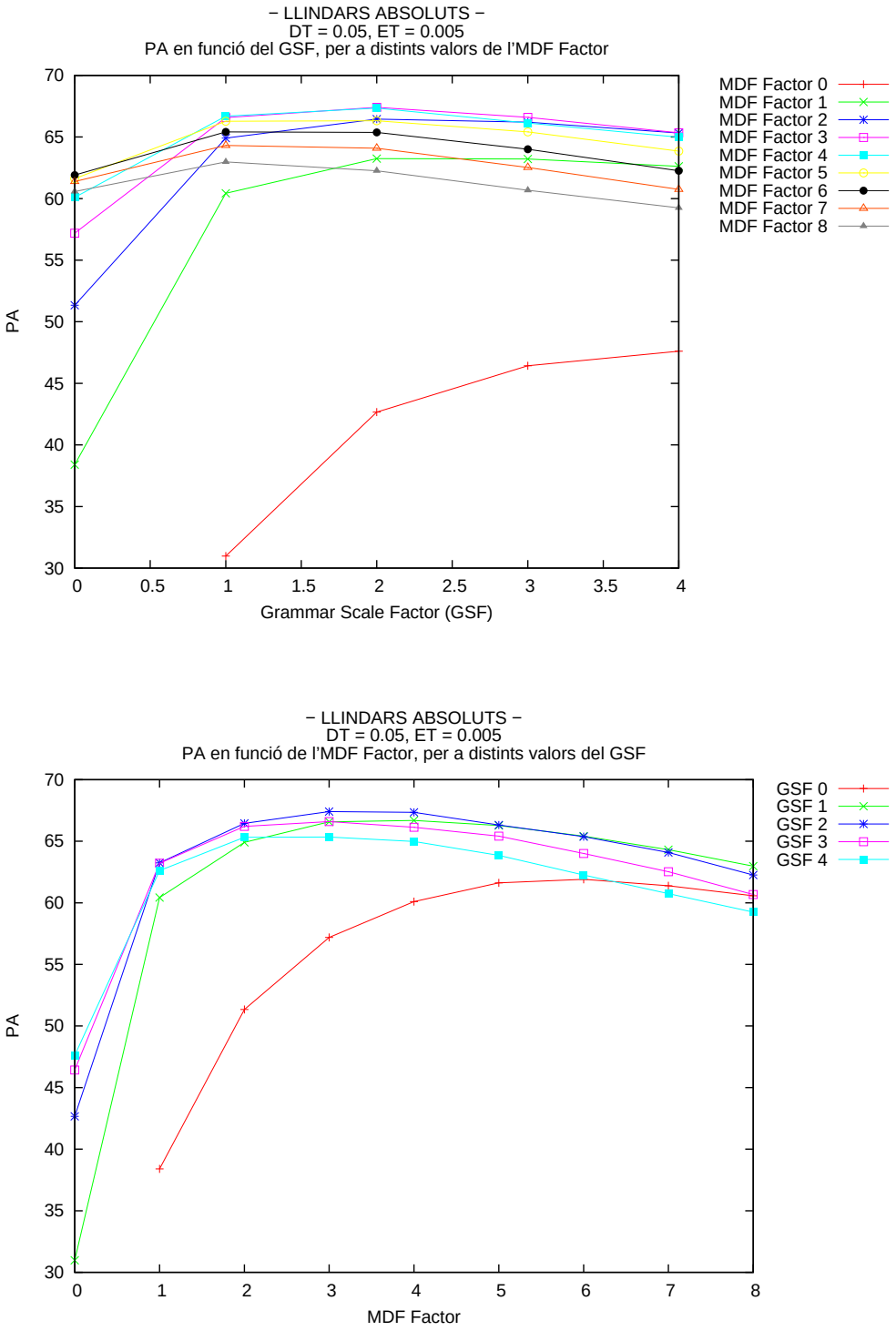


Figura 3.14: PA per a DAF, en funció de l'MDF_{factor} i el GSF.

Model de Llenguatge vs Model de Durada Fonètica vs PhIP

La taula 3.5 resumeix els millors valors obtinguts al llarg d'aquesta experimentació i considerant els tres factors, tant conjuntament com per separat:

Taula 3.5: Valors més elevats del PA per a DAF, per a cada experiment.

$PhIP$	GSF	MDF_{factor}	PA_{DAF}
-5	0	0	62.38%
0	4	0	47.61%
0	0	6	61.91%
-3	2	0	67.40%
-4	0	1	62.50%
0	2	3	67.41%
-2	2	1	67.61%

En primer lloc, cal destacar que no pareix existir una única combinació de valors òptima, ja que els darrers resultats són tots prou similars i per a distints valors dels factors. Sense importar la combinació dels tres factors, per als grafs generats el PA tendeix a ser d'un 67.50% aproximadament.

Resultats amb Altres Combinacions de Llindars

Com s'ha explicat anteriorment, es repeteix tota la experimentació duta a terme però per a distintes combinacions dels llindars, i tant per a llindars absoluts com relatius. Aquesta vegada, els rangs de valors per als factors que van a ser provats són els següents:

- $PhIP \in [-1, -4]$
- $GSF \in [1, 2]$
- $MDF_{factor} \in [1, 4]$

Aquestos rangs són aquells per als quals s'han obtingut els millors resultats en l'experimentació realitzada fins al moment. La Taula 3.6 resumeix els millors resultats obtinguts per a la nova experimentació.

Taula 3.6: PDM vs PA_{ref} (respecte la seqüència de referència) vs PA_{DAF} (obtingut durant la descodificació), per a diverses combinacions de llindars.

Llindar de Detecció	*Llindar d'Extensió*	*Tipus de Llindars*	PDM	PA_{ref} (%)	PA_{DAF} (%)
0.050	0.005	Absoluts	0.22	95.46	66.49
0.090	0.008	Absoluts	0.15	91.31	65.88
0.090	0.010	Absoluts	0.14	90.22	65.38
0.120	0.007	Absoluts	0.13	90.03	66.32
0.050	0.010	Relatius	0.29	96.96	66.58
0.050	0.030	Relatius	0.20	94.10	65.11
0.300	0.010	Relatius	0.18	93.45	66.70
0.500	0.010	Relatius	0.14	90.43	65.91
0.600	0.050	Relatius	0.08	79.48	65.57

Per a llindars absoluts, s'han elegit aquelles combinacions en què el PA_{ref} sempre supera el 90%. Al augmentar el valor dels llindars, el PA_{ref} es redueix i amb ell la densitat dels grafs, passant d'un PDM del 0.22 a un de 0.13. I al contrari del que calia esperar, el PA per a la descodificació (PA_{DAF}) no empitjora gairebé a pesar de contenir una menor part de la seqüència de referència.

En quant als llindars relatius, per a aquells que són menys restrictius la densitat dels grafs és de 0.29% amb un 96.96% de la seqüència de referència i un PA_{DAF} del 66.58%. Si augmentem el valor dels llindars al 60% el de detecció i 5% el d'extensió, el PA_{ref} resulta ser del 79.48% i la densitat dels grafs passa a ser del 0.08. En altres paraules, sacrifiquem una bona part de la seqüència de referència continguda i obtenim uns grafs molt menys densos i que no deriven en una pèrdua del PA_{DAF}.

El millor resultat que s'ha obtingut ha sigut per a llindars relatius, amb uns llindars de detecció i extensió de 0.30 i 0.01 respectivament. Amb aquesta combinació, el PDM és de 0.18.

La conclusió que es pot obtenir és que encara que es perda part de la seqüència de referència, la major part dels arcs i nodes que l'algorisme elimina són superflus, sense cap utilitat de cara a la descodificació.

Resultats de DAF sobre Vectors de Probabilitats

Posant el reconeixedor automàtic de la parla a reconèixer fent ús d'un model de llenguatge a nivell de fonemes, el mateix utilitzat per a fer DAF a partir dels grafs, i considerant com a paraules les pròpies unitats fonètiques, s'obté

com a resultat un $PA = 73.3\%$.

Capítol 4

Discussió i Conclusions

En el present projecte s'ha implementat un algorisme per a la construcció de grafs de fonemes, els quals són una representació intermèdia d'allò que ha sigut pronunciat al llarg d'una locució. Encara que els grafs poden ser emprats amb distintes finalitats, en aquest projecte ens hem centrat en la utilització d'aquestos per a la descodificació acústic-fonètica (DAF). Amb la finalitat de poder avaluar correctament els resultats per a DAF amb l'ús de grafs de fonemes, també s'ha dut a terme una descodificació basada en l'algorisme de Viterbi aplicat sobre la informació proporcionada a partir dels models acústics, tal i com es sol fer amb les tècniques convencionals.

El millor resultat obtingut amb els grafs de fonemes és d'una taxa d'encert del 66.70%, mentre que amb l'algorisme per a DAF convencional aquest resultat ha sigut del 73.3%. No obstant, l'algorisme de Viterbi utilitza tota la informació proporcionada a partir dels models acústics mentre que els grafs de fonemes només empren el 18% d'aquesta informació. En altres paraules, s'ha aconseguit no inserir més d'un 80% la informació que no resulta tan útil de cara a la descodificació a canvi de minvar en un 6.59% el resultat d'aquesta.

Els posteriors mòduls que hagen de processar aquesta informació poden treballar amb qualsevol de les dues representacions intermèdies, però el fet de què una continga un 80% menys de la informació que l'altra pot traduir-se en un gran estalvi en termes de memòria i temps de còmput. A més, els grafs són una estructura ben coneguda i pot resultar més simple treballar amb ells que no amb vectors de probabilitats.

A l'inici del projecte s'ha plantejat l'ús dels grafs de fonemes per a l'alineació del text del discurs amb l'àudio que conté el propi discurs, la qual cosa pot permetre, per exemple, subtitular vídeos de forma automàtica. Encara que en el present projecte no s'han utilitzat els grafs de fonemes amb aquesta finalitat, el fet que l'inici i el final de les pronunciacions detectades ja vinguen

donades en el propi graf pot ser de gran ajuda per a la realització d'aquesta tasca.

Un altre dels objectius plantejats ha sigut la utilització dels grafs per a la indexació de l'àudio i la cerca de termes. En el present projecte no s'ha obtingut cap dada definitiva amb què es puga afirmar si els grafs poden utilitzar-se satisfactòriament o no per a aquesta tasca. No obstant, el fet de què s'elimine la major part de la informació resulta interessant de cara a la investigació del seu ús per a la indexació de fitxers d'àudio i la cerca en aquestos.

En resum, els grafs de fonemes poden ser una bona alternativa als vectors de probabilitats, doncs encara que minven lleugerament el resultat aconsegueixen compactar notablement la informació proporcionada pels models acústics. Per tant, en aplicacions on l'exactitud en el reconeixement no resulte crítica, la utilització de grafs de fonemes pot resultar adequada i suposar un estalvi en quant a temps de còmput i memòria.

Bibliografia

[1] Hanazawa, K., Minami, Y., Furui, S.: An Efficient Search Method for Large-Vocabulary Continuous-Speech Recognition. In: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference, vol. 3, pp. 1787-1790. IEEE (1997)

[2] Nocera, P., Linares, G., Massonié, D., Lefort, L.: Phoneme Lattice Based A* Search Algorithm for Speech Recognition. In: Text, Speech and Dialogue, pp. 301-308. Springer Berlin Heidelberg (2006)

[3] Burget, L., Černocký, J., Fapšo, M., Karafiát, M., Matějka, P., Schwarz, P., Smrž Szöke, I.: Indexing and Search Methods for Spoken Documents. In: Text, Speech and Dialogue, pp. 351-358. Springer Berlin Heidelberg (2006)

[4] Pinto, J., Szoke, I., Prasanna, S.R.M., Hermansky, H.: Fast Approximate Spoken Term Detection from Sequence of Phonemes. In: Proceedings of the ACM SIGIR Workshop on Searching Spontaneous Conversational Speech, pp. 08-45 (2008)

[5] Yu, P., Chen, K., Ma, C., Seide, F.: Vocabulary-Independent Indexing of Spontaneous Speech. In: Speech and Audio Processing, vol. 13, pp. 635-643, IEEE Transactions on (2005)

[6] Ferrieux, A., Peillon, S.: Phoneme-level Indexing for Fast and Vocabulary-Independent Voice/Voice Retrieval. ESCA ETRW Workshop: Accessing Information in Spoken Audio (1999)

[7] Gómez, J. A., Sanchis, E.: Using Word Graphs as Intermediate Representation of Uttered Sentences. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 284-291. Springer Berlin Heidelberg (2012)